ulm university universität
uulm

Universität Ulm
Fakultät für Mathematik und
Wirtschaftswissenschaften

A Bayesian Multi-Population Mortality Projection Model

Masterarbeit

in Wirtschaftsmathematik

vorgelegt von
Lukas Josef Hahn
am 17. Dezember 2014

Gutachter

Jun.-Prof. Dr. Marcus C. Christiansen
Prof. Dr. Hans-Joachim Zwiesler

Acknowledgements

I would like to express my sincere gratitude to my supervisor Jun.-Professor Dr Marcus C. Christiansen for the support and guidance at all stages of this work. His invaluable advice and insightful questions throughout the term improved this thesis significantly. My thanks are further extended to my second reader Professor Dr Hans-Joachim Zwiesler, whose assistance during my studies at Ulm University meant a constant enrichment to my performance in general and the final outcome of this work in particular. I thank all other members of the Faculty of Mathematics and Economics of the University of Ulm and the Institute of Financial and Actuarial Mathematics in Ulm who helped me in preparation of this thesis, most notably Dr Jan-Philipp Schmidt and Dr Matthias Börger.

This work has partly been performed during my time at the Department of Statistics and Actuarial Science of the University of Waterloo in Waterloo, Ontario. I gratefully acknowledge the warm hospitality and financial support during this time. Special thanks are given to my program coordinator Mary Lou Dufton for always providing me with a professional work environment and my supervisor Professor Dr Shoja'eddin Chenouri, whose fruitful comments on statistics and general scientific work have ultimately had a positive impact on this work. I warmly thank the entire team behind the Human Mortality Database (2014) for excellent datasets on mortality, which have become fundamental for the development of the model presented in this thesis. Further thanks are given to Talanx-Stiftung (Talanx foundation) within the Stifterverband für die Deutsche Wissenschaft. Their funding supported me during an extensive time of my graduate studies.

Lastly, and most importantly, I wish to thank my parents, Dr Angela and Norbert Hahn, and my significant other, Jared Best, for their constant care and encouragement. This work would clearly not have been possible without them.

Contents

List of Acronyms

AR	autoregressive
ARMA	Autoregressive Moving Average
BMPMP	Bayesian Multi-Population Mortality Projection
CBD	Cairns-Blake-Dowd
iid	independent and identically distributed
LC	Lee-Carter
MA	moving average
MCMC	Markov Chain Monte Carlo
ML	maximum-likelihood
VAR	Vector Autoregressive
VARMA	Vector Autoregressive Moving Average
VECM	Vector Error Correction Model
w.l.o.g.	without loss of generality

List of Figures

List of Countries

The following countries are considered in the case studies in Chapter 4, ordered by their abbreviations through the corresponding Internet top-level domains (i.e. ISO 3166-1 alpha-2 codes with specific replacement of GB by UK for the United Kingdom). Comments on territorial coverage are made w.r.t. the maximum time horizon for respective analyses in this work.

AT	Austria	Today's territory remained unchanged throughout the time horizon.
CZ	Czech Republic	Today's territory of the Czech Republic, i.e. the Czech lands within Czechoslovakia before 1993 (also known as the Czech Socialist Republic between 1968–1990 and the Czech Republic between 1990–1993) and the independent Czech Republic since 1993 after the dissolution of Czechoslovakia.
DE	Germany	German data are restricted to the dataset for the population in the territory of former West Germany (i.e. the Federal Republic of Germany until 1990 and the old states within the re-unified Federal Republic of Germany since then), i.e. excluding East Germany (i.e. the German Democratic Republic until 1990 and the new states within the re-unified Federal Republic of Germany since then), due to different patterns in mortality in a divided Germany and even after re-unification. The current Internet top-level domain *DE* for the entire republic is conveniently used to denote the territory of former *West Germany*. All case studies involve data for West Germany only, but this exact specification is usually dropped and the population is referred to Germany for simplicity.

ES	Spain	Today's territory of Spain consisting of the Iberian mainland, the Balearic and Canary islands, and the North African exclaves Ceuta and Melilla. The territory remained unchanged throughout the time horizon.
FR	France	Today's territory of France consisting of Metropolitan France (France métropolitaine) only, i.e. the data include European mainland France with all its islands in the Atlantic Ocean, the English Channel and the Mediterranean Sea (including Corsica), but exclude Overseas France, i.e. all overseas departments (Guadeloupe, Guyane, Martinique, Réunion) and all overseas territories and collectivities (including New Caledonia and French Polynesia). The territory remained unchanged throughout the time horizon.
HU	Hungary	Today's territory remained unchanged throughout the time horizon.
IT	Italy	Today's territory remained unchanged throughout the time horizon.
PL	Poland	Today's territory remained unchanged throughout the time horizon.
UK	United Kingdom	Today's territory of the United Kingdom consisting of the four countries England, Scotland, Wales, and Northern Ireland only, i.e. the data include the island of Great Britain and the British north-eastern part of the island of Ireland, but exclude the Kingdom's fourteen overseas territories and the three Crown Dependencies of Guernsey, Jersey, and the Isle of Man, which are not part of the United Kingdom. The territory remained unchanged throughout the time horizon.

List of Symbols

The following symbols are defined as in Chapters 3 and 4. They may be defined differently elsewhere in this work.

a	Acceptance probability in Metropolis-Hastings algorithm
A	$m \times m$ positive definite constant matrix in the prior for the VECM measuring uncertainty in Ω
α	Loading matrix in the VECM containing weights for cointegration relationships of dimension $m \times r$ with full rank r
β	Cointegration matrix in the VECM containing cointegration relationships with dimension $m \times r$ with full rank r
β_l	Lower block matrix of dimension $m - r \times r$ for β in the normal linearisation $\beta = (I_r, \beta_l')'$
c	$m \times r$ matrix $(I_r, 0_{r \times m-r})'$
c_{MH}	Tuning parameter for Σ_{MH}
c_r	Normalising constant in the prior for the VECM
$c_\perp$	$m \times m - r$ matrix $(0_{m-r \times r}, I_{m-r})'$
Cov	Covariance matrix
Γ	$m \times (k-1)m$ matrix $(\Gamma_1, \ldots, \Gamma_{k-1})$
Γ_i	$m \times m$ AR parameter matrix for lag i in the VECM
$\mathit{\Gamma}$	Positive real Gamma function as defined in Definition D.1
$\mathit{\Gamma}_b$	Generalised Gamma function as defined in Definition D.1
d	Number of constants in the VECM for the deterministic trend, equals one in the BMPMP model
diag	Block matrix with elements on the diagonal and zeros elsewhere
D_t	d-dimensional constant of deterministic trends in the VECM, equals the scalar one in the BMPMP model
D_{xpt}	Observed number of deaths for age x, population p at calendar year t

D_{xpgt}	Observed number of deaths for age x, gender g, population p at calendar year t
$\mathcal{D}$	Set of D_{xpt} (or D_{xpgt}) for all x, p, t (and g) with $p \neq p^*$
$\mathcal{D}_t$	Set of D_{xpt} (or D_{xpgt}) for all x, p (and g) with $p \neq p^*$ for calendar year t
ΔK	$m \times T - k$ matrix $\Delta K = (\Delta\kappa_{k+1}, \ldots, \Delta\kappa_T)$
$\Delta\kappa_t$	m-dimensional vector $\kappa_t - \kappa_{t-1}$
exp	Exponential function
E	Expectation
E_{xpt}	Exposure-to-risk for age x, population p at calendar year t
E_{xpgt}	Exposure-to-risk for age x, gender g, population p at calendar year t
$\mathcal{E}$	Set of E_{xpt} (or E_{xpgt}) for all x, p, t (and g) with $p \neq p^*$
$\mathcal{E}_t$	Set of E_{xpt} (or E_{xpgt}) for all x, p (and g) with $p \neq p^*$ for calendar year t
ε	$m \times T - k$ matrix $(\varepsilon_{k+1}, \ldots, \varepsilon_T)$
ε_t	m-dimensional iid multivariate normal error terms in the VECM with zero mean and covariance matrix Ω for calendar year t
f	Probability density function or superscript denoting females
f_A	Tuning function for choice of A in the empirical Bayes approach
g	Index or superscript for gender
η_{xpt}	Linear predictor, i.e. right-hand side, of CBD model for age x, population p at calendar year t
η_{xpgt}	Linear predictor, i.e. right-hand side, of CBD model for age x, population p, gender g at calendar year t
$\mathcal{H}$	Set of hyperparameters $\{\phi, \Gamma, \alpha, \beta, \Omega\}$
i	Index or iteration step
I_m	$m \times m$ identity matrix
IW_a	a-dimensional Inverse Wishart distribution as defined in Definition D.7
k	Lag order in the VECM
K_1	$m \times T - k$ matrix $(\kappa_k, \ldots, \kappa_{T-1})$
K_2	$\left((\Delta\kappa'_k, \ldots, \Delta\kappa'_2)', (\Delta\kappa'_{k+1}, \ldots, \Delta\kappa'_3)', \ldots, (\Delta\kappa'_{T-1}, \ldots, \Delta\kappa'_{T-k+1})'\right)$ of dimension $(k-1)m \times T - k$
$\mathcal{K}$	Set of parameters κ_t for $t = 1, \ldots, T$

$\mathcal{K}_{-t}$	Set of parameters κ_s for $s = 1, \dots, t-1, t+1, \dots, T$
κ_t	m-dimensional time series vector $(\kappa_t^0, \kappa_t^x, \kappa_t^{p_1}, \kappa_t^{p_2}, \dots)$ for calendar year t
κ_t^0	Intercept in CBD model for calendar year t
κ_t^g	Main effect for gender g in CBD model for calendar year t
κ_t^p	Main effect for population p in CBD model for calendar year t
κ_t^{pg}	Interaction for population p and gender g in CBD model for calendar year t
κ_t^x	Main effect for age per one unit increase in x in CBD model for calendar year t
$\kappa_t^{x^2}$	Main effect for quadratic age per one unit increase in x in CBD model for calendar year t
κ_t^{xg}	Interaction for gender g and age per one unit increase in x in CBD model for calendar year t
$\kappa_t^{x^2 g}$	Interaction for gender g and quadratic age per one unit increase in x in CBD model for calendar year t
κ_t^{xp}	Interaction for population p and age per one unit increase in x in CBD model for calendar year t
$\kappa_t^{x^2 p}$	Interaction for population p and quadratic age per one unit increase in x in CBD model for calendar year t
log	Natural logarithm
L	Likelihood function
λ_A	Tuning factor for f_A under the choice $f_A(M) = (\lambda_A^2 m_{ij}^2)_{ij}$ for square matrices $M = (m_{ij})_{ij}$
λ_α	Constant in the prior for the VECM measuring uncertainty in α
λ_b	Constant in the prior for the VECM measuring baseline uncertainty in Γ
λ_l	Constant in the prior for the VECM measuring lag-dependent shrinkage in Γ
m	Dimension of VECM, i.e. number of parameters in CBD model, or superscript denoting males
m_{xpt}	Central death rate for age x, population p at calendar year t
m_{xpgt}	Central death rate for age x, population p, gender g at calendar year t
min	Minimum function

$MN_{a\times b}$	Matrix-Normal distribution of dimension $a \times b$ as defined in Definition D.4
$Mt_{a\times b}$	Matrix-t distribution of dimension $a \times b$ as defined in Definition D.5
μ_t	Prior mean vector for $\kappa_t \mid k$ for $t = 1, \ldots, k$
n_p	Number of populations
N	Number of iterations in MCMC algorithm
N_m	m-variate normal distribution
p	Index or superscript for population
p^*	Index or superscript for reference population (overall sample)
P	Probability measure in the underlying (known but unspecified) probability space
$\mathcal{P}$	List of populations
Poi	Poisson distribution
Π	Matrix product $\alpha\beta'$ in the VECM of dimension $m \times m$ and rank r
q	Constant in prior for the VECM measuring uncertainty in Ω
q_{xpt}	Mortality rate for age x, population p at calendar year t
q_{xpgt}	Mortality rate for age x, population p, gender g at calendar year t
r	Cointegration rank, i.e. $r = \mathrm{rk}(\Pi) = \mathrm{rk}(\alpha) = \mathrm{rk}(\beta)$
rk	Rank of a matrix
s	Index for calendar years
Σ_{MH}	$m \times m$ covariance matrix for Metropolis-Hastings proposals, for simplicity $\Sigma_{\mathrm{MH}} = c_{\mathrm{MH}} A$
Σ	Prior covariance matrix for $\kappa_t \mid k$ for all $t = 1, \ldots, k$ if $\Sigma_1 = \cdots = \Sigma_k$
Σ_t	Prior covariance matrix for $\kappa_t \mid k$ for $t = 1, \ldots, k$
t	Index for calendar years
tr	Trace function, i.e. sum of all diagonal elements of a square matrix
T	Number of calendar years
ϕ	$m \times d$ parameter matrix in the VECM for deterministic trends, m-dimensional vector of ones in the BMPMP model
Φ	$m \times T - k$ matrix $(\phi, \ldots, \phi)$
U	Uniform distribution
Var	Variance
VAR(k)	Vector Autoregressive Model of order k
vec	Vectorisation operator as defined in Definition D.2

w	Posterior weight
W_a	a-dimensional Wishart distribution as defined in Definition D.6
x	Index for age
x_0	Minimum age for CBD model
Ω	$m \times m$ positive definite covariance matrix for all ε_t
$\widehat{\Omega}$	estimate for Ω based on a maximum-likelihood approach
Ω_Γ	$m(k-1) \times m(k-1)$ positive definite component matrix for the covariance matrix of Γ with block matrices $\Omega_{\Gamma_1}, \ldots, \Omega_{\Gamma_{k-1}}$ on its diagonal
Ω_{Γ_i}	$m \times m$ positive definite component matrix $\lambda_b^2 i^{-2\lambda_l} I_m$ for the covariance matrix of Γ_i
$0_{a\times b}$	$a \times b$ zero matrix
$'$	Transpose of a matrix
$\otimes$	Matrix Kronecker product as defined in Definition D.3
$^{(i)}$	Superscript denoting realisation in iteration step i of Metropolis-Hastings algorithm
*	Superscript denoting proposal in Metropolis-Hastings algorithm

1 Introduction

1.1 Mortality Forecasts for Several Populations

Aside from the obvious interest in life expectancies in the myriad of sociological or medical studies, mortality projections have always been an important feature in actuarial science. Valid mortality forecasts are fundamental for pension funds or life insurers, among many other financial institutions, to correctly price annuities, pension plans or life insurances, and to hedge against losses due to longevity risk. Usage of population-specific mortality rates through life tables for the calculation of annuities can be traced back to the 17^{th} century. Several deterministic mortality laws, which are still popular to this day, were established during the 19^{th} and 20^{th} centuries, in particular the prominent Gompertz law for the force of mortality by Gompertz (1825). The late 20^{th} century introduced stochasticity in mortality projection models to account for the general uncertainty associated with systematic and unsystematic risks, i.e. the possibility of errors due to the random nature of mortality events, sampling of historical mortality data, assumptions on continuous mortality measures, modelling approaches, estimation and, in particular, the forecasting techniques and uncertainty about future developments. In this context, the Lee-Carter (LC) model by Lee and Carter (1992) has without doubt evolved into one of the state-of-the-art models and motivated several other approaches comprising stochastic mortality projections. The opportunity of measuring uncertainty in longevity through quantiles and confidence intervals are vital features in actuarial applications.

Due to fast-paced and progressively increasing globalisation, mortality patterns for *different* populations in the Western world have been observed to assimilate over the last decades; see the discussion by Wilson (2001) for instance. Economic development, medical innovations, and international migration cause mortality rates of formerly more

isolated populations of developed countries to converge. As a consequence, such mortality patterns are by far not independent and it has become increasingly clear that even mortality forecasts for individual populations should be modelled based on all available data that contain necessary information. Joint mortality projections shall eliminate biological implausibilities, such as divergent behaviour resulting from individual mortality forecasts, and, due to the increase in available data, improve their statistical properties, particularly for the estimates of uncertainty. It is noteworthy, however, that sudden short-term discrepancies between similar populations are not uncommon as observed in Europe during the early 1990s. Findings in this work further suggest that the assimilation process is less pronounced than commonly stated. Despite their popularity in recent literature, prior convergence hypotheses in design stages of projection models seem to underestimate this complexity.

In addition to the direct effects of globalisation on supra-national mortality patterns, there has been a growing interest in learning about the particular dependencies between populations. Life insurers, pension funds, and other investors are stakeholders in globally organised financial markets. For example, life insurance portfolios may consist of contracts negotiated in different countries, which are priced individually based on national mortality data. From the global risk management point of view, however, worst-case scenarios require mortality projections which take into account the joint movement of mortality rates and, notably, the risk associated with universal shocks on longevity. Similar arguments apply to the common task of modelling mortality in a specified portfolio of populations in possibly different lines of business, which may substantially differ from macro-economic mortality data for the entire population due to adverse selection. In these instances, data quality may be substantially lower than what is observed for the parent population. Again, borrowing the strength from data of other populations incorporates quantification of dependence and improves the statistical properties of the forecasts.

1.2 A Brief Literature Overview

Current literature on actuarial mortality forecasting for developed countries, specifically since the mid 2000s, has shifted its focus on multi-population mortality projection mod-

els due to the aforementioned reasons. Their principal aims are an unbiased measurement of expected mortality and its associated uncertainty. The latter should quantify biologically plausible assimilation of longevity, dependencies between the individual populations, and the statistical improvements due to the increase in available data. Several contributions to this specific problem have been made, already leading to a variety of model approaches, most of which seek to extend successfully proven stochastic single-population models.

Mortality forecasting for single populations under inclusion of stochasticity finds its origin in the seminal paper by Lee and Carter (1992), who model logarithmic mortality rates through an additive model. Besides a deterministic basic pattern for the log-mortality over all ages under consideration, a bilinear term accounts for changes in this pattern over time, weighted differently for distinct ages. Uncertainty about the randomness in mortality is introduced through addition of homoskedastic random noise components with expectation zero, independent for each pair of age and time. The time-dependent changes in mortality patterns are forecast with time series models to construct confidence boundaries for future outcomes in the log-mortality. Much work on the LC model has been done ever since. Several authors postulate a Poisson distribution for the number of deaths to establish maximum-likelihood (ML) estimation, thereby allowing the random errors to be heteroskedastic, see Brouhns et al. (2002) for instance. Under different distribution assumptions, Czado et al. (2005) and Pedroza (2006) use Bayesian statistics for parameter estimation to avoid incoherence within the two-step calibration of the baseline model and the time series forecasts. Other important extensions are the inclusion of cohort effects by Renshaw and Haberman (2006) or overdispersion by, e.g., Delwarde et al. (2007).

The Cairns-Blake-Dowd (CBD) model by Cairns et al. (2006) contributes the main alternative to the LC approach in general stochastic mortality forecasting. Motivated by linear patterns in plots of the logits of observed mortality rates for older people versus age, a Binomial regression with parameters for the intercept and slope is applied with the mortality rates for each calendar year. In a similar manner as before, the bivariate time series of intercept and slope coefficients is forecast into the future. The evolution of the intercept describes general improvements in mortality rates, whereas the change in the slope determines differences in the benefits for distinct age groups. The flexibility

of adjustments in the linear predictor and the well-understood behaviour of generalised linear models has lead to further modifications of this approach, too, such as inclusion of effects for cohorts or quadratic patterns in the mortality logit profiles. A comparison of a wide range of different CBD models and the LC approach is provided by Cairns et al. (2009).

Multi-population mortality forecasting in the stochastic framework dates back to the seminal work by Li and Lee (2005). Their augmented common factor model is an extension of the LC model for several countries with country-specific baseline patterns and additional bilinear terms for each country, which describe individual deviations from the main evolution. Jarner and Kryger (2011), Dowd et al. (2011), and Cairns et al. (2011b) consider models with only two populations, in which time series forecasts are realised w.r.t. the differences in the mortality trends between a population of interest and a larger reference population to ensure a non-divergent behaviour. Zhou et al. (2012) apply multivariate time series models to the bivariate time series of model parameters for two general populations, which is extended to the case of an arbitrary number of populations by Ntamjokouen et al. (2014). Inclusion of country-specific covariates as explanatory variables is considered by Reichmuth and Sarferaz (2008). Other models do not build upon the LC framework. For example, Biatat and Currie (2010) use P-splines to model differences in mortality, but do not project mortality rates into the future, and Ahčan et al. (2014) replicate a population of interest by mixing other populations. It is noteworthy that the literature in actuarial science has seen further approaches that focus on projections of insured amounts rather than the number of deaths, and death-specific models have been in the focus of medical research.

However, despite all possible advances in the works outlined above, a general framework that addresses all problems with mortality projection, as outlined in Section 1.1, has not yet been developed. Such models are generally restricted to a certain maximum number of populations or their nested structures, and are not always as flexible as necessary to be sufficiently practicable.

1.3 Objectives

This work provides contribution to the scientific area of multi-population mortality projection through a model proposal, which addresses the challenge of biologically plausible joint forecasts of dependent populations in a globalised world. After an extensive literature review on stochastic mortality prediction models for single and multiple populations, the proposed model is derived, defined, and applied. In light of benefits and limitations of the reviewed models, the approach is based on a flexible augmented version of the CBD model for higher ages in an arbitrary number of populations. It comprises a hierarchical set-up, in which the model parameters of the CBD model are forecast using a Vector Error Correction Model (VECM), a specific representation of the multivariate Vector Autoregressive (VAR) model. Dependencies between different parameters, and hence between different populations, and biological plausibility in future mortality forces are accounted for not only by cross-correlation terms, but also by additional quantification of cointegration, i.e. common stable trends in the long-run. Application of the universal VECM further removes restrictions on the type of populations that can be analysed.

Along with the model formulation, further attention is devoted to parameter estimation given historical data and incorporation of the different types of uncertainty risks in the model. In addition to the model design, which already accounts for the random nature in mortality, the measurement of other sources of risks in the future developments of mortality is a main objective, and a Bayesian approach is established. The corresponding philosophy of postulating randomness for the unknown parameters leads to quantification of the future predictions via probabilistic distributions, which immediately yield Bayesian credibility regions for mortality forecasts. Moreover, by its very nature, Bayesian statistics aims to detect the characterising properties of such underlying distributions rather than determination of each individual parameter value. Since for any but the smallest number of populations, frequentist estimation methods such as ML lead to singularities in the estimating equations, the Bayesian approach is indeed necessary in this high-dimensional framework to forecast the mortality of an arbitrary and desired number of populations. Further advantages of the Bayesian approach are the reduction of inconsistency in parameter estimation between both levels in the hierarchical set-up and an implicit smoothing of mortality rates.

To keep the model both parsimonious and flexible, extra parameters for cohort effects will not be included. The Bayesian approach to this model is outlined in detail and a Markov Chain Monte Carlo (MCMC) algorithm for numerical estimation is derived. A corresponding routine in the statistical programme language R is made available. Necessary mathematical and probabilistic preliminaries, especially on Bayesian statistics, MCMC theory and multivariate time series analysis, are provided. The model is calibrated with data of different European countries to assess the quality of mortality predictions. Statistical diagnostic tools and comparisons with univariate projections from individual models and frequentist estimation are conducted in the course of two case studies. Core interest lies in careful discussions on how the model performs regarding the desired properties in stochastic mortality forecasting.

1.4 Outline

The work is outlined as follows. Chapter 2 gives a detailed introduction into stochastic mortality forecasting for multiple populations by first introducing standard terminology on mortality and the general framework of stochastic models even for single populations. By explaining existing models in more detail, this chapter motivates both the general need for new multi-population projection models to address the required objectives and the particular modelling choices in the remainder of the work. In Chapter 3, the model is carefully established and the Bayesian estimation procedure is derived, along with a numerical MCMC algorithm and the computation of mortality forecasts based on these results. The model is applied with several European countries via two case studies in Chapter 4. The model fit and the convergence of the MCMC algorithm are analysed and results are compared to univariate and frequentist mortality forecasting procedures. Chapter 5 concludes with a discussion. The Appendix gives background information for a general understanding of Bayesian statistics, MCMC techniques, and the VECM, as well as further technical details where necessary. It further introduces the R package to run the MCMC algorithm along with numerical details.

2 Literature on Mortality Forecast Modelling

Dating back to the first life tables in the 17^{th} century or the first laws of mortality proposed by de Moivre (1725) and Gompertz (1825), the analysis of mortality has always been in the focus of researchers from diverse scientific fields such as medicine, sociology, and economics. In particular, mortality forecasts are of great importance as they heavily influence socio-economic decisions. Any calculations w.r.t. public pension and health care systems build upon projected mortality rates. A broad literature has emerged during the last decades in which various models and projection tools have been suggested. Since the beginning of the 21^{st} century, multi-population as well as Bayesian models have been proposed by several authors. This chapter gives an overview of the stochastic models from recent years, which have evolved into benchmark models in modern mortality forecasting, and a deep insight into simultaneous analysis of several populations and Bayesian approaches to mortality estimation. After defining central measures of mortality in Section 2.1, the subsequent section describes the milestones in stochastic mortality projections and their extensions in detail. Section 2.3 provides descriptions for multi-population models, and Section 2.4 finally introduces the different applications using the Bayesian paradigm.

2.1 Definitions of Mortality

When it comes to studies on mortality, a variety of quantities exists whose notation and terminology are not always consistent throughout the literature. Besides minor changes in subscripts, this work uses the same standardised notation as in the comparison of several stochastic models by Cairns et al. (2009). It is worth mentioning that variables in any other cited literature should be read carefully in order to avoid confusion with def-

initions. The interested reader is referred to Pitacco et al. (2009) for a general overview of mortality measures.

In terms of estimating and forecasting mortality in a certain population, the central quantity of interest is the *force of mortality*, denoted by μ_{xt}, where $x \geq 0$ is some real-valued age and $t \geq 0$ is a point in time. For fixed x and t, it expresses the instantaneous probability of immediate death for an individual aged exactly x at time t, and as a function in age it contains all information of mortality behaviour in the sample under consideration at time t. The force of mortality corresponds to the hazard function known from survival analysis when x and t increase at the same pace given an initial age x_0 at starting point t_0. Typically, μ_{xt} is not directly observable, since mortality data are not recorded on a continuous scale.

In contrast, another important measure of mortality, the *central death rate*, accounts for deaths during a time frame rather than at an exact point in time. It is defined as

$$m_{xt} := \frac{\text{Expected number of deaths during calendar year } t \text{ aged } \lfloor x \rfloor \text{ last birthday}}{\text{Average population during calendar year } t \text{ aged } \lfloor x \rfloor \text{ last birthday}},$$

where $\lfloor x \rfloor$ is the greatest integer not exceeding $x \geq 0$ and the calendar year $t \in \mathbb{N}$ is meant to be the time interval $[t, t+1)$. The average population size in the denominator is usually estimated by the population size in the middle of the calendar year or, to be more precise, the total time lived in calendar year t by people aged $\lfloor x \rfloor$ at their last birthday. The latter estimate is often referred to as the *exposure-to-risk* E_{xt}. The expected number of deaths is naturally approximated by the observed number of deaths D_{xt} in the population under consideration. The resulting estimate D_{xt}/E_{xt} for the central death rate m_{xt} is called the *crude death rate* for x in t.

Finally, a third quantity is the so-called *mortality rate* q_{xt}, which is the probability of dying within calendar year t for an individual aged exactly x at the point in time t.

In order to obtain results for the force of mortality from discrete observations, it is common to impose the assumption that μ_{xt} remains constant over each year of integer

age and over each calendar year, i.e.

$$\mu_{xt} = \mu_{x+\Delta x, t+\Delta t}$$

for all $0 \leq \Delta x < 1$ and $0 \leq \Delta t < 1$, see, e.g., Cairns et al. (2009) or Pitacco et al. (2009). They derive that the force of mortality equals the corresponding central death rate and, indeed, the ML estimate for μ_{xt} is then the crude death rate. Further, it is shown that the relationship

$$q_{xt} = 1 - \exp(-m_{xt}) \tag{2.1}$$

holds for any integer-valued x and t, what they regard an accurate approximation for the mortality rate. This work adopts the assumption of a constant force of mortality and its implications.

2.2 Stochastic Models for Mortality Projection

Besides a variety of deterministic models which extrapolate historical mortality trends into the future (see, e.g., Pitacco et al. (2009)), stochastic models have become popular for mortality forecasting purposes since the early 1990s. Such approaches share the advantage of including the random nature of mortality through underlying probability assumptions. As a consequence, not only point estimates but also confidence intervals for mortality forecasts can be established. Furthermore, these approaches are designed to fulfil standard criteria in mortality modelling such as consistency with historical data, biologically reasonable long-run dynamics or robustness, as outlined in detail by Cairns et al. (2008). This section describes the two main models within this framework and a selection of extensions to these. For a thorough comparison of stochastic mortality models, the reader is referred to the papers by Booth and Tickle (2008), Cairns et al. (2009) and Haberman and Renshaw (2011).

2.2.1 The Lee-Carter Model

In recent years, the *LC model*, first introduced by Lee and Carter (1992), has evolved into the main approach for mortality estimation and forecasting. This stochastic model

describes the central death rate m_{xt} at some age x by the log-bilinear form

$$\log(m_{xt}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{xt} \tag{2.2}$$

with parameters $\alpha_x, \beta_x, \kappa_t$ and error terms ε_{xt}. As a function in age only, α_x describes the basic pattern of m_{xt} averaged over time. Conversely, κ_t is a function in t and expresses the overall evolution of mortality over time. These changes to the underlying mortality scheme are weighted for the different ages through the profile in β_x. Lee and Carter (1992) set up the constraints $\sum_x \beta_x = 1$ and $\sum_t \kappa_t = 0$ in order to achieve uniqueness in the bilinear term. They finally assume the random fluctuations ε_{xt} to be independent with zero mean and variance $\sigma_\varepsilon^2 > 0$. For annual mortality data, Lee and Carter (1992) estimate the parameters by least-squares, using the first-rank approximation from a singular value decomposition of a suitable matrix to deal with the non-linearity in the parameters. Denoting the estimates by $\widehat{\alpha}_x, \widehat{\beta}_x, \widehat{\kappa}_t$, typically improvements in mortality are detected in that the function $\widehat{\kappa}_t$ exhibits a negative trend. Stochastic forecasts for the development in mortality are obtained by applying time series models with these estimates. Lee and Carter (1992) propose a random walk with drift for modelling $\widehat{\kappa}_t$. Point estimates and confidence intervals are then obtained for the future evolution in mortality by using Box-Jenkins approaches (see, e.g., Box et al. (2013)). As a consequence, substituting the remaining parameter estimates in (2.2) and setting the error terms to zero leads to point and interval projections for the expected log-scaled central death rate in the future.

Since the influential work by Lee and Carter (1992), several extensions to the model in (2.2) have been discussed, one of which is an ML estimation procedure suggested by Wilmoth (1993) and Alho (2000). Due to findings by Brillinger (1986), they argue that for some age x and time t, the according number of deaths D_{xt} is independent and approximately Poisson distributed with mean $m_{xt}E_{xt}$, i.e.

$$D_{xt} \sim \text{Poi}(m_{xt}E_{xt})$$

for all x and t. Then ML estimation becomes possible for $\alpha_x, \beta_x, \kappa_t$ using the relation $\log(m_{xt}) = \alpha_x + \beta_x \kappa_t$, i.e. the LC approach as in (2.2) but dropping the additive error term. Due to the work by Brouhns et al. (2002), who apply this technique to Belgian

mortality data, the approach is usually referred to as the *Poisson log-bilinear model*. By excluding ε_{xt} from the model equation, random fluctuations around the logarithm of the central death rate are no longer assumed homoskedastic. This is regarded advantageous since the variance in observed $\log(m_{xt})$ differs between younger and older ages, where relative variability w.r.t. the exposure-to-risk is the highest for old ages due to low sample sizes. Delwarde et al. (2007) generalise the ML approach by replacing the Poisson distribution with the two-parametric Negative Binomial distribution to account for possible overdispersion in the number of deaths, as do Renshaw and Haberman (2003b,c, 2006) through an overdispersed Poisson formulation.

Another major extension of the LC model is the inclusion of cohort effects as suggested by Renshaw and Haberman (2006). It was discovered that the pure LC model is not able to successfully fit certain datasets such as mortality data from England and Wales, see Renshaw and Haberman (2003a). The goodness-of-fit can be significantly increased when for integer-valued age x and calendar year t the cohort effect of the corresponding birth year $t - x$ is taken into account through an additive term γ_{t-x}, i.e.

$$\log(m_{xt}) = \alpha_x + \beta_x^{(1)}\kappa_t + \beta_x^{(2)}\gamma_{t-x}$$

with some distribution assumption on the log-death rates and additional constraints to the new parameters. The resulting model is sometimes referred to as the *Renshaw-Haberman model*, particularly when comparing to simple age-period-cohort effect models of the form $\log(m_{xt}) = \alpha_x + \kappa_t + \gamma_{t-x}$. Many further modifications of the LC model can be found in the literature – see, e.g., Lee (2000) or de Jong and Tickle (2006) –, but will not be discussed here.

Despite its overall success, the LC framework has also been criticised taking into consideration the model criteria by Cairns et al. (2008). For example, as Cairns et al. (2009, 2011a) point out, the quantification of mortality improvements through only one factor κ_t implies perfect correlation among the changes of central death rates for all ages, and the profile β_x may not be smooth. The correlation structure remains simple even for the Renshaw-Haberman model, for which additional problems with robustness are detected. The shortcomings in the LC model and its extensions have therefore ultimately led to various other attempts to project mortality or death rates.

2.2.2 The Cairns-Blake-Dowd Model

With their study on mortality data of the United Kingdom, Cairns et al. (2006) contribute another major approach to stochastic mortality forecasting in addition to the LC model. Based on empirical findings, in this two-factor model, named *CBD model* after its inventors, the core assumption is that for some fixed time $t \geq 0$ and sufficiently high ages x, the logits of the probabilities q_{xt} increase approximately linearly in age. Therefore, they adapt a Binomial generalised linear regression[1] for each t, i.e.

$$\log\left(\frac{q_{xt}}{1-q_{xt}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)} x, \quad x \geq x_0 \tag{2.3}$$

with x_0 being a lower bound for the ages under consideration. The identifiable parameters $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ are estimated via standard ML methods. Then, similar to the LC approach, the estimates $\widehat{\kappa}_t^{(1)}$ and $\widehat{\kappa}_t^{(2)}$ are regarded as a bivariate stochastic process, which is again modelled using time series techniques. The first marginal time series, $\widehat{\kappa}_t^{(1)}$, gives the intercepts in all regression equations, thereby describing the general development in mortality for all ages. As before, this time series usually declines. The second time series, $\widehat{\kappa}_t^{(2)}$, measures changes in the slopes. When some age groups benefit more than others from improvement in mortality rates, this affects the slope. For example, $\widehat{\kappa}_t^{(2)}$ exhibits an increasing trend if younger age groups show faster reduction in mortality than the older age groups do. Hence, the two-factor approach in (2.3) allows for imperfect correlation in changes of mortality rates as distinct from the LC model. This is also advantageous as the CBD model is able to smooth not only the mortality development over time but also the age profile, which is often neglected in standard LC models. On the other side, forecasting the bivariate time series $\left(\widehat{\kappa}_t^{(1)}, \widehat{\kappa}_t^{(2)}\right)$, which finally leads to projections in the logit of mortality rates, becomes more complex than in the univariate case.

For mortality data of the United States, Cairns et al. (2009) find that the plot of estimated $\log\left(q_{xt}/\left(1-q_{xt}\right)\right)$ against age x reveals some curvature. In such cases they suggest to incorporate a quadratic term in (2.3) along with an own set of parameters

[1] Note that the CBD model is not necessarily a *logistic regression* model as the successes and failures describing the mortality rates q_{xt} might not be directly observable. In particular, in this work only the number of deaths D_{xt} (successes) and exposure-to-risk E_{xt} (number of trials) will be available such that q_{xt} must be further linked to the underlying data through (2.1). The resulting link function $m_{xt} \mapsto \log\left(\exp(m_{xt}) - 1\right)$ differs from the logit link so that the more general terminology of *Binomial generalised linear regression* is used.

$\kappa_t^{(3)}$. Pitacco et al. (2009), however, point out that the behaviour of the third time series may not be clear and modelling becomes complicated. Nonetheless, the flexible set-up of the CBD model indeed allows for further modifications such as additional cohort effects, see Cairns et al. (2009) for a thorough comparison of different designs. The CBD model is generally found to be more adjustable with a solid fit, but is only appropriate for the analysis of high ages.

2.3 Simultaneous Mortality Projections for Several Populations

So far, accuracy of the aforementioned stochastic models and their modifications have been assessed for a wide range of mostly developed countries, commonly stratified for both genders. However, models are usually fitted separately to distinct populations and forecasts from each individual model are finally compared – see, e.g., Macdonald et al. (1998), Tuljapurkar et al. (2000) or Booth et al. (2006). Li and Lee (2005) argue that information provided by the interaction within a group of countries is lost when the mortality of one of these countries is modelled individually. Basic patterns in mortality are expected to be consistent among similar countries, and due to globalisation effects differences should vanish over time. Incorporation of such transnational influences should lead to improvements in projections compared to individual studies. Similar arguments are quoted when males and females are separately modelled, or subpopulations, e.g. members of a pension fund in a certain country, shall be compared to the corresponding parent population. This section summarises contribution in recent literature to the topic of joint analyses of more than one population. It should be noted that there also exist various models for simultaneous estimation of other mortality-related quantities such as life expectancy, see Oeppen and Vaupel (2002) for instance.

2.3.1 The Augmented Common Factor Model

Li and Lee (2005) suggest a three-step procedure to mutually model several populations that builds on the LC approach. First, the ordinary model in (2.2) is run for the entire sample comprising all, say, countries. The parameters κ_t and β_x determine the overall development of mortality and according profile of age-specific weights, respectively.

In a second step, the averaged transnational death rate patterns, given by α_x, are replaced by individual patterns α_{xp} for each country p. Estimates for these profiles are the country-specific logarithms of central death rates averaged over time. Finally, the remaining residuals are described by a second bilinear term $\beta_{xp}\kappa_{pt}$, which now depends on the population via the subscript p and enters the model as an additive term. For fixed p, the corresponding κ_{pt} reveals differences in mortality evolution w.r.t. the overall development, and β_{xp} gives population-specific weights for ages concerning these deviations. By analogy to the standard estimation in the LC model, Li and Lee (2005) apply a singular value decomposition to find least-squares estimates. Summarising, the final model, which they call the *augmented common factor model*, is given by

$$\log(m_{xpt}) = \alpha_{xp} + \beta_x\kappa_t + \beta_{xp}\kappa_{pt} + \varepsilon_{xpt}$$

with independent and homoskedastic error terms ε_{xpt}. The authors restrict the model to those countries, whose differences from the overall mortality tend towards a constant level in the long-run. When the estimates for κ_{pt} cannot be satisfactorily modelled by a random walk without drift or an autoregressive (AR) model, they exclude population p from the analysis. Although it is crucial to avoid divergent behaviour in death rates that does not seem plausible, ignoring certain countries may violate the statistical validity of the model, comparable to omitting unwanted observations such as outliers in statistical models in general. The augmented common factor model is even more inappropriate for modelling a group of countries which show some uncommon death rate patterns. For example, Li and Lee (2005) are not able to capture the observations from Bulgaria, Hungary, and Russia, which is debatable because similar countries like the Czech Republic or Lithuania fit to the approach. Since the estimates $\widehat{\kappa}_{pt}$ must be inspected for each population p separately, the augmented common factor model does not serve as a practical framework for a large number of populations.

2.3.2 Models for Two Populations

Based on findings by Booth et al. (2006) that countries with small population sizes appear more difficult to forecast, Jarner and Kryger (2011) propose analyses of small samples along with large, so-called *reference populations* containing the small sample as a subpopulation. In their model, which they call the *spread adjusted international trend*

model due to application in an international context, they use time series methods to forecast both mortality trends of the reference population and deviations from the overall trend in the subpopulation. These deviations, referred to as the *spread*, are modelled in a regression equation, where explanatory variables comprise linear and quadratic effects of age. The time series of parameter estimates are assumed stationary in order to avoid divergent behaviour in mortality evolution and, in particular, a multivariate first-order VAR model with zero mean is used. In their application to mortality data of 19 developed countries with Denmark serving as subpopulation, Jarner and Kryger (2011) show that the assumption of stationarity is indeed fulfilled, implying that Danish mortality rates converge to those of the reference population in the long-run. However, the authors also note that there is no guarantee for this assumption to hold and there may be need for more suitable models when data suggest permanent variability in differences. In light of the previous section, it is not clear whether the model would lead to satisfactory results when countries like Bulgaria, which cannot be adequately captured in Li and Lee (2005), were included into the analysis. Even if the spread adjusted international trend model can be extended to allow for more than one subpopulation, it may not be a good framework for an analysis of a wider range of countries.

The *gravity model* by Dowd et al. (2011) contributes an approach rather similar to the spread adjusted international trend model. The authors also focus on the case where one population significantly exceeds another in size, and it is assumed that interest lies in modelling the smaller population. This is done with the help of the larger population, because both are again believed to behave similarly for biological and socio-economic reasons and statistical properties gain from the increased sample size. By setting up age-period-cohort models for both populations, respectively, the common trend is obtained via a bivariate time series model for the time-dependent parameters governing the mortality evolution for each population over time. In the equation for the small population's innovations in mortality evolution, the process features an additive term to quantify the difference between the parameters of both populations. The corresponding coefficient measures the effect of this difference in a way that in case of large deviations, the time series for the small population is forced to move to the level of the time series for the large population. This mean reversion for the small population becomes stronger the more the levels of the univariate series differ. The effect is comparable to gravity between a planet and its orbit, thereby giving the model its name. As before, this ap-

proach excludes divergence between the two populations in the design stage and benefits from statistical properties through a larger sample for the originally small population of interest, but – again – particular interest lies in one, say, country only and the time series models need to be assessed carefully. It is noteworthy that this approach implicitly postulates a cointegration relationship between the two univariate processes with a predetermined cointegration vector and a loading factor for the submissive population.

Cairns et al. (2011b) set up a mortality projection model for two populations in which parameters are estimated via Bayesian methods. This framework is also designed for modelling a subpopulation with mean-reverting spreads relative to a dominant reference population. This model is described in more detail in the next section, which is entirely devoted to Bayesian models in mortality projections. Zhou et al. (2012) show that it is not always clear which of the small and large populations is the dominant one. Motivated by the above studies, they extend the two-population approach via application of VAR models in plain and VECM form to include symmetric rather than one-sided cross-correlations into the model equations. Hence, as an advantage over the previous models, there is no specific need for a distinction between dominant and submissive populations. As before, non-divergence conditions are incorporated into the VAR model through parameter constraints and into the VECM through the cointegration term with a pre-specified cointegration relationship. Estimation is conducted via ML and comparison of goodness-of-fit checks indicates that the VECM gives the most reasonable results.

Another approach, which makes use of larger datasets to improve the fit for small populations of interest, is given by Plat (2009). Using a reference population, the model is designed for the specific problem of forecasting the insured amount rather than the pure number of deaths in a portfolio of, say, pension funds, and is therefore not considered. In the context of hedging such insurance portfolios, Li and Hardy (2011) compare several extensions of the LC model for two populations, one of which allows for cointegration in a bivariate time series model for κ_t. Several other two-population models, e.g. by Lin et al. (2013) and Zhou et al. (2013), have been suggested in an actuarial context, but are not discussed in detail here. A general alternative to bivariate mortality projection models is proposed by Ahčan et al. (2014). Here the main motivation is that the small population has an insufficient sample size for statistical analyses. A pool of larger reference populations is mixed in an optimal way to replicate the population of interest, and

this sufficiently large counterpart is then projected via standard models as described in Section 2.2. All the models in this section have in common that only two populations can be forecast and, as seen with most of the applications in the respective studies, their use is restricted to cases where interest lies in a single subpopulation rather than dependencies between several populations.

2.3.3 Further Models

Some other contributions on joint analyses of mortality in several populations have been made. Building upon the work by Zhou et al. (2012), Ntamjokouen et al. (2014) apply the VAR model and the VECM to more than two populations. Their application with both genders in nine Canadian provinces give mixed results in that the VECM seems the most appropriate model, but lacks goodness-of-fit notably for males. Due to the brevity of their discussion, the analysis, however, must be considered insufficient for general conclusions on the appropriateness of the VECM. Biatat and Currie (2010) use P-splines to detect similarities and differences between mortality rates of different countries or between males and females. Their analysis, however, is not dedicated to mortality projections into the future and therefore not described in detail here. In comparison, Börger and Aleksic (2011) do make projections on future mortality trends using a stochastic model on mortality improvements rather than rates. It is assumed that the logarithm of annual changes in mortality rates is given by an additive set of parameters for the variables age, period, and cohort. For distinct populations, the parameters are estimated individually and, for projection purposes, different techniques must be applied to the estimates. The principal component of this approach is the forecast of the period parameters, and since a direct methodology does not seem obvious, Börger and Aleksic (2011) derive future values for these parameters by forecasting the life expectancies of the populations under consideration. The authors argue that such a forecast is generally easier than immediate predictions of the mortality improvements, but this method in turn requires a couple of assumptions on the development of life expectancy, which depend on the populations under consideration and raise difficulties on their own. The model is thus applied to Western European countries with large population sizes only, thereby inhibiting this approach from being a general framework for joint analyses of a wide range of countries.

2.4 Bayesian Models for Mortality Forecasting

The models in the previous sections have in common that they principally build upon a two-level hierarchical structure. First, the main model equation expresses the quantity of interest, i.e. a transformation of the mortality rate, in terms of parameters. Then the estimated parameters in this equation are forecast into the future by standard time series models. Since in each of the discussed approaches both submodels are estimated separately, the link between the response variable and the model's underlying dynamic processes may become spurious. Czado et al. (2005) warn against incoherence that these two-step procedures may account for. Moreover, forecasts based on the frequentist Box-Jenkins approaches in the second model stage typically exhibit elliptical confidence boundaries with stable long-term confidence regions, which appears to be unrealistic given the naturally increasing uncertainty about future developments in medicine, economy, and sociology.

In order to avoid such shortcomings, Cairns et al. (2011b) propose to combine both steps into one estimation procedure, which improves consistency within the set of parameters. They indicate that a likelihood-based single-step estimation method could be applied. However, due to its natural inclusion of uncertainty in the parameters, Cairns et al. (2011b) and also Czado et al. (2005) prefer Bayesian inference over frequentist estimation. As stated by Czado et al. (2005), another reason is that the prior belief of mortality rates behaving smooth across ages and time can be integrated into the model framework. When data support this assumption, estimated mortality rates will be smooth, too, thereby making crucial smoothing methods obsolete. The remainder of this section outlines the literature on Bayesian models in mortality forecasting.

2.4.1 Bayesian Approaches to the Lee-Carter Framework

Czado et al. (2005) are the first authors who apply Bayesian statistics to the framework of the LC model. The underlying model is the Poisson log-bilinear model from Section 2.2.1 first introduced by Brouhns et al. (2002). The authors combine the two stages in the LC approach, i.e. the estimation of period effects along with their age profiles and the calibration of an underlying dynamic model for the period effects, into one single step. Contrary to the general model, they assume a first-order AR model

rather than a random walk with drift for κ_t. As mentioned earlier, in combination with the Bayesian methodology this allows for the advantage of smoothing the data within the estimation procedure. Finally, standard priors are used for the unknown hyper-parameters in the AR(1) model. In their application to data of French males, Czado et al. (2005) show that their mortality projections are somewhat more pessimistic, but in general they closely agree with frequentist forecasts. Parameter uncertainty for future forecasts leads to wider credibility bands, which always include the frequentist estimates.

Similar conclusions can be found for the example of US-American males in Pedroza (2006). Here, the Bayesian paradigm is directly applied with the original LC model, i.e. with normally distributed error terms and a random walk with drift for the dynamics process. The findings of wider credibility intervals in both studies suggest that prediction errors in the frequentist version of the LC model are not able to cover all sources of uncertainty. Lee and Carter (1992) themselves suppose in their original work that the fluctuation in the dynamics process accounts for most uncertainty in the final confidence bands. Hence, due to Pedroza (2006), Bayesian models are more suitable in mortality projection as they include all different sources of estimation and prediction errors. Finally, Kogure et al. (2009) compare both models by Czado et al. (2005) and Pedroza (2006) and some variations via application to data of Japanese males. They conclude that in each model framework, a time series with stochastic trend for κ_t performs best, whereas no appropriate results for the differences between the normality and Poisson assumptions can be drawn.

Reichmuth and Sarferaz (2008) provide another Bayesian mortality projection model. Also building on the LC approach, they modify the original model extensively in that they allow for several covariates, e.g. macroeconomic quantities, in addition to the latent variables κ_t. Furthermore, they employ time series models not only for the time-dependent parameters but also for the age parameters to achieve smoothness across ages. With their application to mortality for males in the United States, they conclude that covariates can improve forecasts. As seen in the previous studies, Reichmuth and Sarferaz (2008) also stress the Bayesian property of incorporating all different sources of prediction error. Likewise, the model by Girosi and King (2008) is worth mentioning as it also includes covariates in a Bayesian model. However, the methodology strongly differs from what Lee and Carter (1992) propose, and mortality rates are analysed by

cause of death. Therefore, this model as well as other approaches with cause-specific parameters in medical contexts, see e.g. Bray (2002), are not discussed in detail here.

2.4.2 The Bayesian Mortality Model for Two Populations

The *Bayesian mortality model for two populations* by Cairns et al. (2011b) is the first approach that combines both Bayesian methodology and simultaneous estimation procedures for more than one population. The model is restricted to two populations, which may be distinct or nested. The authors make use of a simple age-period-cohort model, i.e. an additive model with own parameters for each of the age, period, and cohort effects, to keep the focus on the Bayesian approach. Similar to Jarner and Kryger (2011) and Dowd et al. (2011), there is a reference population which is modelled first, where for the other population the spread in central death rates is analysed. The desired behaviour of non-divergence in the long-run between both populations is accounted for by using mean-reverting processes for the underlying dynamics. The set of such models ranges from random walks to AR models of up to second order. Note that the effect on sudden shocks or short-term estrangements has not become clear yet when, instead of data-driven techniques, convergence is postulated in the model's design stage, as done in this and other previously mentioned approaches.

Even if Cairns et al. (2011b) apply mostly non-informative priors, the large number of parameters requires various distribution families, including the Inverse Wishart, Beta, Gamma, and Gumbel distributions. According to the authors, the model therefore remains sensitive towards the assumptions in the prior distributions and results must be treated carefully. Apart from this, Cairns et al. (2011b), however, show that the Bayesian methodology strongly helps in estimating different populations jointly, smoothing the fitted data, and being consistent in the projections. This underlines the appropriateness of Bayesian approaches in mortality forecasting of more than one population.

Up to this point, as far as I am aware, there is no Bayesian approach for more than two populations in the framework of stochastic mortality models described in this chapter. It is worth noticing that Raftery et al. (2013) introduce Bayesian estimation in a stochastic extension of the so far deterministic mortality projection model by the United Nations. Using a hierarchical model on innovations in life expectancy, parameters are country-

specific but follow a common distribution with global hyperparameters, i.e. inference is based on the aggregate of all information. The Bayesian hierarchy guarantees a certain degree of coherence, but this approach is far less concerned about common trends or dependencies between countries because forecasts are still made individually without any quantification of correlation.

3 The Bayesian Multi-Population Mortality Projection Model

As seen in the previous chapter, many achievements have been made in multi-population mortality projection modelling within the last years. As a consequence, different models allow for a rich set of possibilities in mortality forecasting. Nonetheless, the discussion has revealed that, in practice, each model has its own limitations. In order to overcome these shortcomings, a new model is proposed in this chapter. Based on findings in the literature review in Chapter 2, it features the Bayesian paradigm to capture both parameter uncertainty and all sources of prediction errors. Furthermore, the focus is laid on a flexible framework using the CBD model, allowing for an arbitrary number and an arbitrary selection of populations. Accordingly, in the remainder of this thesis, the suggested model is referred to as the *Bayesian Multi-Population Mortality Projection (BMPMP) model*. It should be stressed that – as far as I am aware – no such model approach is found in literature, i.e. the work in this thesis provides substantial contribution to stochastic mortality forecasting. In particular, for the first time, the CBD model is applied in the multi-population framework, and the Bayesian paradigm has not yet been used for the CBD model or a mortality model with more than two populations either. This chapter describes the BMPMP model in full detail and is organised as follows. The first section addresses the targets for the new model, motivated by a summary of shortcomings in established approaches. The actual model itself is defined and explained in detail in Section 3.2. Next, Bayesian estimation and forecasting of the model are presented in Sections 3.3 and 3.4, respectively. Section 3.5 summarises the BMPMP model.

3.1 Model Targets

The main objective of the BMPMP model is to stochastically forecast mortality rates for high ages of an arbitrary number of populations. It must provide biologically plausible joint predictions for flexible selections of populations. Uncertainty of future mortality rates and their inter-dependencies shall be coherently quantified through probability distributions. The BMPMP model addresses its main targets through a combination of different approaches in model design, estimation, and forecasting.

3.1.1 Limitations in Existing Models

To fulfil these targets, a new model is necessary. Clearly, both the LC and CBD frameworks as well as age-period-cohort models have been successfully proven to be state-of-the-art in stochastic mortality forecasting. While the first two approaches bring the advantage of interaction between time-dependent improvements in mortality and their age-dependent weights – and, hence, a principally better fit to historical data over age-period-cohort models –, it is not necessarily clear which one of the LC and CBD models is more suitable. A major disadvantage in the LC methodology is a possible lack in the model fit through an implicit perfect correlation in mortality improvements among different ages. Additionally, the lack of smoothness in the age profiles and their non-identifiability w.r.t. time-dependent quantities may cause undesired results. The CBD model addresses all these issues through its formulation as a generalised linear regression, but is only applicable with ages above a minimum threshold of at least 40 years. Problems with a possible lack of fit in CBD models can be reduced by including a quadratic age effect on mortality rates. The necessity of forecasting the additional coefficient time series will not be an extra burden in an already multivariate model, when several populations need to be analysed. If the modeller is interested in high ages only, which is a reasonable assumption in the context of many life insurance or other financial products, the CBD model serves as solid foundation in stochastic mortality projections.

Up to this point, the discussion has focussed on the fundamental properties of the aforementioned mortality projection models w.r.t. the performance in their intended scope of application: forecasting a single population. However, the introduction of this work stressed the increasing necessity to jointly forecast different populations to

capture common trends and dependencies. Chapter 2 reviewed the extensions of the LC framework and other approaches to address this task of multi-population forecasts. However, it turned out that joint prediction models are far less elaborate. Indeed, existing models for projections of several populations are limited by at least one of the following:

- The model is only applicable to a limited number of populations.
- The model design requires determination of dominant and sub-populations.
- Its assumptions exclude certain populations or combinations of populations, e.g. due to unwanted interaction behaviour in the long-run or the need for a parent population for at least one of the other populations.
- Underlying dynamics processes must be analysed individually, preventing the framework from universal usage.
- Model forecasts have stable long-run confidence boundaries, which are unreasonable considering the increase in uncertainty and risks of regime changes in the future.
- Desired biological plausibility in forecasts is hypothesised in the model design and does not necessarily stem from historical data.
- Populations to be modelled must have a high quality in mortality data and the model may require further explanatory covariates.
- The model requires forecasts of other mortality quantities, which are themselves not easy to handle with.
- The model is not designed for mortality projections.

Based on these points, no such sophisticated tool as in the univariate case to forecast an arbitrary number of possibly heterogeneous populations exists. Consequently, this motivates the BMPMP model. The subsequent section is then devoted to how the model is designed to provide the desired framework in multi-population forecasting, which shall overcome the aforementioned shortcomings.

3.1.2 Benefits of the New Model

Through the flexibility of a generalised linear regression design known from the CBD framework, the BMPMP model allows to jointly forecast an arbitrary number of more than two populations at once. Differences between the distinct populations are accounted for by specific main and interaction effects in the linear predictor. This distinguishes the new approach from the vast majority of projection models, as they are designed to project the spread between only two populations. The Binomial generalised linear regression further allows for simple interpretation and adjustments of the main model equation. The model is only applicable with ages exceeding a minimum threshold of 40 or so years, but has the advantage of immediate smoothing across ages. For an improved model fit, the numbers of deaths are assumed to follow a Poisson distribution.

It is somewhat difficult to collect consistent explanatory variables on all populations, particularly when they are not national but portfolio-specific. It is further non-trivial to project such covariates into the future, which means that such approach cannot be considered universal frameworks. Similar arguments apply to models for which forecasts of other mortality quantities have to be made. By contrast, the BMPMP model is flexible regarding the requirements w.r.t. the available data. For any population to be included in the model, it suffices to have data on the observed number of deaths and estimates for the exposure-to-risk for all ages and calendar years under consideration.

The BMPMP model further employs a VECM for the projection of the underlying dynamics processes. Through its cross-correlation, cointegration, and AR terms, this data-driven time series model is designed to capture dependencies even between less connected populations. In contrast to the augmented common factor model by Li and Lee (2005), which too allows for an unlimited number of populations, the BMPMP model intends to overcome the problems seen for certain Eastern European countries, whose long-run behaviour could not be reasonably forecast into the future. Not only must Li and Lee (2005) exclude certain countries from their analysis, but they further need to specifically investigate the marginal time series for all individual countries. The VECM is more suitable for mortality projection of multiple arbitrary populations, because the multivariate set-up makes such individual analyses obsolete.

Following the arguments in many of the cited studies concerning multi-population forecasting, unwanted long-run interactions are given when mortality rates of different populations diverge over time. However, even if it is biologically plausible that mortality rates converge or keep constant differences over time, some divergent behaviour, at least in the short-run, has been observed for several countries in recent years, e.g. in Eastern Europe in the early 1990s. Furthermore, diseases, natural catastrophes, and wars on the one hand, as well as medical innovations and improvement on the other, have always caused random shocks with long-term effects on mortality in certain populations. The multivariate VECM approach allows for random shocks via Gaussian noise while maintaining a non-divergent long-run behaviour through inclusion of an error correction term. This accounts for cointegration between the univariate time series, i.e. linear combinations of several populations forming stationary processes, whose information is lost in less data-driven techniques. Notably, with a deterministic non-divergence hypothesis, as used in other cited approaches, random shocks might be systematically discounted. Estimation of the error correction term is driven by the data without formulating any restrictive hypotheses in order to achieve biological plausibility. Finally, the VECM is also motivated by Zhou et al. (2012), as it avoids any pre-determination of dominant populations.

Motivated by the studies described in Section 2.4, which have proved successful in capturing a variety of systematic and unsystematic risks, e.g. the purely random nature as well as parameter and prediction uncertainty, the BMPMP model is estimated via Bayesian statistics. This technique as such makes the approach less sensitive towards deviations from the Poisson assumption on the number of deaths through possible overdispersion or omission of cohort effects. A likelihood-based frequentist procedure would again require a two-step estimation method, where the CBD model is estimated first, and then the VECM is applied with the resulting parameters. This study refrains from this approach, because frequentist parameter estimates are expected less coherent.

The quantification of future uncertainty in the frequentist LC and CBD frameworks can additionally be criticised for their typically elliptical behaviour, meaning that after a short time of increasing variability, the confidence boundaries remain stable around the best estimates. These characteristics are not desirable, as uncertainty w.r.t. future projection should naturally increase continuously with time to express the diminishing impact of today's and past developments on the far-away future. It will be apparent

from the case studies in Chapter 4 that the cointegration term in the BMPMP model introduces more reasonable credibility bands of linearly increasing uncertainty.

As a final note, a principal advantage of Bayesian estimation is its ability to handle large datasets and to deliver solid distributions on global parameters if models are not parsimonious or even over-parametrised. Due to the high-dimensionality in multi-population mortality modelling, the BMPMP model would be restricted to a very small and hence undesired number of populations with frequentist approaches, even for a relatively large history of observed mortality data.

3.2 Model Specification

By analogy to other stochastic mortality prediction models, the BMPMP model consists of two hierarchical submodels. First, for each calendar year, the state variable is modelled, which in this instance is done via the CBD approach. Following, the second step is the projection of obtained model parameters into the future via the VECM. This section fully describes and motivates both specifications in the model.

3.2.1 The Cairns-Blake-Dowd Approach

In order to employ the CBD model on the underlying state variables in a multi-population context, the linear predictor in the regression equation of the original approach must be amended. To begin with, the CBD model with quadratic term for the age effect and minimum age x_0 is considered, i.e.

$$\log\left(\frac{q_{xt}}{1-q_{xt}}\right) = \kappa_t^1 + \kappa_t^2(x-\bar{x}) + \kappa_t^3((x-\bar{x})^2 - \widehat{\sigma}^2), \quad x \geq x_0, \tag{3.1}$$

where Cairns et al. (2009) motivate the quadratic pattern by findings on U.S. data. Here, the age x is centred by the average $\bar{x} = n_a^{-1}\sum_i x_i$, where n_a is the total number of different ages x_i under consideration. Similarly, $\widehat{\sigma}^2 = n_a^{-1}\sum_i (x_i - \bar{x})^2$ is the variance of the ages. It is worth mentioning that for the sake of simplicity and flexibility, the model does not include additional parameters for, say, cohort effects as proposed by Cairns et al. (2009). If deviations caused by cohort effects systematically exceed the pure random noise and cannot be captured by the quantification of different sources of

uncertainty in the Bayesian methodology, an extension of the model, which does include explicit cohort parameters, is of course possible. Regarding the minimum age x_0, this value should generally not fall below the age of 40. For lower age groups, the linearity assumption on the logits of mortality rates does generally not hold. In particular, the CBD approach is not designed to fit the so-called *accident bump*. However, for many applications in insurance science, the analysis of mortality rates for people aged 40 or more is sufficient, since life or health insurance contracts are usually concluded just before that age. The CBD framework is often applied with values like $x_0 = 50$ or $x_0 = 60$ for a better performance, see Cairns et al. (2009) for instance.

So far, for a fixed calendar year t, the according mortality rate only depends on age through a linear and quadratic term. The BMPMP model additionally assumes that both the intercept and age effects vary between different populations. Mathematically, these amendments are integrated into the Binomial regression via main effects on the one, and interactions with the age terms on the other hand. Let $n_p \in \mathbb{N}$ be the number of different populations for which mortality projections shall be derived. Taking into consideration the model targets, one assumes without loss of generality (w.l.o.g.) that $n_p > 1$. For some population p with specific mortality rate q_{xpt} in calendar years t, model (3.1) is then extended as follows:

$$\begin{aligned}\log\left(\frac{q_{xpt}}{1-q_{xpt}}\right) &= \kappa_t^0 + \kappa_t^p + \left(\kappa_t^x + \kappa_t^{xp}\right)(x-\bar{x}) \\ &\quad + \left(\kappa_t^{x^2} + \kappa_t^{x^2 p}\right)\left((x-\bar{x})^2 - \widehat{\sigma}^2\right), \quad x \geq x_0. \end{aligned} \tag{3.2}$$

Notation has slightly changed in order to achieve consistent subscripts for the parameters. As x refers to age in this study, the former parameters $\kappa_t^1, \kappa_t^2, \kappa_t^3$ are denoted κ_t^0 for the intercept and $\kappa_t^x, \kappa_t^{x^2}$ for the linear and quadratic age terms, respectively. The new population parameters are consistently referred to through the subscript p. The main effect for population p is κ_t^p, and $\kappa_t^{xp}, \kappa_t^{x^2 p}$ are the interactions with the linear and quadratic terms for age, respectively.

Since there are parameters for each population, a reference population is needed to avoid identifiability problems. Typically, in most statistical models, one certain population acts as reference by setting the according parameters to zero. Similar to the

LC approach, one can also set some constraints, e.g. $\sum_{p=1}^{n_p} \kappa_t^p = 0$ for all t. However, note that this formula implies that all populations are equally weighted, while parameters should be weighted according to population sizes when realistic weights are desired. As an alternative, the BMPMP model uses the overall sample p^*, inferred from the individual populations by simple addition of death and exposure counts, as *reference population*. Consequently, the population parameters measure the differences from average mortality. The BMPMP model then becomes easily interpretable and, for a large number of populations, generally more robust due to the borrowing-strength principle. With the reference population being the overall sample, the number of parameters increases to $3\,(n_p + 1)$. It is worth mentioning that the CBD equation can be interpreted hierarchically. For the reference population p^*, it holds that

$$\log\left(\frac{q_{xp^*t}}{1-q_{xp^*t}}\right) = \kappa_t^0 + \kappa_t^x(x-\bar{x}) + \kappa_t^{x^2}\left((x-\bar{x})^2 - \widehat{\sigma}^2\right), \quad x \geq x_0, \tag{3.3}$$

and, therefore, the intercept and age parameters can be estimated directly from the sum of all populations. In a second step, equation (3.2) is applied with each individual population, given the parameters from (3.3). Throughout this work, the model is set up by this hierarchical structure.

Model (3.2) is a general framework for n_p different populations that do not overlap. Even if the model could be applied with populations that are nested, one should then adjust the linear predictor to avoid incoherence through collinearity. A typical example, which is discussed in the case studies in Chapter 4, is the comparison of males and females in several countries. Denoting by p a particular country, for which the term *population* is adopted, the new dimension of two genders must be integrated separately, e.g. through a subscript g. When a main effect and first-order interactions with both the age and country effects are included, the model becomes

$$\begin{aligned}\log\left(\frac{q_{xpgt}}{1-q_{xpgt}}\right) &= \kappa_t^0 + \kappa_t^p + \kappa_t^g + \kappa_t^{pg} + \left(\kappa_t^x + \kappa_t^{xp} + \kappa_t^{xg}\right)(x-\bar{x}) \\ &\quad + \left(\kappa_t^{x^2} + \kappa_t^{x^2p} + \kappa_t^{x^2g}\right)\left((x-\bar{x})^2 - \widehat{\sigma}^2\right), \quad x \geq x_0,\end{aligned} \tag{3.4}$$

where new parameters are labelled by g and their interpretation is straightforward. With subgroups, the overall samples of both genders are considered first, i.e. the main

gender and gender-age interaction effects are estimated from the reference population. Consequently, interactions of gender with individual populations follow in the second step along with the population-specific main and age effects. With n_p populations, the number of parameters is then $4n_p + 6$. For a general number of subgroups within the populations or an alternative consideration of interaction effects, the well-known generalised linear regression approach allows for easy adjustments.

3.2.2 The Vector Error Correction Model

The aim of the second stage in the BMPMP model is to forecast the parameters from the CBD model w.r.t. time. In the standard version of the CBD model as in (3.1), one obtains a three-dimensional time series, in which univariate time series contain the intercepts and both linear and quadratic age effects. The model is usually forecast via *Box-Jenkins approaches*, e.g. by applying VAR models or, for the sake of convenience, univariate time series models to each marginal parameter type. However, due to the extensively increased number of parameters in the BMPMP model, special care must be devoted to the choice of the time series model for forecasts. The general framework of standard Vector Autoregressive Moving Average (VARMA) models in the Box-Jenkins methodology still remains one of the most reasonable choices due to *Wold's Decomposition Theorem*, as reviewed in Appendix C.1.

Based on findings of mortality forecast models in the literature, it can be expected that the intercept and main age effects show a trend-stationary behaviour over time. Assuming that all other population- or subgroup-specific parameters are not of higher integration order, it seems plausible to concentrate on multivariate time series models for the first differences. Motivated by fundamental results in multivariate time series given in Appendix C, under regular conditions such as non-explosive behaviour, a finite VAR representation without moving average (MA) terms is a reasonable approach. However, first-order differencing in VAR models is substantially different from what is known from the univariate case of AR models. Starting with a VAR model of lag order k, which includes non-stationary time series, the *reverse characteristic polynomial* for unit root detection is now a matrix-valued function. Whereas in the univariate case, the number of unit roots reveals the order of integration, the multivariate case allows the marginal time series to have integration order strictly less than the number of unit roots. In this

case, the unsophisticated technique of marginally differencing the univariate time series of the CBD parameters can distort possible stationarity of long-run relationships among the marginals. As explained in Appendix C.2 in more detail, marginal time series may not be stationary, although linear combinations of them can still be. Such time series are said to be *cointegrated*. In particular, combinations of the intercept and the main effect for females or between different populations-specific parameters are expected to have a stable equilibrium. In Appendix C.3, it is shown that information on cointegration in a VAR model carries over to a singular matrix, which is obtained by evaluation of the characteristic polynomial at 1. It is derived how differencing the vector-valued time series indeed leaves this matrix – which will later be denoted as Π – as an additional parameter in the equation for first differences, in contrast to loosing this information when marginally differencing the individual time series. The resulting model of correct first differences is the VECM, and starting with this representation of a VAR model is always a valuable approach, as described in Appendix C.3. The analysis of cointegrated VAR models goes back to a series of pioneering papers, starting with the work by Granger (1981), and further exploration by Engle and Granger (1987) with much contribution from several authors in the years thereafter. Most notably, Johansen (1988, 1991) and Johansen and Juselius (1990, 1992) develop an ML estimation framework, referred to as the *Johansen procedure*, which is widely used nowadays. A thorough overview on the VECM can be found in Johansen (1995) and Lütkepohl (2007), for instance. Appendix C reviews all necessary preliminaries.

Mathematically, the VECM can be defined in different but equivalent formulations. In the following, the *transitory version* of the cointegrated VAR is described, where for the alternative *long-run specification*, the reader is referred to the appendix or the literature cited above. Let $\mathcal{K}$ be the multivariate time series of all unknown parameters with values $\kappa_t = (\kappa_t^0, \kappa_t^x, \kappa_t^{p_1}, \kappa_t^{p_2}, \dots)$ for $t = 1, \dots, T$, with T being the number of calendar years for which CBD models are run. With $m \in \mathbb{N}$ one denotes the dimension of the time series, which automatically coincides with the number of parameters in each CBD model from the previous section. The according time series of first-order differences $\kappa_t - \kappa_{t-1}$ is of length $T - 1$ and denoted by $\Delta\kappa_t$. The VECM of order $k \in \mathbb{N}$ with initial values

$\kappa_1, \ldots, \kappa_k$ is

$$\Delta\kappa_t = \phi D_t + \sum_{i=1}^{k-1} \Gamma_i \Delta\kappa_{t-i} + \alpha\beta'\kappa_{t-1} + \varepsilon_t, \quad t = k+1, \ldots, T, \tag{3.5}$$

or, equivalently,

$$\kappa_t = \phi D_t + \sum_{i=1}^{k-1} \Gamma_i \Delta\kappa_{t-i} + (I_m + \alpha\beta')\,\kappa_{t-1} + \varepsilon_t, \quad t = k+1, \ldots, T, \tag{3.6}$$

with I_m being the $m \times m$ identity matrix. The error terms ε_t are independent and identically distributed (iid) multivariate normal distributed with zero mean and covariance matrix $\Omega \in \mathbb{R}^{m\times m}$. A vector of time-varying constants D_t with some fixed dimension $d \in \mathbb{N}$ enables the user to include deterministic, e.g. linear or seasonal, trends. This influence is measured by the according parameter matrix $\phi \in \mathbb{R}^{m\times d}$. For the BMPMP model, the VECM with time-consistent $D_t = 1$ and $d = 1$ is applied. Further parameters are the $k-1$ AR coefficient matrices $\Gamma_i \in \mathbb{R}^{m\times m}$, which describe the impact of recent changes in κ on the current difference. Adjustment of k obviously leads to different time horizons, and, in particular, the VECM of order 2 is second-order Markovian and excludes any history before the previous change in the time series. Finally, the cointegration term is of special interest, as the current change $\Delta\kappa_t$ also depends on the according starting point κ_{t-1} through the parameter matrix $\Pi := \alpha\beta' \in \mathbb{R}^{m\times m}$, which is assumed to have rank $r \in \{0, 1, \ldots, m\}$. The matrix Π is decomposed into the matrices α and β, which are both of dimension $m \times r$ with full rank r. For β, the upper $r \times r$ block matrix is assumed to be the identity matrix such that a unique representation of Π is obtained. The decomposition of Π allows for the following interpretations. First, $\beta'\kappa_{t-1}$ represents a new r-dimensional time series containing different linear combinations of the univariate time series in κ. The VECM assumes that these univariate time series are stationary; therefore, the original time series is said to be cointegrated of rank r. Second, the parameters in α then explain to what extent these derivations of the current level in κ govern $\Delta\kappa_t$. It is worth mentioning that, similar to the order k, the cointegration rank r is generally not known and must be additionally estimated.

It is convenient to formulate the model in (3.5) with $D_t = 1$ and $d = 1$ in the compact matrix form

$$\Delta K = \Phi + \Gamma K_2 + \alpha\beta' K_1 + \varepsilon \tag{3.7}$$

with $m \times T - k$ matrices $\Phi = (\phi, \ldots, \phi)$, $\Delta K = (\Delta\kappa_{k+1}, \ldots, \Delta\kappa_T)$, $K_1 = (\kappa_k, \ldots, \kappa_{T-1})$, $\varepsilon = (\varepsilon_{k+1}, \ldots, \varepsilon_T)$, an $m \times (k-1)m$ matrix $\Gamma = (\Gamma_1, \ldots, \Gamma_{k-1})$, and a $(k-1)m \times T - k$ matrix

$$K_2 = \left(\left(\Delta\kappa'_k, \ldots, \Delta\kappa'_2\right)', \left(\Delta\kappa'_{k+1}, \ldots, \Delta\kappa'_3\right)', \ldots, \left(\Delta\kappa'_{T-1}, \ldots, \Delta\kappa'_{T-k+1}\right)' \right).$$

In this representation, the t-th columns of both matrices on the left- and right-hand side belong to the vectorised form of the VECM at time t. Model equation (3.7) is less interpretable than the previous one; however, it is more useful w.r.t. theoretical results in Bayesian statistics. The set of underlying hyperparameters $\{\phi, \Gamma, \alpha, \beta, \Omega\}$ is abbreviated by $\mathcal{H}$ for convenience. The decomposition of β is written as $\beta = (I_r, \beta_l')'$ with the lower block matrix $\beta_l \in \mathbb{R}^{(m-r)\times r}$. Under this so-called *linear normalisation*, one can write $\beta = c + c_\perp \beta_l$ with $c = (I_r, 0_{r\times m-r})' \in \mathbb{R}^{m\times r}$ and $c_\perp = (0_{m-r\times r}, I_{m-r})' \in \mathbb{R}^{m\times(m-r)}$.

Summarising, models (3.4) and (3.5) establish the BMPMP approach introduced in this study. Bayesian methods are used to estimate all parameters in both models; they will be explained in the following section. The only exceptions are the lag order and the cointegration rank, which are assumed to be known beforehand. The according variables k and r are left open such that the analyst can specify the quality fit through two single adjustment parameters.

3.3 Bayesian Estimation

Similar to other approaches in mortality forecasting, the BMPMP model consists of two model equations. In the first stage, mortality data are modelled separately for each calendar year, where the second stage introduces the time dimension in that parameters are regarded realisations from stochastic processes. Whereas ML methods usually estimate the parameters subsequently, the Bayesian approach allows for a one-step estimation method, which has positive impact on coherence between parameters. Different

sources of risk, due to pure randomness, estimation errors, or future regime changes, are specifically accounted for by the likelihood, prior distributions, and posterior predictive methods, respectively. This section describes the Bayesian estimation procedure for the BMPMP model in detail. First, the likelihood for the underlying mortality data as well as the prior distributions for all parameters of the previous section are determined. Then the posterior distributions with according simulation algorithms are derived. For this section, the reader is expected to bring a broad understanding of principals in Bayesian methodology. Otherwise, Appendix A provides a sufficient introduction to this topic based on the standard textbook by Gelman et al. (2013).

3.3.1 Likelihood for the Underlying Data

Following Appendix A, *Bayes' Theorem* plays a central role in Bayesian estimation. Prior distributions for model parameters are updated to posterior distributions through information of the underlying observations, which in this context are the number of deaths and according exposure-to-risk for all ages, populations, and calendar years under consideration. The typical notation from the previous chapter is expanded w.r.t. the additional population dimension, i.e. as D_{xpt} and E_{xpt} are denoted the number of deaths and the exposure-to-risk for age x, population p, and calendar year t. Due to the common assumption of a constant force of mortality as in Cairns et al. (2009), it is then assumed that

$$D_{xpt} \sim \text{Poi}\left(m_{xpt} E_{xpt}\right),$$

thereby following the arguments of Czado et al. (2005) that a Poisson distribution best expresses the natural mortality behaviour. From (2.1) it follows that

$$D_{xpt} \sim \text{Poi}\left(-\log\left(1 - q_{xpt}\right) E_{xpt}\right). \tag{3.8}$$

As usual, the numbers of deaths are furthermore assumed independent between different ages, populations, and calendar years. It is worth mentioning that (3.8) is additionally assumed to hold for the reference population p^* with some sample mortality rate q_{xp^*t}, number of deaths $D_{xp^*t} = \sum_{p \neq p^*} D_{xpt}$ and exposure-to-risk $E_{xp^*t} = \sum_{p \neq p^*} E_{xpt}$. Independence is still assumed to hold for different ages and calendar years; however, there is clearly a strong dependence on the individual subpopulations. The mortality rate q_{xpt}

is the response variable in the CBD model, and some algebra gives that

$$q_{xpt} = \frac{\exp(\eta_{xpt})}{1 + \exp(\eta_{xpt})},$$

where η_{xpt} denotes the linear predictor, i.e. the right-hand side of equation (3.2). Substituting this expression into (3.8) reveals

$$D_{xpt} \sim \text{Poi}\left(\log\left(1 + \exp(\eta_{xpt})\right) E_{xpt}\right)$$

and leads to the *likelihood*

$$\begin{aligned}
&L\left(\mathcal{K} \mid D_{xpt}, E_{xpt}\right) \\
&= P\left(D_{xpt} \mid \mathcal{K}, E_{xpt}\right) \\
&= \frac{1}{D_{xpt}!}\left(m_{xpt}E_{xpt}\right)^{D_{xpt}} \exp\left(-m_{xpt}E_{xpt}\right) \\
&= \frac{E_{xpt}^{D_{xpt}}}{D_{xpt}!}\left[\log\left(1 + \exp(\eta_{xpt})\right)\right]^{D_{xpt}} \exp\left(-\log\left(1 + \exp(\eta_{xpt})\right) E_{xpt}\right) \\
&\propto \left[\log\left(1 + \exp(\eta_{xpt})\right)\right]^{D_{xpt}} \left(1 + \exp\left(\eta_{xpt}\right)\right)^{-E_{xpt}},
\end{aligned}$$

where L indeed depends on $\mathcal{K}$ through the linear predictor η_{xpt}. Let now

$$\begin{aligned}
\mathcal{D} &= \{D_{xpt} \;\forall\; x, p, t,\; p \neq p^*\}, \\
\mathcal{E} &= \{E_{xpt} \;\forall\; x, p, t,\; p \neq p^*\}
\end{aligned}$$

be the sets of all individually observed data for the numbers of deaths and exposure-to-risk, respectively. Since all D_{xpt} in $\mathcal{D}$ are mutually independent, one can also write more compactly

$$\begin{aligned}
&L\left(\mathcal{K} \mid \mathcal{D}, \mathcal{E}\right) \\
&= P\left(\mathcal{D} \mid \mathcal{K}, \mathcal{E}\right) \\
&= \prod_x \prod_p \prod_t P\left(D_{xpt} \mid \mathcal{K}, E_{xpt}\right) \\
&\propto \prod_x \prod_p \prod_t \left(\left[\log\left(1 + \exp(\eta_{xpt})\right)\right]^{D_{xpt}} \left(1 + \exp\left(\eta_{xpt}\right)\right)^{-E_{xpt}}\right).
\end{aligned}$$

3.3.2 Prior Distributions

Apart from the first k calendar years, *prior distributions* for the parameters $\mathcal{K}$ in the CBD equations are iteratively given by the dynamics process, i.e.

$$\kappa_t \mid \kappa_{t-k}, \ldots, \kappa_{t-1}, \mathcal{H}, k, r \sim N_m \left(\phi + \sum_{i=1}^{k-1} \Gamma_i \Delta\kappa_{t-i} + \left(I_m + \alpha\beta'\right) \kappa_{t-1}, \Omega \right)$$

for $t = k+1, \ldots, T$, where N_m denotes the m-variate normal distribution. The parameters $\kappa_1, \ldots, \kappa_k$ require own prior distributions, as there are not sufficiently many preceding values available for the AR part. For all values but the first, a possible choice would be the above normal prior based on the VECM representation restricted to the available history. However, the missing information on history in these cases alter the conditional assumptions and interpretation, which may introduce bias. To avoid such problems, for the first k values, an m-variate normal distribution with fixed mean vectors μ_t and $m \times m$ covariance matrices Σ_t, respectively, i.e.

$$\kappa_t \mid k \sim N_m \left(\mu_t, \Sigma_t\right), \quad t = 1, \ldots, k, \tag{3.9}$$

is suggested. Ideally, these moments are determined based on prior beliefs or experience. If this is not possible, mean and variance can be estimated from the underlying data. Due to similar magnitudes in variability, it is reasonable to use one constant covariance matrix $\Sigma = \Sigma_1 = \cdots = \Sigma_k$. However, even if such a so-called *empirical Bayes approach* is practicable, it should be noted that, in general, data should not influence the prior assumptions on the distributions of parameters due to the Bayesian paradigm.

Regarding the hyperparameters $\phi, \Gamma, \alpha, \beta, \Omega$ of the underlying dynamics process, priors for the VECM are needed. Since the cointegration parameters α and β multiplicatively affect one another, Bayesian analysis is not straightforward. Several approaches have been suggested since the 1990s, which aim to overcome possible problems with inconsistency as well as local and global identification issues. Particular attention must be given to the choice of the prior for the cointegration term, as supposedly non-informative priors turn out to distribute probability mass in an unreasonable way in the so-called *cointegration space*, i.e. the space spanned by the columns of β. However, a well-defined uniform prior over this space is introduced via the *Grassman approach* due to Villani (2005). The

BMPMP model incorporates the prior by Warne (2006), a slightly generalised version of the standard Grassman prior by Villani (2005). The remainder of this section is devoted to the specific formulation of this prior and its marginal distributions in the context of the BMPMP model. Appendix C.5 provides a detailed discussion of Bayesian techniques and their challenges in the context of the VECM. It particularly motivates the Grassman approach through a discussion of theoretical background on cointegration spaces. For an even more general overview on different Bayesian approaches to cointegration and the Grassman approach, the reader is referred to Koop et al. (2006) and Villani (2005), respectively.

The general *reference prior* due to Warne (2006) is

$$\begin{aligned} &f\left(\phi, \Gamma, \alpha, \beta, \Omega \mid k, r\right) \\ &= c_r \left|\Omega\right|^{-(m+q+r+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left[\Omega^{-1}\left(A + \frac{1}{\lambda_\alpha^2}\alpha\beta'\beta\alpha'\right)\right]\right) f\left(\Gamma \mid \Omega, k\right) \end{aligned} \tag{3.10}$$

with constants $\lambda_\alpha > 0, q \geq m$, and a positive definite matrix $A \in \mathbb{R}^{m\times m}$, and applies to the hyperparameters given that the lag order k and cointegration rank r are fixed. Here, $\operatorname{tr}(M)$ denotes the trace of a quadratic matrix M, i.e. the sum over all diagonal elements. The *normalising constant* c_r is

$$c_r = \left|A\right|^{q/2} \frac{\Gamma_r\left(m\right)}{\Gamma_m\left(q\right)\Gamma_r\left(r\right)} \frac{2^{-qm/2}\pi^{-m(m-1)/4}}{\left(2\pi\lambda_\alpha^2\right)^{mr/2}\pi^{(m-r)r/2}},$$

where $\Gamma_b\left(a\right)$ for $a, b \in \mathbb{N}_0$ with $a \geq b$ is a generalised form of the Gamma function as defined in Definition D.1. It is worth mentioning that f is an improper density function, as it remains constant for different values for ϕ. When the marginal prior for $\Gamma \mid \Omega, k$ is chosen to be constant, too, (3.10) yields the original reference prior by Villani (2005). This work adopts the generalisation by Warne (2006), who defines a proper distribution for $f\left(\Gamma \mid \Omega, k\right)$ in a similar manner as the classical *Minnesota prior* for AR terms in VAR models. For constants $\lambda_b, \lambda_l > 0$, define the diagonal $m(k-1) \times m(k-1)$ matrix $\Omega_\Gamma = \operatorname{diag}\left(\Omega_{\Gamma_1}, \ldots, \Omega_{\Gamma_{k-1}}\right)$ through the $k-1$ block matrices

$$\Omega_{\Gamma_i} = \frac{\lambda_b^2}{i^{2\lambda_l}} I_m, \quad i = 1, \ldots, k-1,$$

of dimension $m \times m$, respectively. Then Warne (2006) lets $\Gamma \mid \Omega$ be Matrix-Normal distributed with

$$\Gamma \mid \Omega, k \sim MN_{m(k-1)\times m}\left(0, \Omega_\Gamma, \Omega\right), \tag{3.11}$$

see Definition D.4. As outlined in the appendix, it holds for the vectorisation of $\Gamma \mid \Omega$ that

$$\operatorname{vec}\left(\Gamma\right) \mid \Omega, k \sim N_{m^2(k-1)}\left(0, \Omega \otimes \Omega_\Gamma\right),$$

where $\operatorname{vec}(\cdot)$ and $\otimes$ are the vectorisation and Kronecker matrix product operators defined in Appendix D.2. Obviously, Γ is conditionally independent of α, β, ϕ given Ω, r, k. The generalisation by Warne (2006) hence does not distort the results by Villani (2005) for $f\left(\Gamma \mid \Omega\right) = 1$ and, consequently, for the marginal distribution of Ω, it follows that

$$\Omega \sim IW_m\left(A, q\right),$$

where IW_m denotes the Inverse Wishart distribution, see Definition D.7. In Bayesian methodology, it is common to formulate probabilistic statements about variance parameters by their inverse, the so-called *precision*, i.e. Ω^{-1} has the conjugate m-dimensional Wishart prior from Definition D.6 with parameters A and q. Moreover, for the profile matrix α, the conditionally marginal prior is Matrix-normal with

$$\alpha \mid \beta, \Omega, r \sim MN_{m\times r}\left(0, \lambda_a^2\Omega, \left(\beta'\beta\right)^{-1}\right)$$

or, equivalently,

$$\operatorname{vec}\left(\alpha\right) \mid \beta, \Omega, r \sim N_{mr}\left(0, \left(\beta'\beta\right)^{-1} \otimes \lambda_\alpha^2\Omega\right).$$

For the lower block matrix β_l of $\beta = \left(I_r, \beta_l'\right)'$, it further holds that

$$\beta_l \mid r \sim Mt_{(m-r)\times r}\left(0, I_{m-r}, I_r, 0\right),$$

where Mt denotes the Matrix-t distribution as in Definition D.5. In addition to these marginal results derived by Villani (2005), it is shown by Warne (2006) that the marginal

prior for Γ is also Matrix-t distributed with

$$\Gamma \sim Mt_{m\times m(k-1)}\left(0, A^{-1}, \Omega_\Gamma, q-m\right).$$

The marginal distributions show that the constants A and q determine the prior for Ω and α. For $q \geq m+2$, Villani (2005) shows that $\mathrm{E}(\Omega) = (1/(q-m-1))\,A$, and hence, for given A, the uncertainty about unexplained variability in the model decreases in q, because the expected error covariance matrix converges to the zero matrix. The constant λ_α affects the uncertainty of α, since its columns have zero mean and covariance matrices $\lambda_\alpha^2\,\mathrm{E}(\Omega)$. A larger value therefore implies larger uncertainty. Since the magnitude of this covariance matrix depends on the expected value of Ω, Villani (2005) suggests to determine A and q first and to adjust λ_α in a second step. For the prior assumptions on Ω, the analyst needs to quantify the expected variances and, to be more informative, also covariances for the latent time series a priori, which will generally be hard to conceive. It is hence practicable to use an empirical Bayes approach based on an ML estimate such as $A = f_A(\widehat{\Omega})$, where f_A is a *tuning function* for the modeller, and $\widehat{\Omega}$ can be the empirical covariance matrix of the time series in levels or, if there is too much variability, the corresponding covariance matrix for the time series in differences. A particular choice could be $f_A(M) = (\lambda_A^2 m_{ij}^2)_{ij}$ for a square matrix $M = (m_{ij})_{ij}$ with some *tuning factor* $\lambda_A > 0$. In either case, one should use the minimum value $q = m+2$ for maximum uncertainty in order to reduce the effect of such an improper data-driven technique. Another choice for the first two constants, found in the literature, is $A = 0$ and $q = 0$, which yields an uninformative diffuse marginal prior for Ω. In light of the Minnesota prior for AR coefficients, the remaining constants λ_b and λ_l determine the uncertainty in the parameters $\Gamma_1, \ldots, \Gamma_{k-1}$, because Warne (2006) shows that it is $\mathrm{E}(\Gamma) = 0$ for $q \geq m+1$ and $\mathrm{Cov}(\mathrm{vec}(\Gamma)) = \Omega_\Gamma \otimes \mathrm{E}(\Omega)$ for $q \geq m+2$. The uncertainty in Γ is hence driven by Ω_{Γ_i} for $i = 1, \ldots, k-1$, whose overall magnitude is given by the *baseline constant* λ_b for given $\mathrm{E}(\Omega)$, whereas the lag constant λ_l measures the *shrinkage* towards the zero covariance matrix and, hence, to more certainty about vanishing AR coefficients with increasing order. Due to the dependence on the expected value of Ω, these two constants should be determined along with λ_α. For parsimony in the model, the chosen prior for $\Gamma \mid \Omega$ implies conditional independence between the different Γ_i, but through inclusion of further constants, the prior could be even more generalised to allow for dependence between the AR matrices. The above prior and its parametrisation are

motivated and interpreted in more detail by Villani (2005) and Warne (2006).

As a final note, the prior for $\mathcal{H}$ has so far been conditional on fixed values for k and r. As outlined in Appendix C.4, it is common to determine these values in a first step using, e.g., information criteria or subsequent hypothesis tests. In a Bayesian framework, k and r should also be considered random and chosen based on a posterior distribution that is driven by the data and prior beliefs. Since both constants affect the number of hyperparameters in $\mathcal{H}$, direct evaluation of these values along with all other hyperparameters is difficult. Villani (2005) assumes that the lag order k is either known or determined via fore-run Bayesian approaches – see the paper for references. The cointegration rank r can also be pre-determined through an individual Bayesian analysis, but Villani (2005) shows how to compute posterior probabilities for r replacing the prior in (3.10) by

$$f\left(\phi, \Gamma, \alpha, \beta, \Omega \mid k\right) = f\left(\phi, \Gamma, \alpha, \beta, \Omega \mid k, r\right) f\left(r \mid k\right)$$

with a prior $f\left(r \mid k\right)$ over all possible cointegration ranks $r = 0, 1, \ldots, m$. The approach requires computation of the marginal likelihoods $P(\mathcal{D} \mid \mathcal{E}, r)$ for all $r = 0, \ldots, m$, and apart from the cases $r = 0$ and $r = m$, these distributions can only be approximated via the MCMC results under corresponding specification of r. Using a joint prior $f(k, r)$ instead of $f\left(r \mid k\right)$, Warne (2006) extends this approach to posterior evaluation of all possible pairs (k, r), which additionally requires numerical evaluation of the marginal likelihoods $P(\mathcal{D} \mid \mathcal{E}, k, r)$. Due to the high-dimensionality of the time series under consideration, such approaches are not yet feasible in reasonable time. In fact, possible values for k are principally restricted to 0 or 1, since operations regarding the matrix Γ become intractable. For simplicity, the value of r will also be assumed known, although a posterior analysis as outlined in the above references is possible when the model is estimated parallel for different r.

3.3.3 Posterior Distributions

The aim of the Bayesian estimation procedure for the BMPMP model is to identify *joint posterior distributions* for the parameters $\mathcal{K}$ and the hyperparameters $\mathcal{H}$. Due to the complexity in both the prior distributions and likelihood, no closed-form results on joint

posterior distributions can be derived. However, MCMC methods are applied in the context of the BMPMP model in order to approximate the distribution. As described in Appendix B in more detail, the general idea is to run sampling algorithms, which iteratively simulate new realisations of the posterior distribution for the parameters. For the BMPMP model, both the *Gibbs* and *Metropolis-Hastings algorithms* from Appendices B.2 and B.3 are used. Due to the high-dimensional complexity in the BMPMP model, emphasis must be given to a careful empirical diagnosis of convergence to the desired limiting distribution in the Markov chain, see Appendix B.4 for details.

First, simulation of the hyperparameters in $\mathcal{H}$ is examined. In the Bayesian VECM estimation via the Grassman approach by Villani (2005) and its extension by Warne (2006), it is possible to derive marginal posterior distributions for each hyperparameter given all other parameters. The existence of *full conditional distributions* therefore allows for applying the Gibbs sampler. In each step i, one has to simulate from the following distributions given the current realisations for the other parameters. Beginning with the outcomes from the last step $i-1$, denoted by the respective superscripts, each hyperparameter is visited individually, and a new realisation with superscript i is simulated. For hyperparameters that have already been visited, the new state of the chain is used in the successive simulations of other hyperparameters. For theoretical and practical details of Gibbs sampling, the reader is referred to Appendix B.2.

To begin with, the new realisation $\Omega^{(i)}$ for Ω is drawn from the marginal posterior distribution

$$\Omega \mid \phi^{(i-1)}, \Gamma^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r \sim IW_m\left(S_\Omega^{(i)}, n_\Omega\right),$$

where

$$\begin{aligned} S_\Omega^{(i)} &= \varepsilon^{(i-1)}\varepsilon^{(i-1)\prime} + A + \left(1/\lambda_\alpha^2\right)\alpha^{(i-1)}\beta^{(i-1)\prime}\beta^{(i-1)}\alpha^{(i-1)\prime} + \Gamma^{(i-1)}\Omega_\Gamma^{-1}\Gamma^{(i-1)\prime}, \\ n_\Omega &= T - k + q + r + m(k-1), \\ \varepsilon^{(i-1)} &= (\Delta K)^{(i-1)} - \Phi^{(i-1)} - \Gamma^{(i-1)}K_2^{(i-1)} - \alpha^{(i-1)}\beta^{(i-1)\prime}K_1^{(i-1)}. \end{aligned}$$

Next, $\phi^{(i)}$ is drawn through simulation of

$$\phi \mid \Omega^{(i)}, \Gamma^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r \sim N_m\left(\mu_\phi^{(i)}, \Omega^{(i)}\right),$$

where

$$\mu_\phi^{(i)} = \sum_{t=k+1}^{T}\left(\Delta\kappa_t^{(i-1)} - \sum_{j=1}^{k-1}\Gamma_j^{(i-1)}\Delta\kappa_{t-j}^{(i-1)} - \alpha^{(i-1)}\beta^{(i-1)\prime}\kappa_{t-1}^{(i-1)}\right)$$

is the vector resulting from row-wise summation over the matrix

$$(\Delta K)^{(i-1)} - \Gamma^{(i-1)}K_2^{(i-1)} - \alpha^{(i-1)}\beta^{(i-1)\prime}K_1^{(i-1)}.$$

The i-th realisation for Γ is simulated from a Matrix-normal distribution, which can be expressed through the vectorisation operator in terms of a multivariate normal distribution. In particular,

$$\mathrm{vec}\,(\Gamma) \mid \Omega^{(i)}, \phi^{(i)}, \alpha^{(i-1)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r \sim N_{m^2(k-1)}\left(\mu_\Gamma^{(i)}, \Sigma_\Gamma^{(i)}\right),$$

where

$$\mu_\Gamma^{(i)} = \mathrm{vec}\left[\left((\Delta K)^{(i-1)} - \Phi^{(i)} - \alpha^{(i-1)}\beta^{(i-1)\prime}K_1^{(i-1)}\right)K_2^{(i-1)\prime} \times \left(K_2^{(i-1)}K_2^{(i-1)\prime} + \Omega_\Gamma^{-1}\right)^{-1}\right]$$

and

$$\Sigma_\Gamma^{(i)} = \left(K_2^{(i-1)}K_2^{(i-1)\prime} + \Omega_\Gamma^{-1}\right)^{-1} \otimes \Omega^{(i)}.$$

The new value for α is obtained through its vectorised form via

$$\mathrm{vec}\,(\alpha) \mid \Omega^{(i)}, \phi^{(i)}, \Gamma^{(i)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r \sim N_{mr}\left(\mu_\alpha^{(i)}, \Sigma_\alpha^{(i)}\right),$$

where

$$\mu_{\alpha}^{(i)} = \text{vec}\left[\left((\Delta K)^{(i-1)} - \Phi^{(i)} - \Gamma^{(i)} K_2^{(i-1)}\right) K_1^{(i-1)\prime} \beta^{(i-1)} \times \left[\beta^{(i-1)\prime}\left(K_1^{(i-1)} K_1^{(i-1)\prime} + \frac{1}{\lambda_{\alpha}^2} I_m\right)\beta^{(i-1)}\right]^{-1}\right]$$

and

$$\Sigma_{\alpha}^{(i)} = \left[\beta^{(i-1)\prime}\left(K_1^{(i-1)} K_1^{(i-1)\prime} + \frac{1}{\lambda_{\alpha}^2} I_m\right)\beta^{(i-1)}\right]^{-1} \otimes \Omega^{(i)}.$$

Finally, the vectorised lower block matrix $\text{vec}\,(\beta_l)$ has the conditional posterior distribution

$$\text{vec}\,(\beta_l) \mid \Omega^{(i)}, \phi^{(i)}, \Gamma^{(i)}, \alpha^{(i)}, \mathcal{K}^{(i-1)}, k, r \sim N_{r(m-r)}\left(\mu_{\beta_l}^{(i)}, \Sigma_{\beta_l}^{(i)}\right),$$

where

$$\begin{aligned}
\mu_{\beta_l}^{(i)} &= \Sigma_{\beta_l}^{(i)} \left(I_r \otimes c_{\perp}\right)' \left(\Sigma_{\beta}^{(i)}\right)^{-1} \left(\mu_{\beta}^{(i)} - \text{vec}\,(c)\right), \\
\Sigma_{\beta_l}^{(i)} &= \left[\left(I_r \otimes c_{\perp}\right)' \left(\Sigma_{\beta}^{(i)}\right)^{-1} \left(I_r \otimes c_{\perp}\right)\right]^{-1}, \\
\mu_{\beta}^{(i)} &= \Sigma_{\beta}^{(i)} \,\text{vec}\left[K_1^{(i-1)} \left((\Delta K)^{(i-1)} - \Phi^{(i)} - \Gamma^{(i)} K_2^{(i-1)}\right)' \left(\Omega^{(i)}\right)^{-1} \alpha^{(i)}\right], \\
\Sigma_{\beta}^{(i)} &= \left[\left(\alpha^{(i)\prime}\left(\Omega^{(i)}\right)^{-1} \alpha^{(i)}\right) \otimes \left(K_1^{(i-1)} K_1^{(i-1)\prime} + \frac{1}{\lambda_{\alpha}^2} I_m\right)\right]^{-1}.
\end{aligned}$$

Now all five hyperparameters are updated, where for Γ, α, and β, the stacked vectors must be rearranged into matrix format again. The conditional posteriors require simulations from the Inverse Wishart and the multivariate Normal distributions, both of which are easily available in statistical standard software.

With the new states for the hyperparameters at hand, in the second step, the old realisations for the parameters in $\mathcal{K}^{(i-1)}$ must be replaced. Due to the dependencies between different κ_t resulting from the underlying dynamics process, full conditionals cannot be derived for these parameters, making the Gibbs sampler intractable. Consequently, the Metropolis-Hastings algorithm is applied with the parameters from the CBD model. The remainder of this section derives the particular algorithm in the context of the BMPMP model, where general theoretical and practical details on Metropolis-Hastings sampling are reviewed in Appendix B.3. For a fixed calendar year $t \in \{1, \ldots, T\}$, let

$$\begin{aligned}
\mathcal{K}_{-t} &= \{\kappa_1, \ldots, \kappa_{t-1}, \kappa_{t+1}, \ldots, \kappa_T\}, \\
\mathcal{D}_t &= \{D_{xpt} \; \forall \; x, p, \; p \neq p^*\}, \\
\mathcal{E}_t &= \{E_{xpt} \; \forall \; x, p, \; p \neq p^*\}.
\end{aligned}$$

From Bayes' Theorem it then follows that

$$f\left(\kappa_t \mid \mathcal{D}, \mathcal{E}, \mathcal{H}, \mathcal{K}_{-t}, k, r\right) \propto f\left(\mathcal{D} \mid \mathcal{E}, \mathcal{K}, k, r\right) f\left(\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r\right).$$

Due to (3.8), the underlying data $\mathcal{D}_t$ for one calendar year t depend on the mortality rate q_{xpt} only, which itself is purely described by κ_t. Using the mutual independence of the numbers of deaths among different calendar years, it holds for the likelihood – when interpreted as a function in κ_t – that

$$\begin{aligned}
&f\left(\mathcal{D} \mid \mathcal{E}, \mathcal{K}, k, r\right) \\
&\propto f\left(\mathcal{D}_t \mid \mathcal{E}_t, \mathcal{K}_t, k, r\right) \\
&\propto \prod_x \prod_p \left[\log\left(1 + \exp(\eta_{xpt})\right)\right]^{D_{xpt}} \left(1 + \exp\left(\eta_{xpt}\right)\right)^{-E_{xpt}},
\end{aligned}$$

where dependence on κ_t is captured by η_{xpt}. For the conditional distribution of $\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r$, it follows from the VECM representation that κ_t with $t > k$ depends only on the k preceding values through $\kappa_t = \phi + (I_m + \alpha\beta')\kappa_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta\kappa_{t-i} + \varepsilon_t$. Of course, no preceding values but marginal priors are given for $\kappa_1, \ldots, \kappa_k$. The conditional

distribution for any κ_t can be most generally written as

$$
\begin{aligned}
& f\left(\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r\right) \\
&= \frac{f\left(\mathcal{K}, \mathcal{H}, k, r\right)}{f\left(\mathcal{K}_{-t}, \mathcal{H}, k, r\right)} \\
&\propto f\left(\mathcal{K}, \mathcal{H}, k, r\right) \\
&\propto f\left(\mathcal{K} \mid \mathcal{H}, k, r\right) \\
&= f\left(\kappa_T \mid \mathcal{K}_{-T}, \mathcal{H}, k, r\right) f\left(\mathcal{K}_{-T}, \mathcal{H}, k, r\right) \\
&= f\left(\kappa_T \mid \mathcal{K}_{-T}, \mathcal{H}, k, r\right) f\left(\kappa_{T-1} \mid \kappa_1, \ldots, \kappa_{T-2}, \mathcal{H}, k, r\right) f\left(\kappa_1, \ldots, \kappa_{T-2}, \mathcal{H}, k, r\right) \\
&= \ldots \\
&= \prod_{s=k+1}^{T} f\left(\kappa_s \mid \kappa_1, \ldots, \kappa_{s-1}, \mathcal{H}, k, r\right) f\left(\kappa_1, \ldots, \kappa_k, \mathcal{H}, k, r\right) \\
&\propto \prod_{s=k+1}^{T} f\left(\kappa_s \mid \kappa_{s-k}, \ldots, \kappa_{s-1}, \mathcal{H}, k, r\right) \prod_{s=1}^{k} f\left(\kappa_s\right).
\end{aligned}
$$

Using the residuals $\varepsilon_s = \kappa_s - \phi - \left(I_m + \alpha\beta'\right)\kappa_{s-1} - \sum_{i=1}^{k-1} \Gamma_i \Delta\kappa_{s-i}$, $s = k+1, \ldots, T$, from this general result it follows

$$
\begin{aligned}
f\left(\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r\right) &\propto \prod_{s=k+1}^{k+t} f\left(\kappa_s \mid \kappa_{s-k}, \ldots, \kappa_{s-1}, \mathcal{H}, k, r\right) f\left(\kappa_t\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\sum_{s=k+1}^{k+t} \varepsilon_s' \Omega^{-1} \varepsilon_s + \left(\kappa_t - \mu_t\right)' \Sigma_t^{-1} \left(\kappa_t - \mu_t\right)\right)\right)
\end{aligned}
$$

for $t = 1, \ldots, k$,

$$
\begin{aligned}
f\left(\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r\right) &\propto \prod_{s=t}^{k+t} f\left(\kappa_s \mid \kappa_{s-k}, \ldots, \kappa_{s-1}, \mathcal{H}, k, r\right) \\
&\propto \exp\left(-\frac{1}{2} \sum_{s=t}^{k+t} \varepsilon_s' \Omega^{-1} \varepsilon_s\right)
\end{aligned}
$$

for $t = k+1, \ldots, T-k$, and

$$f\left(\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r\right) \propto \prod_{s=t}^{T} f\left(\kappa_s \mid \kappa_{s-k}, \ldots, \kappa_{s-1}, \mathcal{H}, k, r\right)$$
$$\propto \exp\left(-\frac{1}{2}\sum_{s=t}^{T} \varepsilon_s' \Omega^{-1} \varepsilon_s\right)$$

for $t = T-k+1, \ldots, T$. Combining the results for the likelihood and the conditional distribution finally leads to

$$\begin{aligned} & f\left(\kappa_t \mid \mathcal{D}, \mathcal{E}, \mathcal{H}, \mathcal{K}_{-t}, k, r\right) \\ & \propto \exp\left[\sum_x \sum_p \left[D_{xpt} \log\left(\log\left(1+\exp(\eta_{xpt})\right)\right) - E_{xpt} \log\left(1+\exp(\eta_{xpt})\right)\right]\right] \\ & \times \begin{cases} \exp\left(-\frac{1}{2}\left(\sum_{s=k+1}^{k+t} \varepsilon_s' \Omega^{-1} \varepsilon_s + \left(\kappa_t - \mu_t\right)' \Sigma_t^{-1} \left(\kappa_t - \mu_t\right)\right)\right), & t \leq k \\ \exp\left(-\frac{1}{2}\sum_{s=t}^{k+t} \varepsilon_s' \Omega^{-1} \varepsilon_s\right), & k < t \leq T-k \\ \exp\left(-\frac{1}{2}\sum_{s=t}^{T} \varepsilon_s' \Omega^{-1} \varepsilon_s\right), & t > T-k \end{cases} . \end{aligned}$$

As mentioned earlier, the constants $\Sigma_1, \ldots, \Sigma_k$ may equal one constant covariance matrix Σ. Direct sampling from this distribution is not straightforward, hence the Gibbs sampler is not available for updating the CBD parameters. Instead, the Metropolis-Hastings sampling is applied, where at step i, the algorithm walks through all calendar years t as follows.

- Generate a *candidate*

$$\kappa_t^* \sim N_m\left(\kappa_t^{(i-1)}, \Sigma_{\text{MH}}\right) \tag{3.12}$$

with positive definite covariance matrix Σ_{MH}, which is known beforehand or determined within the *burn-in period* of the MCMC algorithm.

- Compute the *acceptance probability*

$$a\left(\kappa_t^{(i-1)}, \kappa_t^*\right) = \min\left\{1, \frac{f\left(\kappa_t^* \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}{f\left(\kappa_t^{(i-1)} \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}\right\}.$$

  Note that the fraction is computable since the unknown proportionality factors cancel out. The updates are made conditional on

$$\mathcal{K}_{-t}^{(i-1/i)} = \left\{\kappa_1^{(i)}, \ldots, \kappa_{t-1}^{(i)}, \kappa_{t+1}^{(i-1)}, \ldots, \kappa_T^{(i-1)}\right\}$$

  of all other parameter realisations from the last or, if already updated in step i, from the current iteration, as well as conditional on the fresh values $\mathcal{H}^{(i)} = \left\{\phi^{(i)}, \alpha^{(i)}, \beta^{(i)}, \Gamma^{(i)}, \Omega^{(i)}\right\}$ for the hyperparameters. Details are provided in Appendix E.2.

- Generate a random probability $u \sim U([0,1))$. If $u \leq a\left(\kappa_t^{(i-1)}, \kappa_t^*\right)$ then set $\kappa_t^{(i)} = \kappa_t^*$, otherwise let $\kappa_t^{(i)} = \kappa_t^{(i-1)}$.

If m is large, the Metropolis-Hastings algorithm may work inefficiently due to poor multivariate proposals κ_t^*, which result in low acceptance rates. In order to increase the number of movements in the Markov chain, it is suggested to split the parameters κ_t^* into single or groups of elements, for which the algorithm is then applied individually. This simplification is known as *single-component Metropolis-Hastings algorithm*, see Gilks (2005) for instance. In addition, as discussed in Appendix E.3, it is numerically advantageous. The covariance matrix Σ_{MH} for the proposal distribution may be chosen as a diagonal matrix with one unique entry or, due to different magnitudes in the univariate time series, may depend on the different parameters. A well-working choice is to set $\Sigma_{\text{MH}} = c_{\text{MH}} A$, for which the different magnitudes are captured by the constant A, and tuning of the acceptance rates is further achieved through manipulation of the factor c_{MH}. The single-component Metropolis-Hastings algorithm simply extracts the diagonal entries as marginal variances.

Initial values for $\mathcal{K}$ in the MCMC algorithm are their corresponding ML estimates from the CBD model, such that convergence can be expected to be reached faster with a much shorter burn-in period. Note that the CBD model is a Binomial generalised linear re-

gression with D_{xpt} events out of E_{xpt} trials and link function $m_{xpt} \mapsto \log(\exp(m_{xpt}) - 1)$, comprising the logit link for q_{xpt} as in (3.2) and the relation between q_{xpt} and m_{xpt} given by (2.1). For the hyperparameters, reasonable starting values are A for Ω as well as the $m \times r$ matrix $(I_r, 0_{r \times m-r})'$ for β. Furthermore, it is set $\Gamma^{(0)} = 0 \in \mathbb{R}^{m \times (k-1)m}$, $\alpha^{(0)} = 0 \in \mathbb{R}^{m \times r}$ and $\phi^{(0)} = 0 \in \mathbb{R}^m$.

As a summary, an MCMC method, which combines both the Gibbs and the Metropolis-Hastings algorithms, is applied with Bayesian parameter estimation in the BMPMP model. Implementation is straightforward, whereas this procedure may be time-consuming when the number of populations becomes large. Following Geweke (1996), the Gibbs sampling technique for the hyperparameters of the VECM, based on the reference prior by Villani (2005), meets the minimum conditions to guarantee theoretical convergence. Due to the normality assumption in the VECM and the normality of the proposal values for the parameters $\mathcal{K}$, *ergodicity* can further be established for the Metropolis-Hastings algorithm using statements on sufficient conditions as in Robert and Casella (2004). Despite the theoretical results, whether or not practical convergence has happened must be assessed by the results from the procedure as outlined in Appendix B.4. An example follows in the next chapter.

3.3.4 Weighted Posterior

In case of large values for the time horizon T or the number of populations n_p, a very large amount of observations is supposed to be fitted by a comparably easy model framework. Low magnitudes for both the likelihood and priors for the time series parameters cause the posterior distribution for $\mathcal{K}$ to be almost *degenerate*, i.e. probability mass for some κ_t is highly concentrated around one point with vanishing variance. This result stems from the simplifying model assumptions, e.g. Poisson distributed numbers of deaths and the VECM structure for the time series, which do not allow for flexibility in case of overdispersion, model misspecification, and data anomalies. Considering posterior predictive forecasting based on the marginal posterior distribution for the hyperparameters, as described in the next section, a degenerated posterior for $\mathcal{K}$ is not of major concern. Therefore, in the course of the case studies conducted in Chapter 4, the amendments proposed in this section will not be of further interest. However, in case a valid approximation of this posterior is desired for the calibration window, several techniques

are available in the literature, for example the inclusion of independent and Gamma distributed nuisance parameters θ_{xpt} with mean 1, such that $D_{xpt} \sim \text{Poi}(E_{xpt} q_{xpt} \theta_{xpt})$. This results in a simple *Poisson-Gamma two-stage model*, in which the random factors measure unexplained variability not covered by the systematic part $E_{xpt} q_{xpt}$. Such approaches are well-known in different statistical applications, for instance see Wakefield (2007) for a review on random effects models in spatial data analysis.

A disadvantage of two-stage models in this Bayesian framework is the large number of additional parameters that have to be estimated. Motivated by work on *weighted likelihood* approaches in Bayesian inference, see Agostinelli and Greco (2012) or Newton and Raftery (1994), one can down-weight the posteriors by an exponent $w \in (0, 1]$ chosen by the modeller, i.e. instead of $f(\kappa_t \mid \mathcal{D}, \mathcal{E}, \mathcal{H}, \mathcal{K}_{-t}, k, r)$ sample from

$$c_w f(\kappa_t \mid \mathcal{D}, \mathcal{E}, \mathcal{H}, \mathcal{K}_{-t}, k, r)^w \propto f(\mathcal{D} \mid \mathcal{E}, \mathcal{K}, k, r)^w \, f(\kappa_t \mid \mathcal{H}, \mathcal{K}_{-t}, k, r)^w \tag{3.13}$$

with normalising constant c_w, which only depends on w and need not be known, as it cancels out in the Metropolis-Hastings algorithm. This naturally leads to higher variability in the posterior distribution, because large values for f are down-weighted and values smaller than 1 are increased. Thereby, variances larger than what is observed by the pure model and likelihood specification, can be accommodated. A motivating example is a nominal normal posterior with mean μ and variance σ^2. It is easy to see that taking the density to power $w \in (0, 1]$ leads again to the normal distribution with same mean and variance $\sigma^2 / w \geq \sigma^2$. One can think of w as a weight for both the likelihood and the prior: it expresses the certainty w.r.t. these model specifications compared to a uniform distribution for the data D_{xpt} over the range of, say, $[0, E_{xpt}]$, and a non-informative prior for κ_t, because the posterior is proportional to the product of the term on the right-hand side of (3.13) and $(\prod_{x,p,t} E_{xpt}^{-1})^{1-w} \cdot 1^{1-w}$. The result of incorporating w is an easy way to introduce flexibility to the model framework, which allows for a wide enough parameter space for the time series parameters $\mathcal{K}$. Only one additional parameter must be pre-selected by the analyst, capturing the prior belief about certainty w.r.t. model assumptions and, hence, following the Bayesian paradigm. Finally, it is easy to see that for the BMPMP model, the modification of the posterior

carries through all calculations and leads to the acceptance probability

$$a\left(\kappa_t^{(i-1)}, \kappa_t^*\right) = \min\left\{1, \exp(w)\frac{f\left(\kappa_t^* \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}{f\left(\kappa_t^{(i-1)} \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}\right\}.$$

3.4 Bayesian Forecasting

Once a sufficiently large sample of the posterior distribution is simulated, the estimated parameters are used to forecast the mortality rates of the underlying populations into the future. In many real-life examples, the mortality and crude death rates exhibit certain random noise, which makes a precise short-term prediction rather difficult. However, the incorporation of first-order AR components and the cointegration terms is expected to lead to global estimation of the dependency structure between the marginal time series. This allows for comparably sophisticated forecasts in the long-run as will be seen in the application of the BMPMP model in the subsequent chapter. Due to the complexity of the parameters and their joint distribution, forecasts are based on simulation rather than analytical computations.

Besides the prediction of future values, i.e. the main goal of this model, the following forecast procedure can also be applied for model validation. In the latter case, following Appendix A.3, the posterior distribution of the hyperparameters is used to predict the time series over the calendar years, whose observations were used as the training dataset. For a good model fit, the probabilistic behaviour of the resulting forecast should naturally coincide with what was observed from the original data. Of course, this procedure tends to overfit the data, since the model parameters are estimated to exactly fit the particularly observed paths. More sophisticated model validation tools can be applied, such as division of the observed data into two independent training and test sets, to reduce the impact of overfitting.

Mortality rates are forecast via prediction of the VECM model parameters $\mathcal{K}$ into the future and final computation of q_{xpt} via the corresponding CBD equation from Section 3.2.1. Although one specific calendar year serves as starting point, no exact initial value for the VECM is given. First, crude death rates are subject to random noise due to measurement errors and natural fluctuations, and second – and even more important

– the time series representation in $\mathcal{K}$ is latent and must be estimated. For the purpose of easy model forecasts, one could simply use the ML estimate for the parameters in the CBD model of the corresponding calendar year, i.e. κ_T when predictions start with the last observed mortality data. In light of the entire Bayesian philosophy, however, it is preferred to use the already realised posterior sample for the corresponding parameter κ_T as a distribution for the starting value. The uncertainty in both death rates and latent time series representation is then inherently integrated. Particular improvement in expressing uncertainty is seen for the very first years of predictions due to the auto-correlation in $\mathcal{K}$.

For any realised starting value of the time series, a set of all different hyperparameters is drawn from the joint posterior distribution. Following the normal assumption of the VECM, the entire time series can be successively generated through realisations of the m-dimensional normal distribution, where mean and covariance matrix are determined by the previous step and the hyperparameters. Simulation of the hyperparameters is easily done by random sampling from the corresponding joint MCMC sample, given that the Markov chain has converged and a pseudo-independent sample is obtained, eventually after deleting a sufficiently large burn-in period. Due to the *Ergodic Theorem*, the simplest way to do so is to walk through the entire path of the realised Markov chain and to use the hyperparameter values $\mathcal{H}^{(i)} = \left\{\phi^{(i)}, \alpha^{(i)}, \beta^{(i)}, \Gamma^{(i)}, \Omega^{(i)}\right\}$ at each stage i for the time series simulation. As the number of MCMC steps is already large, one simulation of the time series for each set of hyperparameters naturally leads to a sample size for the predicted values of the same size, but several simulations of the VECM per realised set of hyperparameters are of course possible, too.

Having obtained the forecasts over a pre-determined time span, particular interest with this model then lies in the joint distributions of the estimated mortality rates. In comparison to marginal or low-dimensional mortality forecast models, the BMPMP model has its specific strength in detection of future dependencies between different populations. Furthermore, due to the normal assumption in the VECM, simulation is even straightforward when certain scenarios for some of the marginal time series are assumed. If there is a strong prior assumption of the, say, overall development of mortality, an according deterministic sample path for the intercept can be postulated. Such assumptions are reasonable when other scientific studies forecast a general trend much better

than the BMPMP approach based on biological and social factors. The BMPMP model can then be used to predict the marginal influences of the global development on the individual populations. Since the multivariate normal distribution in the VECM implies normal conditional distributions, simulation of the remaining time series is done in the same manner as before, and the resulting effects on the mortality rates can be interpreted w.r.t. the given assumptions on the global mortality progression. In particular, if $\kappa_t = (\kappa_t^{(1)\prime}, \kappa_t^{(2)\prime})'$ is divided into an m_1-dimensional sub-vector $\kappa_t^{(1)}$ and an m_2-dimensional $\kappa_t^{(2)}$ with $m = m_1 + m_2$, one makes use of the fact that $\kappa_t^{(1)} \mid \kappa_t^{(2)} = k_t^{(2)}$ is m_2-variate normal distributed with mean

$$\begin{aligned}
&\phi^{(1)} + \left((I_{m_1}, 0) + \alpha^{(1)}\beta'\right)\kappa_{t-1} + \sum_{i=1}^{k-1} \Gamma_i^{(1)} \Delta\kappa_{t-i} \\
&\quad + \Omega_{12}\Omega_{22}^{-1}\left(k_t^{(2)} - \phi^{(2)} - \left((0, I_{m_2}) + \alpha^{(2)}\beta'\right)\kappa_{t-1} - \sum_{i=1}^{k-1} \Gamma_i^{(2)} \Delta\kappa_{t-i}\right)
\end{aligned}$$

and variance

$$\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21},$$

where $(I_{m_1}, 0) \in \mathbb{R}^{m_1 \times m}, (0, I_{m_2}) \in \mathbb{R}^{m_2 \times m}, \phi = (\phi^{(1)\prime}, \phi^{(2)\prime})'$ with $\phi^{(1)} \in \mathbb{R}^{m_1}, \phi^{(2)} \in \mathbb{R}^{m_2}$, $\alpha = (\alpha^{(1)\prime}, \alpha^{(2)\prime})'$ with $\alpha^{(1)} \in \mathbb{R}^{m_1 \times r}, \alpha^{(2)} \in \mathbb{R}^{m_2 \times r}$, $\Gamma_i = (\Gamma_i^{(1)\prime}, \Gamma_i^{(2)\prime})'$ with $\Gamma_i^{(1)} \in \mathbb{R}^{m_1 \times m}, \Gamma_i^{(2)} \in \mathbb{R}^{m_2 \times m}$ for all $i = 1, \ldots, k-1$, and

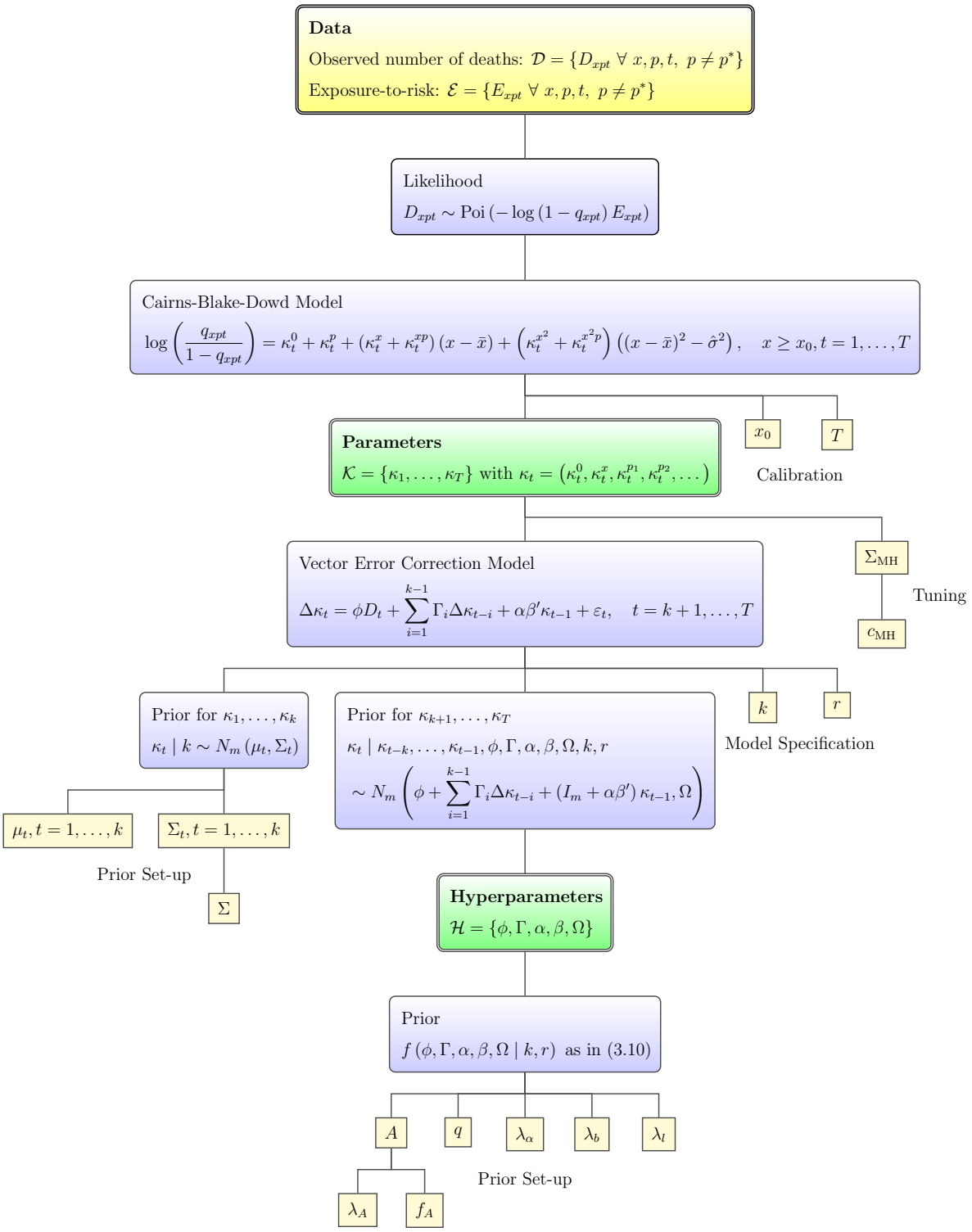
$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

with $\Omega_{11} \in \mathbb{R}^{m_1 \times m_1}, \Omega_{12} = \Omega_{21}' \in \mathbb{R}^{m_1 \times m_2}, \Omega_{22} \in \mathbb{R}^{m_2 \times m_2}$. Note that a scenario-dependent forecast will not be further discussed in this work.

3.5 Model Summary

Figure 3.1 summarises the BMPMP model.

Figure 3.1: The Bayesian Multi-Population Mortality Projection Model



The Bayesian estimation algorithm at iteration i is summarised a follows.

- Gibbs sampler for the hyperparameters: Simulate from

$$\begin{aligned}
\Omega \mid \phi^{(i-1)}, \Gamma^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r &\sim IW_m\left(S_\Omega^{(i)}, n_\Omega\right), \\
\phi \mid \Omega^{(i)}, \Gamma^{(i-1)}, \alpha^{(i-1)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r &\sim N_m\left(\mu_\phi^{(i)}, \Omega^{(i)}\right), \\
\mathrm{vec}\left(\Gamma\right) \mid \Omega^{(i)}, \phi^{(i)}, \alpha^{(i-1)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r &\sim N_{m^2(k-1)}\left(\mu_\Gamma^{(i)}, \Sigma_\Gamma^{(i)}\right), \\
\mathrm{vec}\left(\alpha\right) \mid \Omega^{(i)}, \phi^{(i)}, \Gamma^{(i)}, \beta^{(i-1)}, \mathcal{K}^{(i-1)}, k, r &\sim N_{mr}\left(\mu_\alpha^{(i)}, \Sigma_\alpha^{(i)}\right), \\
\mathrm{vec}\left(\beta_l\right) \mid \Omega^{(i)}, \phi^{(i)}, \Gamma^{(i)}, \alpha^{(i)}, \mathcal{K}^{(i-1)}, k, r &\sim N_{r(m-r)}\left(\mu_{\beta_l}^{(i)}, \Sigma_{\beta_l}^{(i)}\right),
\end{aligned}$$

  with parameters as given in Section 3.3.3.

- Metropolis-Hastings sampler for the parameter κ_t:
  - Simulate $\kappa_t^* \sim N_m\left(\kappa_t^{(i-1)}, \Sigma_{\mathrm{MH}}\right)$.
  - Compute

$$a\left(\kappa_t^{(i-1)}, \kappa_t^*\right) = \min\left\{1, \exp(w)\frac{f\left(\kappa_t^* \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}{f\left(\kappa_t^{(i-1)} \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}\right\}.$$

  - Generate $u \sim U\left([0,1)\right)$. If $u \leq a\left(\kappa_t^{(i-1)}, \kappa_t^*\right)$ then set $\kappa_t^{(i)} = \kappa_t^*$, otherwise let $\kappa_t^{(i)} = \kappa_t^{(i-1)}$.

4 Case Studies in European Mortality Forecasting

In this chapter, the BMPMP model is calibrated with mortality data of different European countries to assess the model's ability to fulfil the desired targets outlined in Section 3.1. Particular attention is devoted to investigate the flexibility to model an arbitrary number of arbitrary populations. In both sections of this chapter, the BMPMP model is applied with gender-specific mortality rates of five European countries, respectively. This number clearly exceeds the restriction of only two populations for most of the existing multi-population mortality projection models, but remains small enough for illustration purposes. While Section 4.1 comprises a standard example with data from commonly investigated Western European countries, a selection of Central European countries, which could not be satisfactorily modelled in the augmented common factor model by Li and Lee (2005), is included in Section 4.2. Besides discussion of the choice of priors, convergence diagnostics of the MCMC algorithm, and general goodness-of-fit analyses, the main focus lies on diagnostic assessment and interpretation of the output taking into consideration the multi-population context, including a comparison of the joint forecast with results from univariate Bayesian CBD models for each country. Furthermore, the second distinct feature of the BMPMP model, the Bayesian paradigm, is compared to its frequentist counterpart where possible, i.e. only for such model specifications, for which the high-dimensional maximisation problem is non-singular.

All data in this chapter are obtained through the Human Mortality Database (2014), available at `http://www.mortality.org`, from the underlying data sources referenced therein. The Human Mortality Database provides a rich set of mortality-related figures for a wide range of mostly developed countries. Original data are collected from the national statistical bureaus or institutes. For each country, the observed number of deaths D_{xgt} and the corresponding exposure-to-risk E_{xgt} are available for both genders $g = m, f$,

where m and f denote male and female, respectively, for all ages $x = 0, 1, \ldots, 109$ and all remaining ages, denoted by 110+. All data will be taken as provided by the Human Mortality Database (2014) and considered absolutely comparable. In particular, differences in data quality and methodologies of how the data were derived or handling changes of territorial claims will not be discussed in this work – the interested reader is referred to the list of all considered countries in the preface of this work for details on the territorial coverage and the documentation in the Human Mortality Database (2014) for further information. The time horizon for t is country-specific and varies between a decade and hundreds of years. For most countries, the post-war period is sufficiently covered such that the BMPMP model can be at least calibrated for the last 60 or so years.

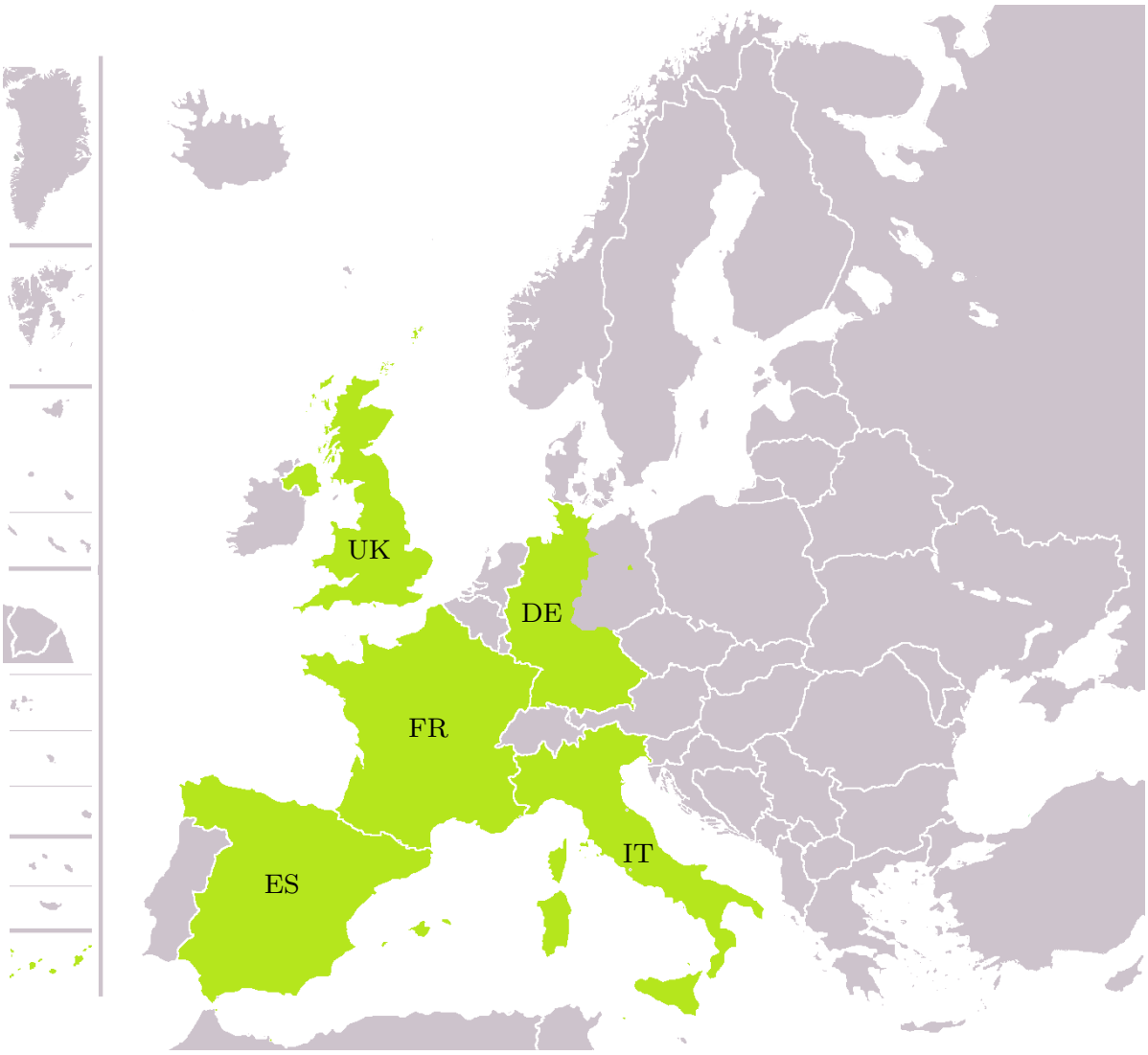
For all the following case studies, numerical results are obtained through the author's `bmpmp` package in the freely available statistical programming software R. It is designed for flexible and efficient estimation and output of the BMPMP model for joint forecasts of different countries with distinction of both genders. The routines in the package allow the modeller to create required datasets using the data from the Human Mortality Database (2014), to calibrate and to forecast the BMPMP model, and to automatically obtain important graphical output for the analysis. Further tools, such as ML estimation, are available for comparison purposes. The `bmpmp` package is available upon request to the author and described in detail in Appendix F.

4.1 Case Study 1: The Big Five

In this section, the BMPMP model is calibrated with gender-specific data on the so-called *Big Five*, i.e. the five largest countries in Western Europe: France, Germany, Italy, Spain, and the United Kingdom. See Figure 4.1 for a map. Due to different patterns in mortality because of differences in social, economic, and medical circumstances in a divided Germany until re-unification and even thereafter, the data for Germany are restricted to the population of West Germany (i.e. the then Federal Republic of Germany) until 1990 and the corresponding territory (i.e. the old states within the re-unified Federal Republic of Germany) since then, but referred to simply as Germany for convenience. For all other countries, see the information in the preface on details of territorial coverage. In order to interpret country-specific parameters as deviations from

Figure 4.1: Map of the Big Five

Shown in green are the so-called *Big Five*, i.e. the five largest countries in Western Europe: France (FR), Germany (DE, data only for West Germany in this case study), Italy (IT), Spain (ES), and the United Kingdom (UK), within Europe. For details on territorial coverage for these five countries, see the list of countries in the preface.



the average mortality in all considered countries, the overall sample is used as reference and sometimes loosely called *European* population. The ages under consideration are restricted to the interval $[40, 100]$. Data availability differs for the countries under consideration. The combined sample of all countries is available from 1956 to 2009. The model is calibrated w.r.t. the time horizon from 1956 to 1995. The remaining data are used for external out-of-sample validation. Due to regime changes within this calibration period, in the latter course of this case study, the model is estimated based on data from 1981–2009 and forecast until the year of 2100.

4.1.1 Model Equations

Following the gender-specific equation in (3.4), the CBD equation in the BMPMP model for this case study is set up as follows:

$$\begin{aligned}\log\left(\frac{q_{xpgt}}{1-q_{xpgt}}\right) &= \kappa_t^0 + \kappa_t^p + \kappa_t^g + \kappa_t^{pg} + \left(\kappa_t^x + \kappa_t^{xp} + \kappa_t^{xg}\right)(x - x_0) \\ &\quad + \left(\kappa_t^{x^2} + \kappa_t^{x^2 p} + \kappa_t^{x^2 g}\right)(x - x_0)^2, \quad x \geq x_0, \end{aligned} \tag{4.1}$$

with $t = 1, \ldots, 40$ representing the years $1956, \ldots, 1995$, $x = 40, \ldots, 100$, $g = m, f$, and $p \in \mathcal{P} = \{\mathrm{DE}, \mathrm{ES}, \mathrm{FR}, \mathrm{IT}, \mathrm{UK}, \mathrm{EU}\}$, where the countries are abbreviated by their Internet top-level domains[1], and $p^* = \mathrm{EU}$ stands for the overall sample of all countries. In contrast to the common equations in Section 3.2.1, this case study centres the age variables at $x_0 = 40$ instead of the average $\bar{x}$ and uses $\widehat{\sigma}^2 = 0$, i.e. the reference for the quadratic age parameters is the absolute squared differences w.r.t. the age of 40 rather than the excess of the empirical variance in age. For uniqueness in parametrisation, it is assumed that $\kappa_t^{\mathrm{EU}} = \kappa_t^m = \kappa_t^{\mathrm{EU},m} = \kappa_t^{x,\mathrm{EU}} = \kappa_t^{x^2,\mathrm{EU}} = \kappa_t^{x,m} = \kappa_t^{x^2,m} = 0$ for all t. There are five populations with 61 age groups for both genders in each of the 40 calendar years, i.e. the dataset consists of 24,400 entries. On the other hand, for each calendar year, the CBD equation adds one intercept, one main effect for gender, two main effects for age (for the linear and quadratic terms, respectively), and five main effects for the countries. In addition, second-order interactions between age, gender, and population include further 17 parameters. Therefore, one ends up with 26 parameters in each CBD equation,

[1] ISO 3166-1 alpha-2 codes with specific replacement of GB by UK for the United Kingdom: DE – Germany (restricted to the territory of former West Germany), ES – Spain, FR – France, IT – Italy, UK – United Kingdom.

i.e. 1,040 parameters in total. The CBD stage in the BMPMP model in this example is indeed parsimonious with a parameter-to-data quota of 4.3%. Note that individual CBD models for all ten populations, comprised of all five countries and both genders, would actually require 1,200 parameters.

The VECM equation is then given as outlined in Section 3.2.2 with the linear normalisation for β. The lag order is chosen to be $k = 2$ to allow for one AR parameter in the VECM representation. For numerical efficiency, the lag order is not further increased, as any additional AR coefficient matrix introduces $26^2 = 676$ additional parameters. For the cointegration rank, the choice of $r = 5$ corresponds to the prior belief of at most five stationary linear combinations of the marginal time series, respectively. The number of parameters in the VECM becomes 26 for ϕ, 130 for α and 105 for β, 676 for Γ_1, and $26(26+1)/2 = 351$ for Ω due to symmetry, i.e. 1,104 in total. The number of parameters in both stages of the BMPMP model is of similar magnitude, and the VECM parameters may even exceed the CBD parameters. Hence, frequentist estimation becomes problematic or, as seen with this example, impossible. The alternative of independent models would see a much lower number of parameters, but neglect any quantification of dependencies. Although the remainder of this case study mainly focuses on this particular specification of the model with $k = 2$ and $r = 5$, due to the difficulty of even presenting such high-dimensional results for one model estimation, comments on other model specifications will be made w.r.t. results from calibration of model set-ups by all combinations of $k = 1, 2$ and $r = 0, 1, 5$ in Section 4.1.9.

4.1.2 Choice of Priors

Starting with the constants for the hyperparameters in the VECM, the values for $A, q, \lambda_\alpha, \lambda_b$ are set as follows. For the expected covariance matrix, the empirical Bayes approach with $A = f_A(\widehat{\Omega})$ is applied. Here, $\widehat{\Omega}$ is the empirical covariance matrix of the time series in differences, which is tuned by $f_A(M) = (\lambda_A^2 m_{ij}^2)_{ij}$ for $M = (m_{ij})_{ij}$ with $\lambda_A = \sqrt{10}$. If this choice of A becomes numerically singular, its off-diagonal elements are set to zero. The choice of $q = m + 2$ for maximum uncertainty is made to reduce the impact of the empirical Bayes approach. The value for λ_α is chosen to be one such that α has covariance matrix $\mathrm{E}(\Omega)$. Finally, it is $\lambda_b = 5$ to increase uncertainty in Γ_1. For the case $k = 1$, this parameter is not required, and for any choice of k, there is at

most only one lag for Γ, so that in this case study no determination of λ_l as shrinkage factor is required.

The CBD parameters for the first two calendar years, which are not immediately modelled through the VECM, require additional priors for their marginal distributions. Again, an empirical Bayes approach is used, in which μ_1 and μ_2 equal the ML estimates of the CBD parameters of the first two years in levels. The covariance matrices Σ_1 and Σ_2 are set equal to A. For the models with $k = 1$, the prior choices are only made for μ_1 and Σ_1.

4.1.3 Initialisation of the Algorithm and Starting Values

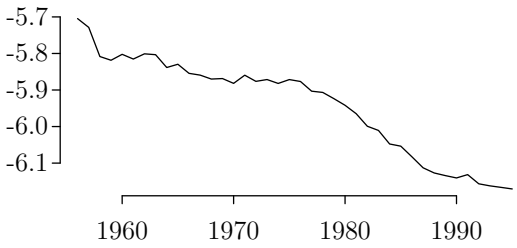
In this example, the single-component version of the Metropolis-Hastings algorithm is applied. The proposal covariance matrix is set to be $\Sigma_{\text{MH}} = c_{\text{MH}} A$ with tuning factor $c_{\text{MH}} = 10$ to achieve higher acceptance rates by increasing the innovation steps. The single-component algorithm only requires the diagonal entries of Σ_{MH} for the marginal proposals. Finally, initial values for the CBD parameters are the respective ML estimates on the Binomial generalised linear regression with events D_{xpgt} out of E_{xpgt} trials and link function $m_{xpgt} \mapsto \log(\exp(m_{xpgt}) - 1)$, and the starting values for the hyperparameters are chosen as outlined in Section 3.3.3. The MCMC algorithm is run for a total of $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100, i.e. only every 100-th iteration is stored to reduce memory space for the high-dimensional output.

The starting values for $\mathcal{K}$ are plotted in Figure 4.2 for a general understanding in model interpretation. The upper six plots show the ML estimates of the intercept and all gender- and age-related main and interaction effects as time series for $t = 1, \ldots, 40$. The remaining lower plots show the main effects of the distinct populations as well as their interactions with gender and age. Since the values of each of the latter types of parameters are directly comparable with each other, as they represent deviations from the overall reference population, the plots for each parameter category contain the according time series for all populations p.

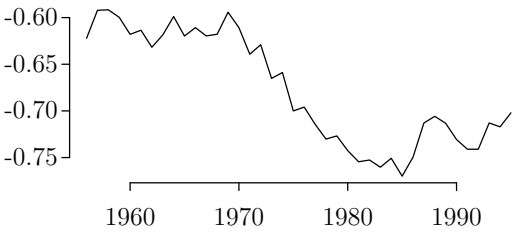
The intercept, which is to be interpreted as the logit mortality rate of a 40-year-old male in the reference population, clearly exhibits the negative trend known from studies on

Figure 4.2: Starting values for $\mathcal{K}$

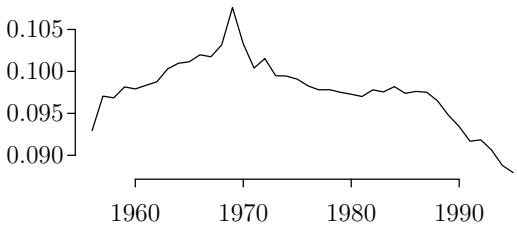
Shown are the ML estimates $\kappa_t^{(0)}$ of the CBD parameters as time series in t for $1956, \ldots, 1995$.



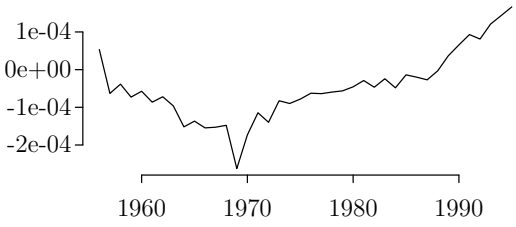
(a) Intercept $\kappa_t^{0(0)}$



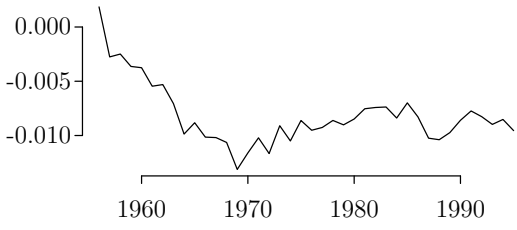
(b) Gender effect for females $\kappa_t^{f(0)}$



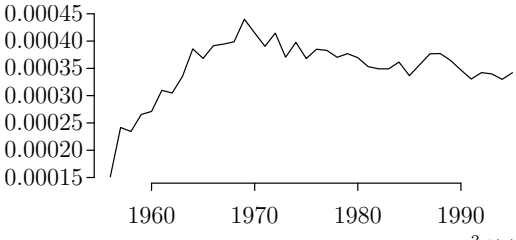
(c) Linear age effect $\kappa_t^{x(0)}$



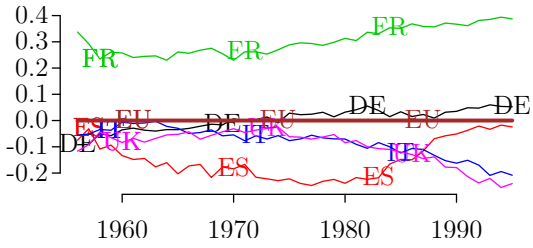
(d) Quadratic age effect $\kappa_t^{x^2(0)}$



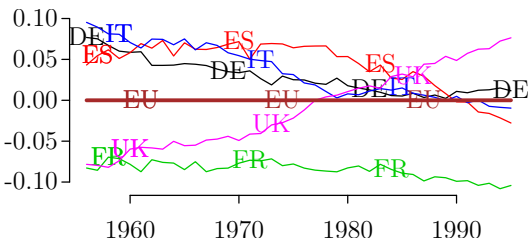
(e) Gender and linear age interactions $\kappa_t^{xf(0)}$



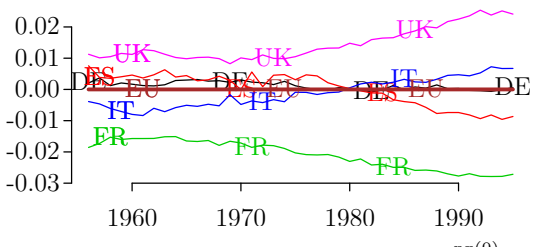
(f) Gender and quadratic age interactions $\kappa_t^{x^2 f(0)}$



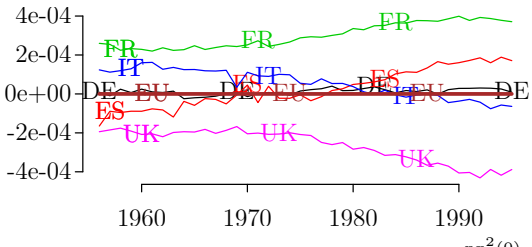
(g) Main country effects $\kappa_t^{p(0)}$



(h) Gender and country interactions $\kappa_t^{pf(0)}$



(i) Linear age and country interactions $\kappa_t^{px(0)}$



(j) Quadratic age and country interactions $\kappa_t^{px^2(0)}$

mortality in developed countries cited in Chapter 2. Moreover, the increased pace of mortality improvements seen for the second half of the calibration period is consistent with applications of the standard CBD model, see Cairns et al. (2006) for instance, and suggests that model calibration may be eventually improved by shifting the beginning of the time horizon to the late 1970s. The negative values in the time series for the effect of females reflect the biological fact of lower mortality among women, where the exact values have to be interpreted for females aged 40 in the reference population on the logit scale. The 1970s clearly mark a time of even further improvements in mortality for females in this age group on top of the general pattern given by the intercept.

While identifying a regime change for the age-related parameters around the year of 1970 with opposite movements beforehand, since then the linear age effect declines, and the quadratic effect becomes stronger. The interactions with gender remain merely stable over time since 1970, so that above effects are apparent for both genders with a constantly more quadratic curve for females. The linear and quadratic age effects are multiplied by values between 0–60 and 0–3600, respectively, and thus the magnitudes of the time series must be interpreted accordingly. Furthermore, interpretation of these parameters must be done simultaneously, indicating that in addition to the general improvement via the intercept, the linearity in the logits of mortality rates gets lost over time in favour of a quadratic curve. As a result, the middle age groups experience a larger extent of mortality improvement compared to low age groups around 40 years and high age groups around 100 years. Quadratic effects were found to be significant with data from males of the United States by Cairns et al. (2006), but an important consequence of this finding for European data is that quadratic patterns in the logits of mortality rates seem to be present for other populations in the developed world. Since the effect of quadratic age patterns seems to become more important over time, this suggests that mortality prediction models in the CBD framework with only linear age effects may be too simplistic in the description of age-dependent mortality improvements.

Finally, the country-specific plots describe their deviations from the overall population. For each country, its four specific time series adjust the intercept $\kappa_t^{0(0)}$ and the main effects for gender and age, i.e. $\kappa_t^{f(0)}, \kappa_t^{x(0)}, \kappa_t^{x^2(0)}$. Vanishingly small values for Germany, which is the largest country in terms of population among all five, indicate that it is well-represented by the reference population. In contrast, for 40-year-old males, French

mortality is constantly higher, while Spanish mortality uses to be lower until 1990. Relative improvements for Italy and the United Kingdom emerge towards the end of the calibration window. Effects for females are opposite, i.e. compared to the overall reference, there is a constantly larger gap between French men and women and a historically smaller gap for Germany, Italy, and Spain. For the United Kingdom, the specific trend of men catching up with women is outstanding. The two last plots reveal that French mortality data exhibit the strongest quadratic pattern throughout the time horizon among all five countries, while the logits of mortality rates for the United Kingdom are still linear. Interestingly, the country-specific deviations from the reference population do not seem to converge over the calibration period. Hypotheses of convergence stated a priori, as known from other models, are questioned by the BMPMP model in favour of this data-driven approach.

The marginal time series in Figure 4.2 are obviously non-stationary, but can be justified to be trend-stationary when analysing plots of their first differences (not shown). Direct comparison of the main linear and quadratic age effects or comparison of such time series with their respective interactions with gender – just to name a few – reveal a strong correlation among many of the marginal time series. It is evident that a joint time series approach, which in the BMPMP model is given by the VECM, is vital and suggests that estimation of the covariance matrix Ω is of major importance. Even more, the strong correlation further motivates the choice of this model, because linear combinations of different marginal time series in levels, e.g. $\kappa_t^{x(0)} - \kappa_t^{x^2(0)}$ or $\kappa_t^{xf(0)} - \kappa_t^{x^2f(0)}$, seem to be indeed stationary.

4.1.4 Convergence Diagnostics

Before performing any type of statistical inference, the BMPMP model and the output of its MCMC algorithm must be diagnosed for validity. The aim of the MCMC procedure is to approximate the posterior distribution of the model parameters by an extensive sample of pseudo-independent realisations. Although theoretical ergodicity could be established for the BMPMP model, it must be verified that the Markov chain indeed converged to this limiting distribution within the given iterations, and that the sample is sufficiently large for the Ergodic Theorem to make autocorrelation negligible. Failure in obtaining the correct posterior distribution falsifies results from the statistical

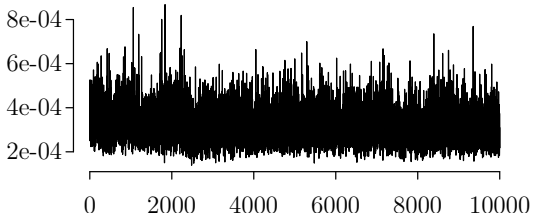
analysis. However, as far as the purpose of mortality forecasts with the BMPMP model is concerned, primary focus must be devoted to the marginal posterior distribution of the set of hyperparameters. Bayesian forecasting is equivalent to obtaining the posterior predictive distribution of future values of the CBD parameters through posterior sampling of the hyperparameters only, rather than sampling from the posterior distribution of the given CBD parameters.

While referring to Appendix B.4 for a general introduction to different concepts of convergence diagnostics in the MCMC framework, the convergence analysis in this case study is mainly based on graphical assessment of the marginal parameter distributions. Two of the most important tools, plots of the paths of marginal parameters w.r.t. the MCMC iterations as well as plots of the current ergodic means versus the iteration index, will be displayed in detail. However, convergence of the marginal distribution can only indicate that the joint distribution of all parameters and hyperparameters might have reached its equilibrium. The total number of more than 2,000 parameters and hyperparameters makes the Markov chain such high-dimensional that convergence of the joint distribution is practically impossible to verify. Moreover, this high number allows the discussion of only selected convergence diagnostics plots as an explanatory representation of the full set of parameters. As mentioned earlier, results in this section are presented for the model with $k = 2$ and $r = 5$, noting that plots for the five other model set-ups, which cannot be shown in this work, generally display the same behaviour.

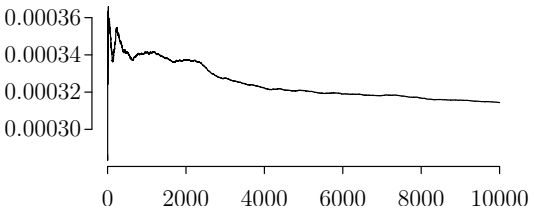
The following seven figures, i.e. Figures 4.3 to 4.9, show paths of every 100-th iteration for selected marginal hyperparameters and parameters (left panels) and the corresponding evaluation of ergodic means (right panels). As typical in Bayesian analysis, convergence for the covariance matrix Ω, analysed in Figure 4.3, is additionally diagnosed for the precision Ω^{-1} in order to better assess the stability of equilibriums, see Figure 4.4. For the analysis of convergence in the cointegration space, immediate parameters of the matrix $\Pi = \alpha\beta'$ instead of entries in both sub-parameters are plotted in Figure 4.5. Remaining plots are given for the hyperparameters ϕ and $\Gamma = \Gamma_1$ in Figures 4.6 and 4.7, respectively. On the first hierarchy level, i.e. for the CBD parameters $\mathcal{K}$, Figures 4.8 and 4.9 depict convergence diagnoses for selected entries of κ_{20} (i.e. the time series at calendar year 1975) and κ_{40} (i.e. the time series at calendar year 1995), respectively.

Figure 4.3: Convergence diagnostics for Ω

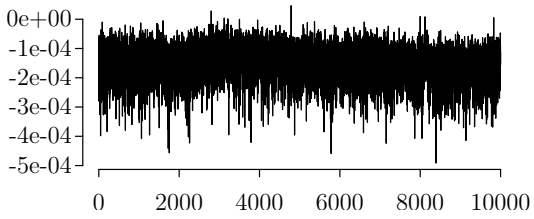
Left panel: Realised marginal paths of selected entries of Ω after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



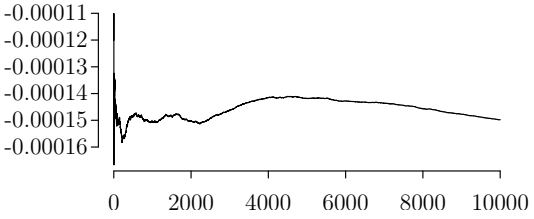
(a) MCMC path for Ω_{11}



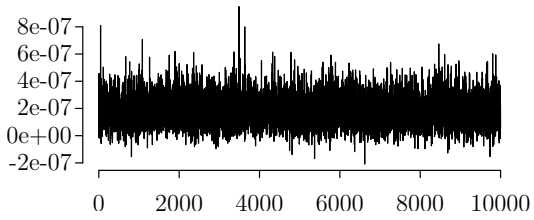
(b) Cumulative plot of ergodic mean for Ω_{11}



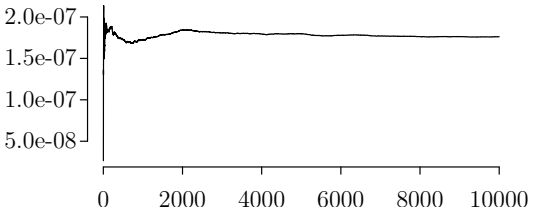
(c) MCMC path for Ω_{21}



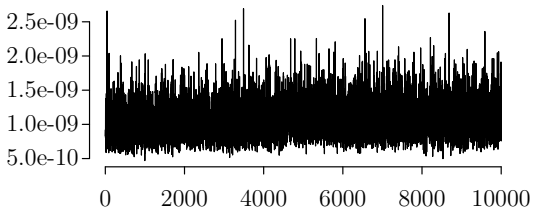
(d) Cumulative plot of ergodic mean for Ω_{21}



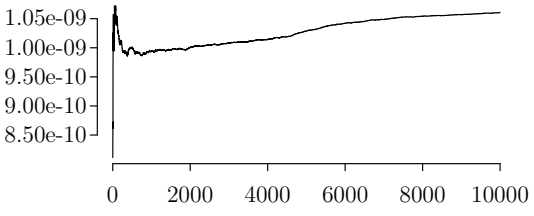
(e) MCMC path for Ω_{41}



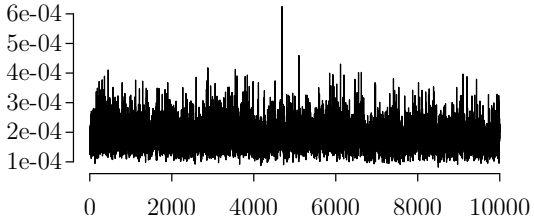
(f) Cumulative plot of ergodic mean for Ω_{41}



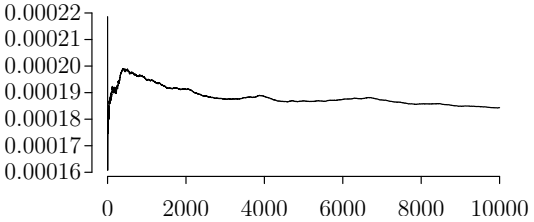
(g) MCMC path for Ω_{44}



(h) Cumulative plot of ergodic mean for Ω_{44}



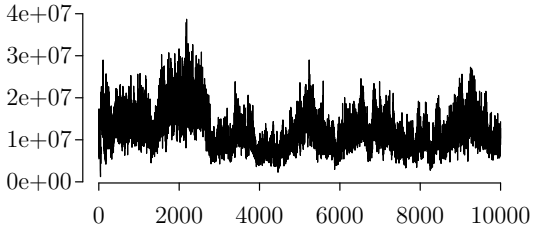
(i) MCMC path for Ω_{77}



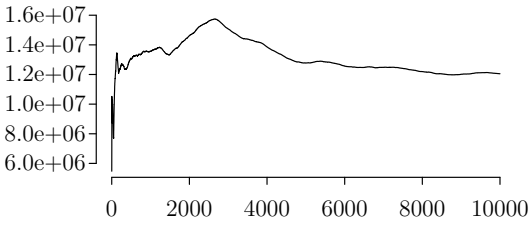
(j) Cumulative plot of ergodic mean for Ω_{77}

Figure 4.4: Convergence diagnostics for Ω^{-1}

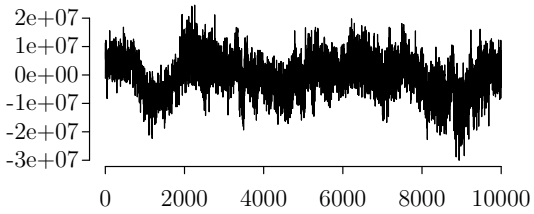
Left panel: Realised marginal paths of selected entries of Ω^{-1} after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means.
Indices: Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



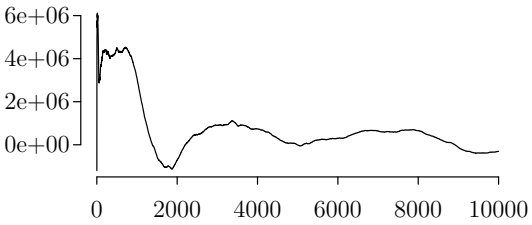
(a) MCMC path for Ω_{11}^{-1}



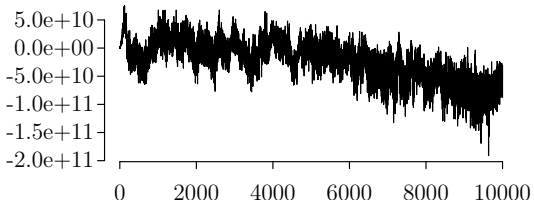
(b) Cumulative plot of ergodic mean for Ω_{11}^{-1}



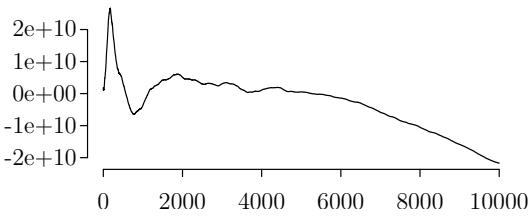
(c) MCMC path for Ω_{21}^{-1}



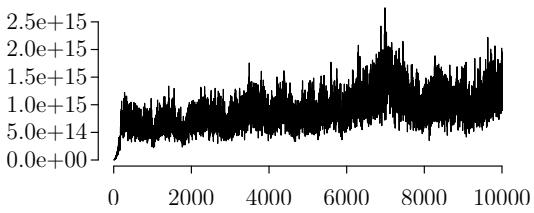
(d) Cumulative plot of ergodic mean for Ω_{21}^{-1}



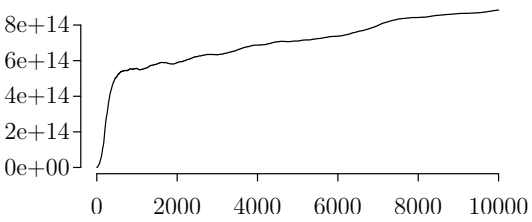
(e) MCMC path for Ω_{41}^{-1}



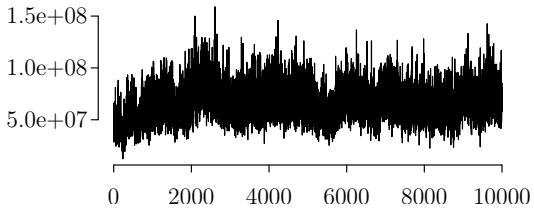
(f) Cumulative plot of ergodic mean for Ω_{41}^{-1}



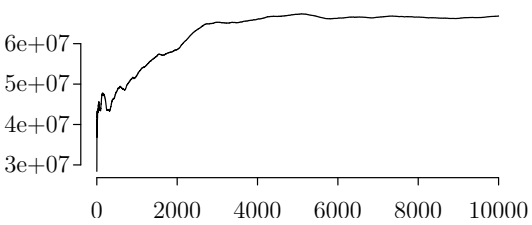
(g) MCMC path for Ω_{44}^{-1}



(h) Cumulative plot of ergodic mean for Ω_{44}^{-1}



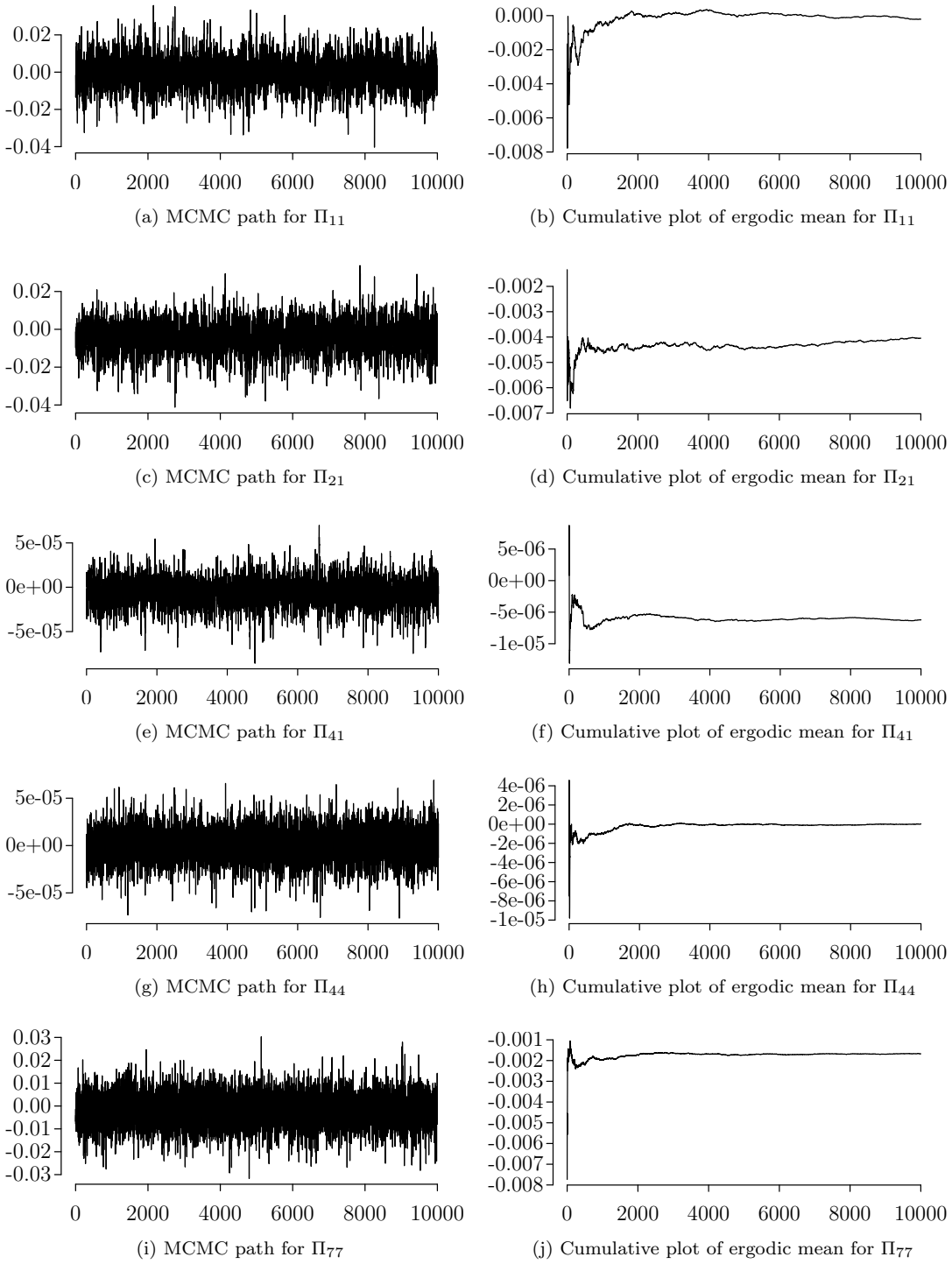
(i) MCMC path for Ω_{77}^{-1}



(j) Cumulative plot of ergodic mean for Ω_{77}^{-1}

Figure 4.5: Convergence diagnostics for $\Pi = \alpha\beta'$

Left panel: Realised marginal paths of selected entries of Π after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



(a) MCMC path for Π_{11}

(b) Cumulative plot of ergodic mean for Π_{11}

(c) MCMC path for Π_{21}

(d) Cumulative plot of ergodic mean for Π_{21}

(e) MCMC path for Π_{41}

(f) Cumulative plot of ergodic mean for Π_{41}

(g) MCMC path for Π_{44}

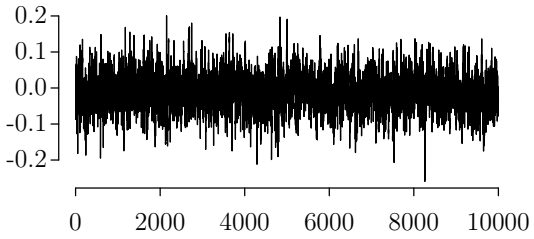
(h) Cumulative plot of ergodic mean for Π_{44}

(i) MCMC path for Π_{77}

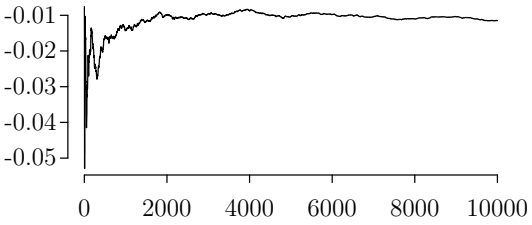
(j) Cumulative plot of ergodic mean for Π_{77}

Figure 4.6: Convergence diagnostics for ϕ

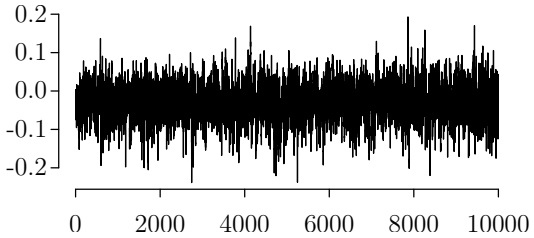
Left panel: Realised marginal paths of selected entries of ϕ after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$, $17 - \kappa_t^{x,\mathrm{DE}}$.



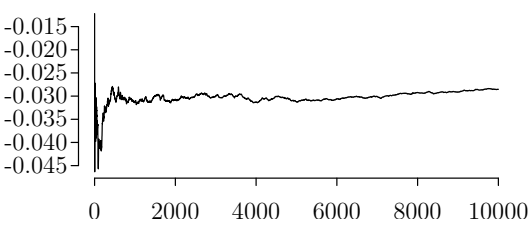
(a) MCMC path for ϕ_1



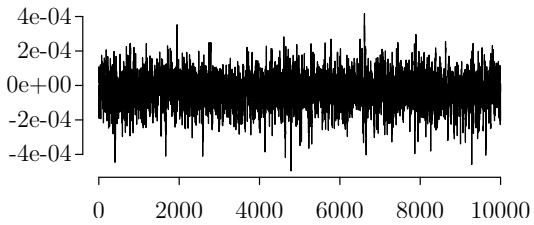
(b) Cumulative plot of ergodic mean for ϕ_1



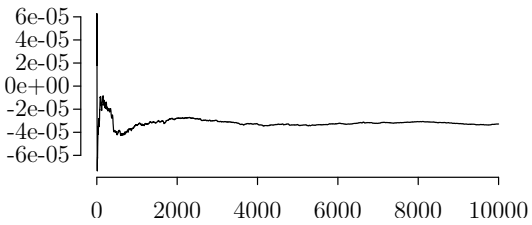
(c) MCMC path for ϕ_2



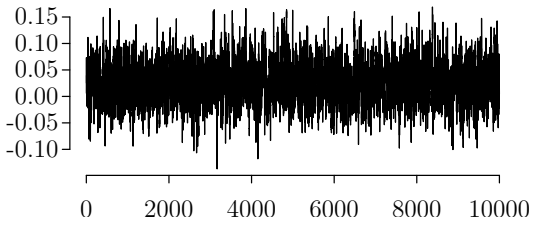
(d) Cumulative plot of ergodic mean for ϕ_2



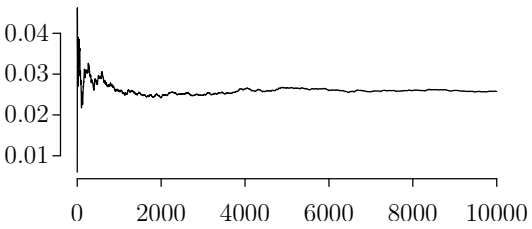
(e) MCMC path for ϕ_4



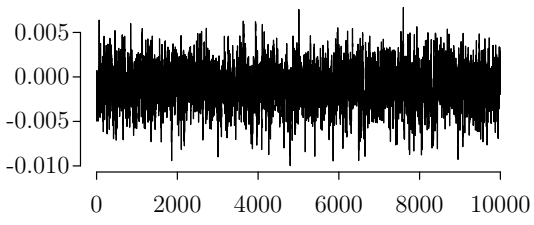
(f) Cumulative plot of ergodic mean for ϕ_4



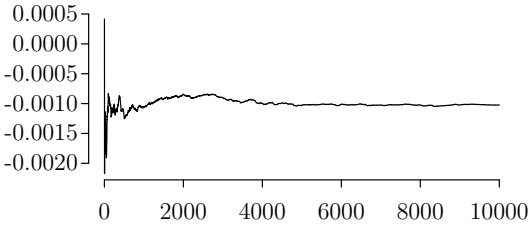
(g) MCMC path for ϕ_7



(h) Cumulative plot of ergodic mean for ϕ_7



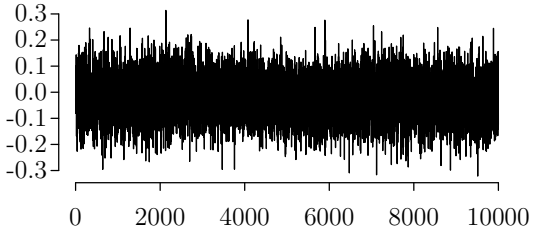
(i) MCMC path for ϕ_{17}



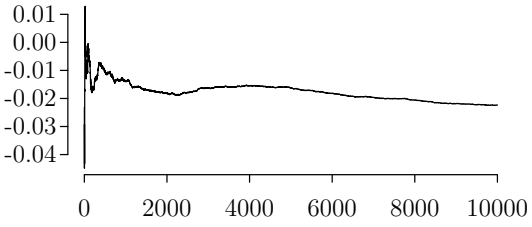
(j) Cumulative plot of ergodic mean for ϕ_{17}

Figure 4.7: Convergence diagnostics for $\Gamma = \Gamma_1$

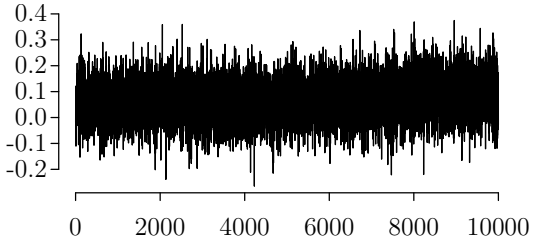
Left panel: Realised marginal paths of selected entries of Γ after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



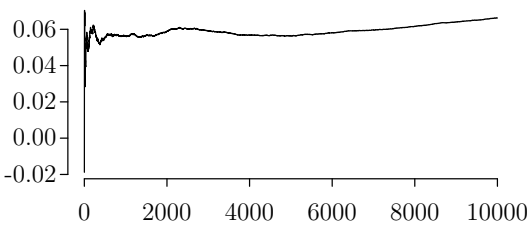
(a) MCMC path for Γ_{11}



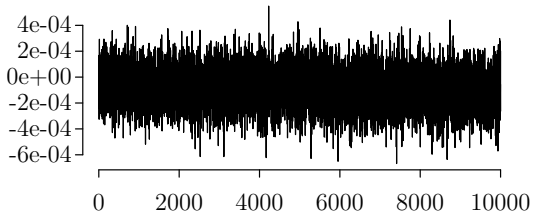
(b) Cumulative plot of ergodic mean for Γ_{11}



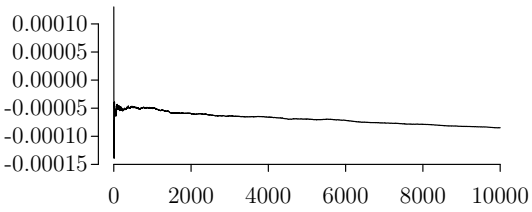
(c) MCMC path for Γ_{21}



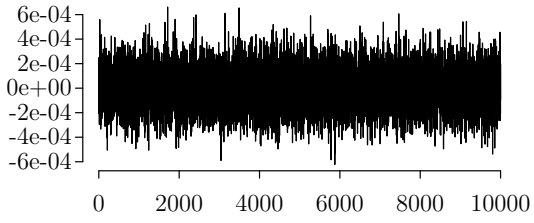
(d) Cumulative plot of ergodic mean for Γ_{21}



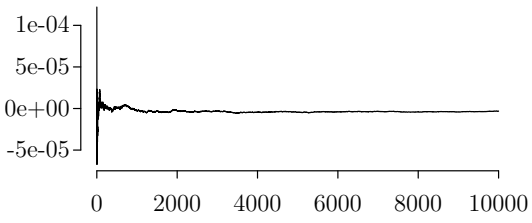
(e) MCMC path for Γ_{41}



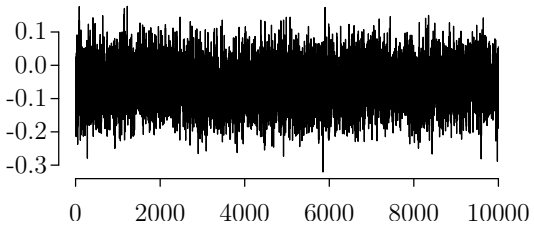
(f) Cumulative plot of ergodic mean for Γ_{41}



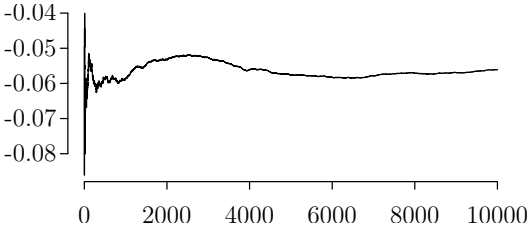
(g) MCMC path for Γ_{44}



(h) Cumulative plot of ergodic mean for Γ_{44}



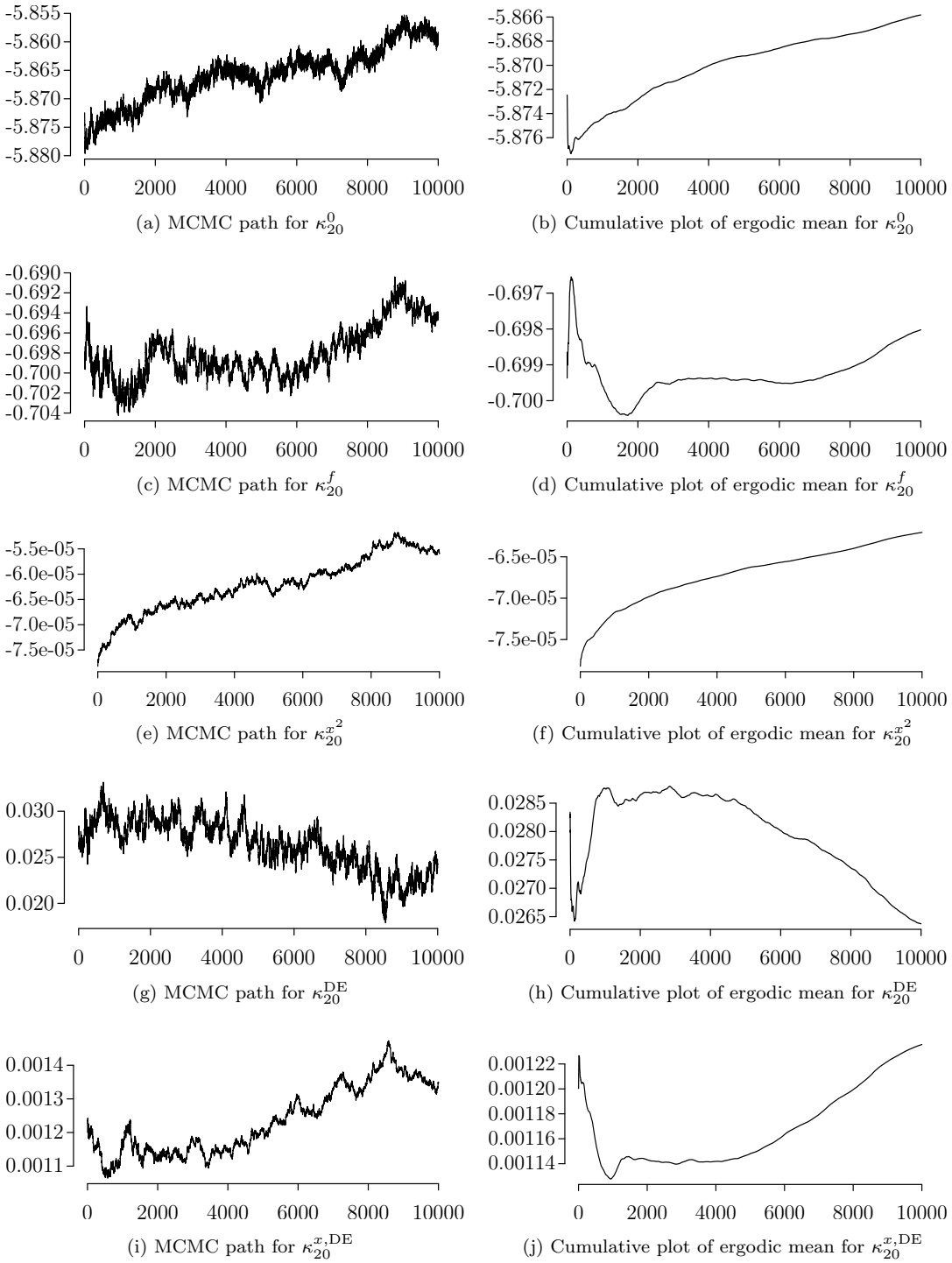
(i) MCMC path for Γ_{77}



(j) Cumulative plot of ergodic mean for Γ_{77}

Figure 4.8: Convergence diagnostics for κ_{20}

Left panel: Realised marginal paths of selected entries of κ_{20} after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means.



(a) MCMC path for κ_{20}^{0}

(b) Cumulative plot of ergodic mean for κ_{20}^{0}

(c) MCMC path for κ_{20}^{f}

(d) Cumulative plot of ergodic mean for κ_{20}^{f}

(e) MCMC path for $\kappa_{20}^{x^2}$

(f) Cumulative plot of ergodic mean for $\kappa_{20}^{x^2}$

(g) MCMC path for $\kappa_{20}^{\mathrm{DE}}$

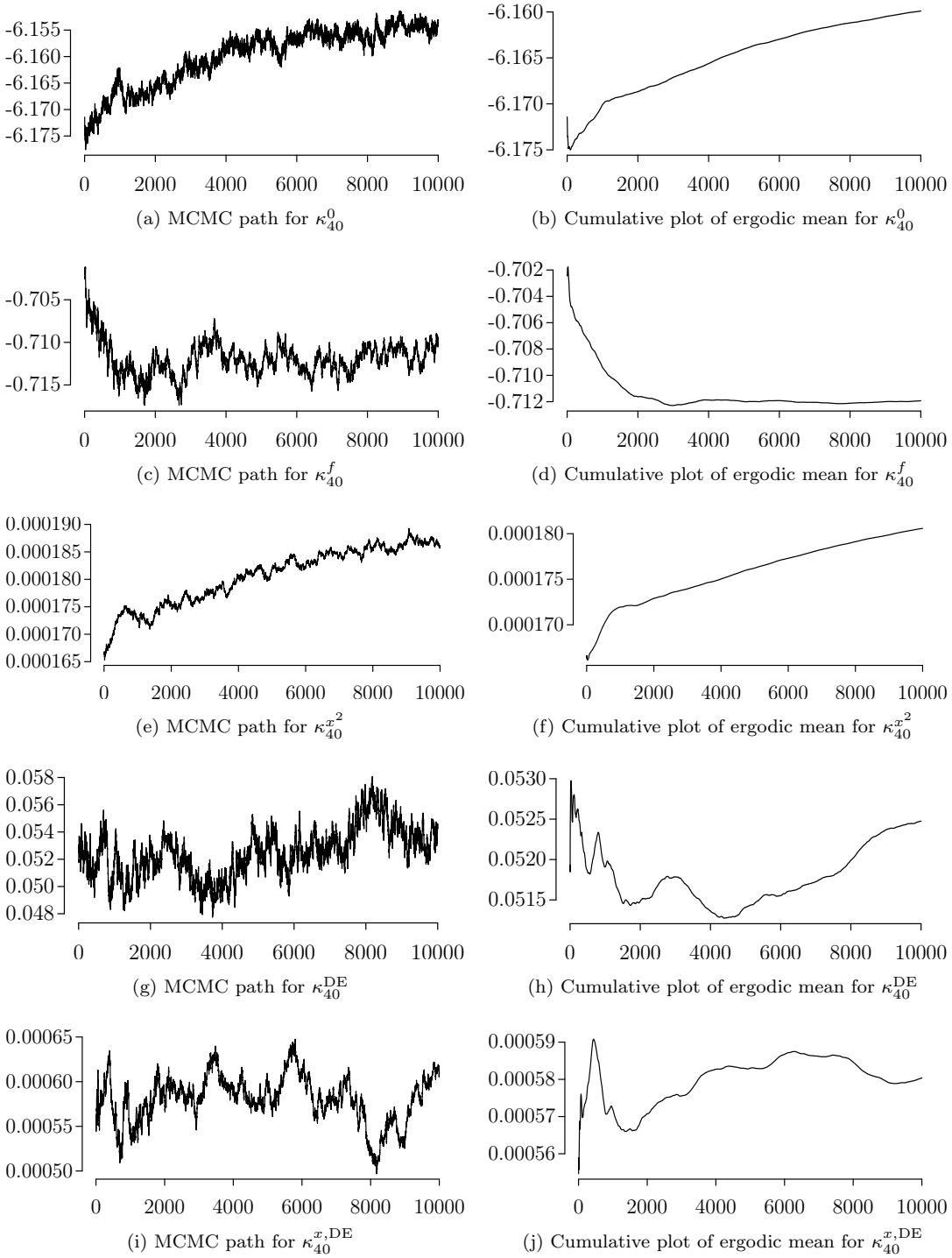
(h) Cumulative plot of ergodic mean for $\kappa_{20}^{\mathrm{DE}}$

(i) MCMC path for $\kappa_{20}^{x,\mathrm{DE}}$

(j) Cumulative plot of ergodic mean for $\kappa_{20}^{x,\mathrm{DE}}$

Figure 4.9: Convergence diagnostics for κ_{40}

Left panel: Realised marginal paths of selected entries of κ_{40} after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means.



(a) MCMC path for κ_{40}^{0}

(b) Cumulative plot of ergodic mean for κ_{40}^{0}

(c) MCMC path for κ_{40}^{f}

(d) Cumulative plot of ergodic mean for κ_{40}^{f}

(e) MCMC path for $\kappa_{40}^{x^2}$

(f) Cumulative plot of ergodic mean for $\kappa_{40}^{x^2}$

(g) MCMC path for $\kappa_{40}^{\mathrm{DE}}$

(h) Cumulative plot of ergodic mean for $\kappa_{40}^{\mathrm{DE}}$

(i) MCMC path for $\kappa_{40}^{x,\mathrm{DE}}$

(j) Cumulative plot of ergodic mean for $\kappa_{40}^{x,\mathrm{DE}}$

The graphical assessment of convergence among the hyperparameters starts with the plots shown in Figures 4.5 to 4.7. The left panels in each of the figures indicate that all marginal time series start within the range of the eventual posterior distributions, and no serial autocorrelation or changes in variability seem to be present after thinning. Due to the large amount of noise, the right panels are evaluated for a better understanding of stability and determination of burn-in periods. In general, all marginal hyperparameters seem to have sufficiently converged, see for example Π_{77}, or nearly converged as for, e.g., Γ_{21}, where a slow upward movement of the ergodic mean seems to be still ongoing. A sample of the Markov chain after a burn-in period of, say, 500,000 values would already give a promisingly good approximation of the marginal posterior distributions. The length of such a sample can further be expected to yield a reasonably valid approximation of the joint posterior distribution; however, a longer run of the MCMC algorithm would naturally be favourable. Plots of selected entries in Ω, as shown in Figure 4.3, draw a similar picture, but with increasing concern regarding insufficient length of the Markov chain to exhibit undoubted convergence. In particular, corresponding plots for Ω^{-1} in Figure 4.4 such as for, e.g., Ω_{44}^{-1} reveal the necessity for more iterations to establish full convergence and larger samples to compensate the higher magnitudes of serial autocorrelation and changes in variability. With the purpose of illustrating several examples in this work, this MCMC algorithm is not further conducted, and a burn-in period of 750,000 values will be used in the later course of this case study to obtain rough approximations of the posterior distribution of all hyperparameters. However, the MCMC estimation should be run far longer to obtain precise results in any real-life application, since joint convergence in this high-dimensional set-up intuitively requires even more time to be reached. Plots for various other selected marginal hyperparameters and parameters for all choices of k and r were also analysed by the author supporting previous conclusions, but cannot be discussed here in detail.

With the choice of a burn-in period of 750,000 iterations, in addition to the previously discussed graphical output for each hyperparameter, the pivotal comparison of ergodic means

$$\frac{\bar{t}_a - \bar{t}_b}{\sqrt{\widehat{\mathrm{Var}}(\bar{t}_a - \bar{t}_b)}}$$

is computed as explained in Appendix B.4, where $\bar{t}_a$ and $\bar{t}_b$ denote the sample means at the iterations 750,000–850,000 and 900,000–1,000,000 after thinning, respectively. Based on plots of the autocorrelation functions (not shown), the thinned samples of the hyperparameters are satisfactorily pseudo-independent such that the variance of the difference in sample means is estimated via the sum of variance estimators for iid sample means. The more hyperparameters have already converged, the lower are the rejection rates under the approximate standard normal distribution. The null hypothesis of convergence is rejected for 46.4% of the entries of Ω, 69.2% for Ω^{-1}, 0.6% for Π, 15.4% for ϕ, and 5.9% for Γ on a nominal confidence level of $\alpha^* = 2.47 \cdot 10^{-5}$, chosen such that the Bonferroni method yields a global type I error rate of $\alpha = 0.05$. These numbers support previous findings in that a longer MCMC run is required for the Markov chain to reach an equilibrium for all entries in the covariance matrix and, to some extent, ϕ. Rejection rates for the hyperparameters Π and Γ do not lead to evidence against convergence.

The hierarchical structure of the BMPMP model naturally requires an extensively long duration for the posterior sample to yield a sufficient sample for the parameters in the CBD model. Hence, turning to the convergence plots for $\mathcal{K}$ in Figures 4.8 and 4.9, it is not surprising to see less satisfying results than what was observed for the hyperparameters. Obvious ongoing trends indicate that the Markov chain has not yet converged for the majority of parameters. This is supported by a rejection of the null hypothesis of convergence for 89.3% of parameters in $\mathcal{K}$ when applying above test with $\alpha^* = 4.93 \cdot 10^{-5}$. The plots for κ^f_{20} further warn against false conclusions from the possibility of long-lasting metastability. In addition, the dependence in the simultaneous Bayesian estimation of the VECM let starting values be out of range and lead to higher autocorrelation. Altogether, the analysis of these and similar plots for other parameters and model set-ups clearly shows that the full posterior distribution for the joint set of $\mathcal{H}$ and $\mathcal{K}$ is not yet available. In conclusion, the marginal posterior distribution of the set of hyperparameters may be affected through the simultaneousness of the estimation procedure. However, previous results regarding $\mathcal{H}$ remain fruitful in that this study can work with a rough approximation. More importantly, a sufficiently large sample obtained in a reasonably long run of the MCMC algorithm can at least give a true sample of the marginal posterior distribution for the set of hyperparameters. Indeed, for the desired purpose of the BMPMP model, namely forecasting mortality trends into the future, the hierarchical design of the model only requires the posterior distribution of

$\mathcal{H}$ to be able to simulate the posterior predictive distribution of $\mathcal{K}$ for future calendar years. By way of contrast, the immediate marginal posterior distribution of $\kappa_1, \ldots, \kappa_{40}$, on which the model is calibrated, need not be known in Bayesian hierarchical modelling. Non-convergence of the CBD parameters and, even more, the fact that convergence may not be reached in computationally feasible time is hence negligible, as long as convergence for the hyperparameters can be well-justified, which in turn appears viable.

Note that the convergence diagnostics for $\mathcal{K}$ already show that the posterior distribution will not include its starting values, i.e. the ML estimates of the CBD model. This is in line with findings by Czado et al. (2005) on their Bayesian modelling approach of the LC model and results from the different estimation methodologies: whereas the hierarchical framework is estimated simultaneously in Bayesian statistics, the frequentist estimates stem from a sequential procedure with the potential risk of incoherent results.

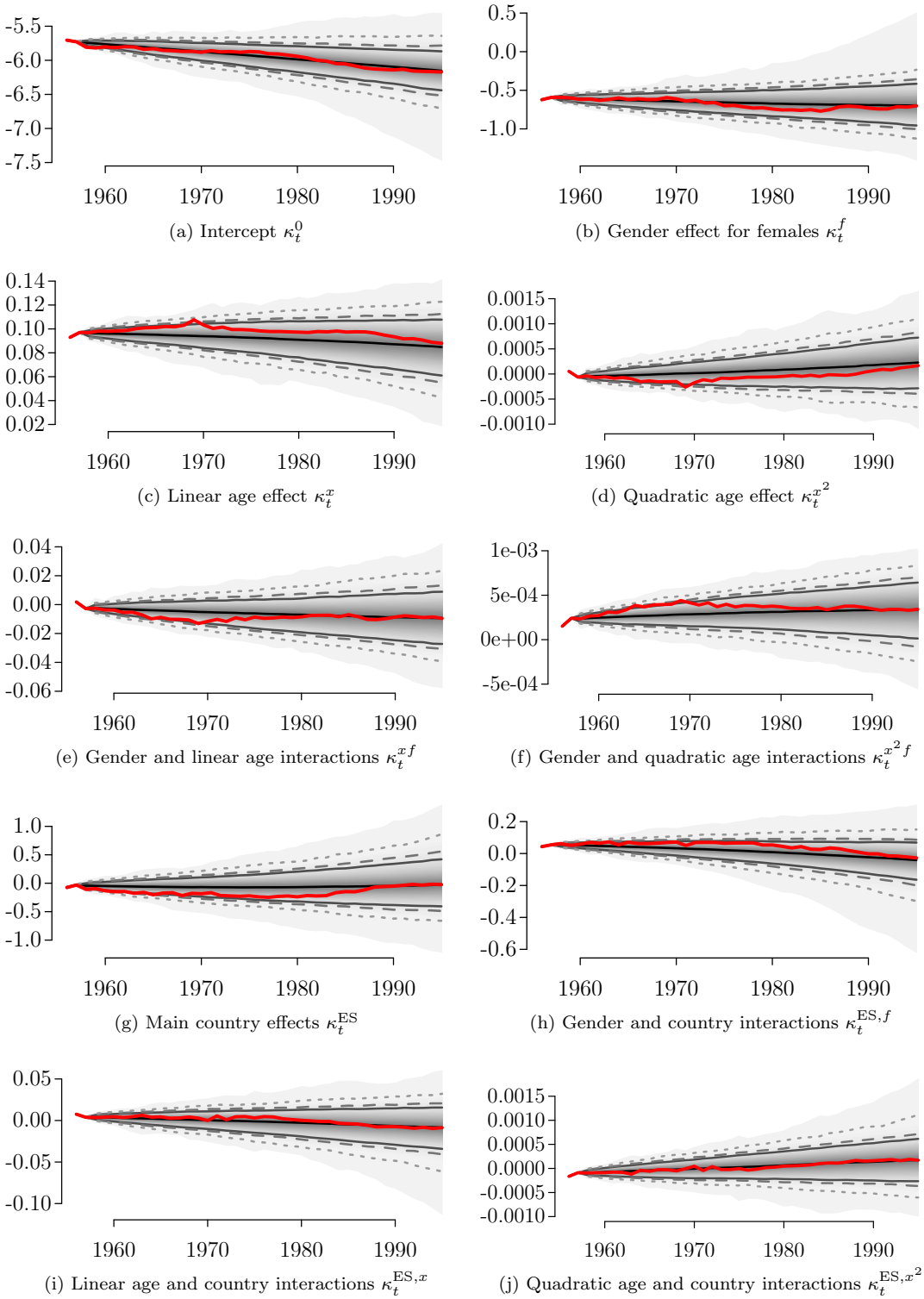
4.1.5 Posterior Predictive Checking

Under the conclusion that the MCMC algorithm yields at least a rough, if not good, approximation of the hyperparameters in $\mathcal{H}$, the BMPMP model itself must now be checked for the general ability to fit the observed data in $\mathcal{D}$ and $\mathcal{E}$ well. If there is evidence of lack of fit, estimators for model parameters are biased, and any inference, particularly forecasts of mortality into the future, is invalid. Diagnostics of goodness-of-fit in this Bayesian framework is done via posterior predictive checking, which comprises comparison of the observed quantities and the posterior distribution of replicated values through graphical tools or computation of Bayesian discrepancy tests and corresponding p-values. A general overview on diagnostic quantification techniques in Bayesian statistics is given in Appendix A.4.

As the Bayesian version of standard internal prediction validation, the hyperparameters $\mathcal{H}$ are first simulated from their posterior distribution, i.e. values are sampled from the MCMC output after discarding the burn-in period. In the following, the entity of all 2,500 thinned realisations after the burn-in period is used. Then, starting with analogously sampled realisations from the posterior of κ_1 and κ_2, for each realisation of $\mathcal{H}$, the parameters in $\mathcal{K}$ are simulated via the VECM, thereby giving a sample of the posterior predictive distribution of the CBD parameters over the calibration period.

Figure 4.10: Posterior predictive checking for $\mathcal{K}$

Shown are fancharts of selected marginal posterior predictive distributions for κ_t as time series in t for $1956, \ldots, 1995$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. The country-specific parameters are illustrated for the example of Spain. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Red lines denote ML estimates (starting values).



(a) Intercept κ_t^0

(b) Gender effect for females κ_t^f

(c) Linear age effect κ_t^x

(d) Quadratic age effect $\kappa_t^{x^2}$

(e) Gender and linear age interactions κ_t^{xf}

(f) Gender and quadratic age interactions $\kappa_t^{x^2 f}$

(g) Main country effects κ_t^{ES}

(h) Gender and country interactions $\kappa_t^{\mathrm{ES},f}$

(i) Linear age and country interactions $\kappa_t^{\mathrm{ES},x}$

(j) Quadratic age and country interactions $\kappa_t^{\mathrm{ES},x^2}$

Figure 4.11: Posterior predictive checking for η_{xpgt} for the age of 60

Shown are fancharts of marginal posterior predictive distributions for η_{xpgt} with $x = 60$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1956, \ldots, 1995$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Green lines denote crude estimates from observed data.

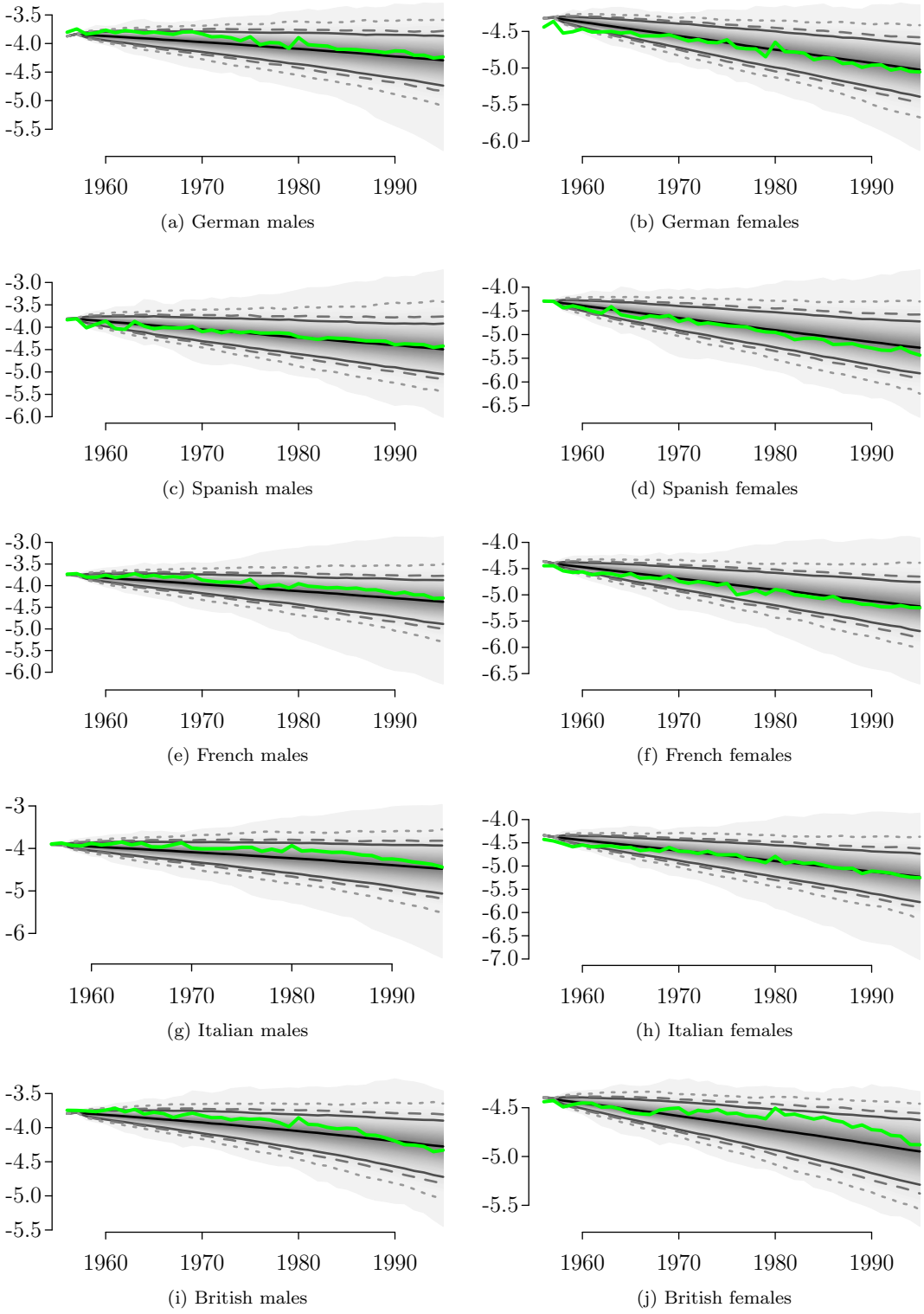
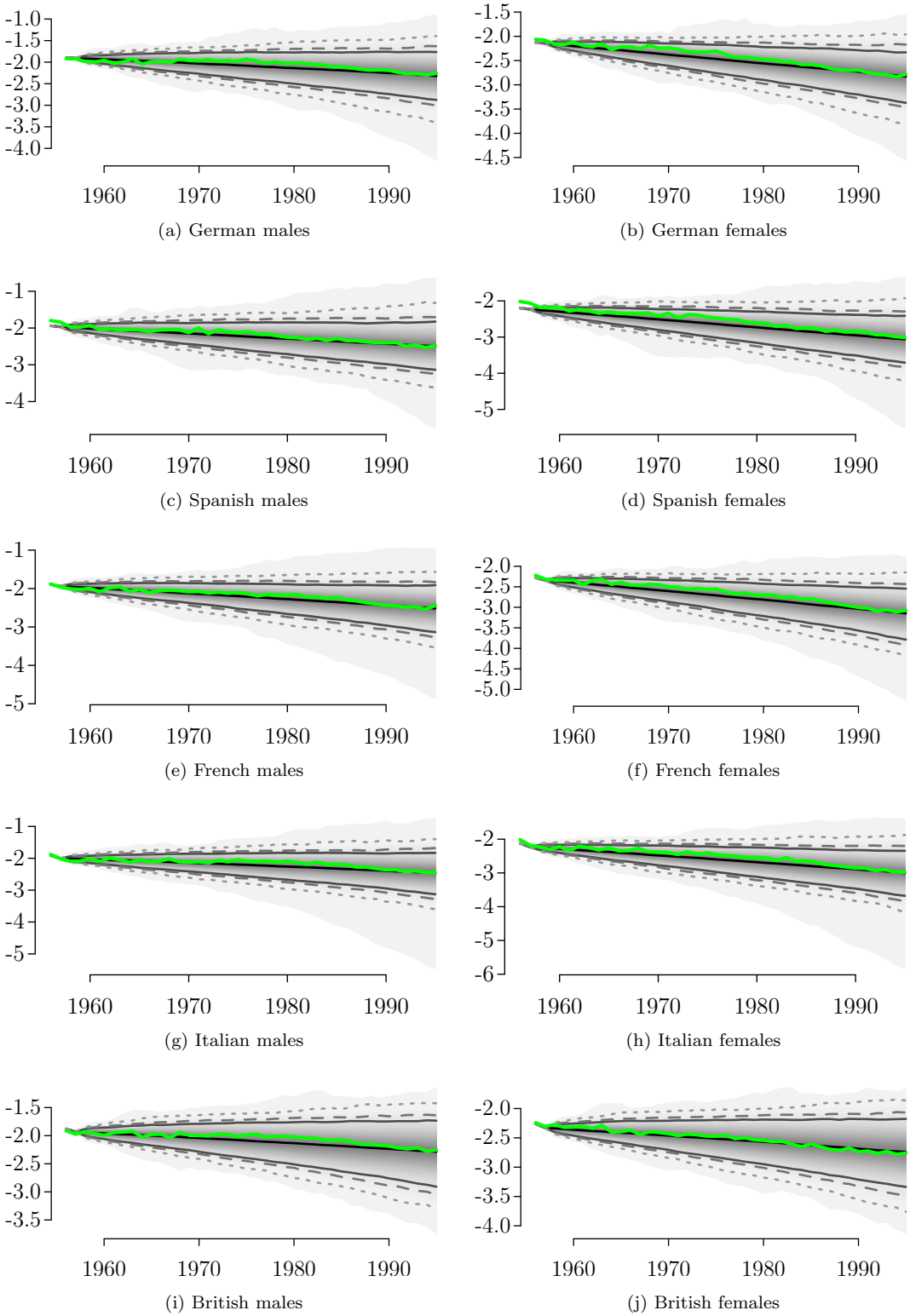


Figure 4.12: Posterior predictive checking for η_{xpgt} for the age of 80

Shown are fancharts of marginal posterior predictive distributions for η_{xpgt} with $x = 80$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1956, \ldots, 1995$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Green lines denote crude estimates from observed data.



(a) German males

(b) German females

(c) Spanish males

(d) Spanish females

(e) French males

(f) French females

(g) Italian males

(h) Italian females

(i) British males

(j) British females

Figure 4.10 shows the evaluation of selected marginal time series in $\mathcal{K}$. Since these parameters are latent, their posterior predictive distribution is checked for inclusion of the ML estimates, i.e. their starting values, which is the case on high credibility levels. In Figures 4.11 and 4.12, the resulting linear predictors η_{xpgt} as time series in t are compared to their crude estimates from the raw data for each combination of all five countries and both genders for the fixed values of $x = 60$ and $x = 80$, i.e. the ages 60 and 80. The graphical output clearly shows a very close agreement between the observed data and the medians of the posterior distributions. A major deviation can only be detected for British 60-year-old females, where the observed data are still found to lie within the pointwise 90% credibility interval. In fact, as the nominal credibility intervals for the posterior distribution appear to be wider than what they should be, based on the comparison with the crude estimates, these results exhibit a noticeable magnitude of optimism, which is commonly observed in internal validation analyses.

Although in the literature on other Bayesian mortality prediction models, the analysis of model fit is restricted to the assessment of graphical and tabular MCMC output, it is good statistical practice to conduct quantitative goodness-of-fit tests to assess the appropriateness of the chosen model. For example, since the sum of all mutually independent D_{xpgt} is again Poisson distributed with the mean being the sum of all individual means, the standardised residual for the overall death count, i.e.

$$T(\mathcal{D}, \mathcal{E}, \mathcal{K}) = \frac{\sum_{x=40}^{100} \sum_{p \in \mathcal{P}} \sum_{g \in \{m,f\}} \sum_{t=1}^{40} \left(D_{xpgt} - \log\left(1 + \exp(\eta_{xpgt})\right) E_{xpgt}\right)}{\sqrt{\sum_{x=40}^{100} \sum_{p \in \mathcal{P}} \sum_{g \in \{m,f\}} \sum_{t=1}^{40} \log\left(1 + \exp(\eta_{xpgt})\right) E_{xpgt}}},$$

is used as a summary test statistic. For each set of posterior predictive values $\mathcal{K}$, it is evaluated at the observed number of deaths and exposure-to-risk. The reference distribution is the posterior distribution of this test statistic obtained by replacing the observed D_{xpgt} by their posterior predictive realisations D^{rep}_{xpgt}. The Bayesian posterior predictive p-value corresponds to the posterior probability of observing more extreme outcomes for the replicated than for the historical data, i.e. $P(T(\mathcal{D}^{rep}, \mathcal{E}, \mathcal{K}) \geq T(\mathcal{D}, \mathcal{E}, \mathcal{K}) \mid \mathcal{D}, \mathcal{E})$. Indeed, the p-value is computed to be 0.41 and, hence, not close to either extreme value 0 or 1. On the other hand, conducting this test with the immediate posterior values for $\mathcal{K}$ leads to a p-value of 0.95, indicating lack of fit when the model's main goal was estimation of historical values. In fact, Czado et al. (2005) point out that any type of –

what they call – "robust", i.e. hierarchical and parsimonious, mortality projection model generally shows a poor fit. Due to the large sample size of $5 \cdot 2 \cdot 61 \cdot 40 = 24{,}400$ death counts, it is not surprising to see statistical tests leading to rejection of the comparably simplistic two-level hierarchical BMPMP model with its simplifying Poisson assumption. For further illustration, the general omnibus test from (A.1) in Appendix A.4 is computed, which in this case study becomes

$$\begin{aligned} T(\mathcal{D}, \mathcal{E}, \mathcal{K}) &= \sum_{x=40}^{100} \sum_{p \in \mathcal{P}} \sum_{g \in \{m,f\}} \sum_{t=1}^{40} \frac{\left(D_{xpgt} - \mathrm{E}(D_{xpgt} \mid E_{xpgt}, \mathcal{K})\right)^2}{\mathrm{Var}(D_{xpgt} \mid E_{xpgt}, \mathcal{K})} \\ &= \sum_{x=40}^{100} \sum_{p \in \mathcal{P}} \sum_{g \in \{m,f\}} \sum_{t=1}^{40} \frac{\left(D_{xpgt} - \log\left(1 + \exp(\eta_{xpgt})\right) E_{xpgt}\right)^2}{\log\left(1 + \exp(\eta_{xpgt})\right) E_{xpgt}}, \end{aligned}$$

using the properties of the Poisson distribution. Plugging in the posterior predictive or immediate posterior values $\mathcal{K}$, this classical goodness-of-fit test statistic is evaluated at the observed number of deaths and, to obtain the reference distribution, at the posterior predictive realisations for D_{xpgt}. The Bayesian p-value corresponds to the posterior excess probability of $T(\mathcal{D}^{rep}, \mathcal{E}, \mathcal{K})$ compared to $T(\mathcal{D}, \mathcal{E}, \mathcal{K})$ and turns out to be zero for either choice of $\mathcal{K}$, i.e. the omnibus test yields the strongest evidence possible against the BMPMP model. However, based on the previously discussed graphical output, by similarity to other models in mortality forecasting, it is concluded that the statistical evidence against the model should not be used to rule out its obvious ability to produce realistic stochastic mortality forecasts.

The previous discussion yields a Bayesian posterior check for the entity of the BMPMP model. Since the marginal posterior distribution of the parameters $\mathcal{K}$ was used in particular cases, it can be further understood as a specific check of the CBD model. It may be of additional interest to the analyst how the VECM, i.e. the second stage in the two-level hierarchy, performs on its own. However, model diagnostics, in light of the well-understood techniques described in Appendix A.4 for the separate VECM, cannot be performed, since the underlying data are the latent parameters from the CBD model and subject to change. Classical statistical tests on residual autocovariance or the normality of the white noise, such as the Portmanteau, Lagrange-Multiplier, and Lomnicki-Jarque-Bera tests described in Appendix C.6, are hence not available for posterior predictive checking, as the data do not remain fixed. However, to get some insight

in the performance of the VECM, the joint posterior distribution for the residual auto- and cross-correlation matrix is derived from the full joint distribution on both $\mathcal{H}$ and $\mathcal{K}$. A good model fit would lead to posterior residual autocorrelations scattering around zero, but the analysis (not shown) reveals that marginal posterior distributions for these values usually do not contain zero. This indication of lack of fit is not regarded conclusive due to the latent behaviour of $\mathcal{K}$, its non-convergence in the MCMC algorithm, and the general problems with robust mortality prediction models. Details and further investigation of the VECM is therefore omitted here.

4.1.6 External Validation

With the choice of the calibration window 1956–1995, an out-of-sample dataset with death counts for the years 1996–2009 is available, such that the goodness of forecasting of the BMPMP model can be assessed through the Bayesian counterpart of external validation, thereby eliminating the optimism seen in the results from the previous section. For illustration purposes, the forecast period is further extended until 2014. Again, the posterior predictive distribution for the future values $\kappa_t, t = 41, \ldots, 59$, is derived through sampling realisations from the posterior for $\mathcal{H}$ and subsequent simulation of κ_t using the VECM. The posterior distribution for the hyperparameters is approximated with the same MCMC output as before. Starting values for the time series at the now initial years 1994 and 1995 are drawn from the posterior distribution of $(\kappa_{39}, \kappa_{40})$. Figure 4.13 shows the evaluation of the marginal time series from Figure 4.10 until 2014, along with its ML estimates over the calibration period for comparison. Figures 4.14 and 4.15 depict the resulting linear predictors η_{xpgt} as time series in t with the same choices for x as before. Here, for all plots, both the historical and future crude estimates until 2009 from the raw data are given for evaluation of the goodness of prediction.

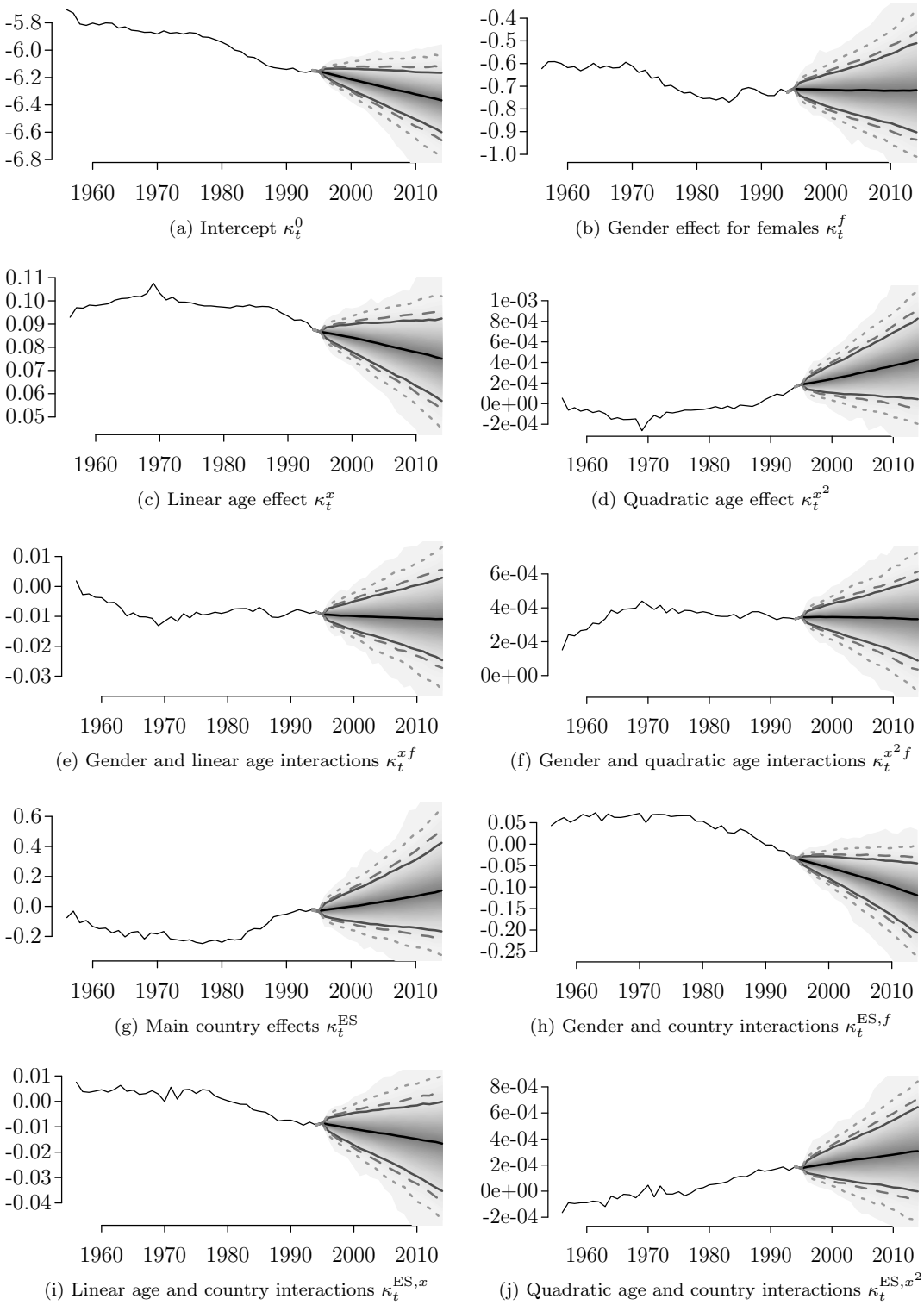
The plots in Figure 4.13 display a generally well-behaving forecast nature for the latent time series in $\mathcal{K}$. Historical trends are found to continue over the years 1996–2014 with linearly increasing credibility bands, i.e. a continuously growing domain for the probability mass of all parameters to account for future uncertainties. The slopes of future realisations in plots for, e.g., the intercept (a), linear age effects (c), and the parameters for Spain (g–j) reveal that the Bayesian estimates do not prolong the most current trend, but rather exhibit all kinds of developments that the time series has experienced over

the calibration period. For example, the pace of decline for the time series of $\kappa_t^{\mathrm{ES},f}$ in (h) seems to be lower than what would be expected based on the development over the course of the most recent twenty years. The smaller magnitude of the slope is explained by the stable, if not increasing, pattern of the time series for the years 1956–1975. These results signalise the apparently specific feature of the BMPMP model to be strongly influenced by the entire calibration period. Strong sensitivity towards the calibration period is a common concern found for most stochastic mortality projection models; see Cairns et al. (2009) for a general survey on models in the LC and CBD frameworks. As a consequence, the time before the regime change, discovered in the discussion regarding the starting values in Section 4.1.3, should be excluded for an improvement in mortality projections.

Figures 4.14 and 4.15 show that the linear predictors η_{xpgt} for both ages $x = 60$ and $x = 80$ carry over the characteristics of the forecasts for the marginal time series in $\mathcal{K}$. The linearly increasing fans in all plots correspond to the desired nature of continuously growing uncertainty in future mortality predictions. The overly optimistic credibility bands in the internal validation have become more realistic such that their width is indeed required to cover crude estimates in most of the cases; for instance, see the plots for German females with a noticeable amount of noise within the 90% credibility interval. However, except for Spain and France, mortality rates for males are overestimated for Germany, Italy, and – most notably – the United Kingdom, where crude estimates of the linear predictor for $x = 60$ lie outside the lower 90% credibility boundary. Forecast properties of the linear predictors resemble the previous finding for the marginal time series regarding the sensitivity towards the calibration period. The increased pace in the decline of linear predictors seen for the aforementioned countries suggest a calibration period that excludes the years until 1980. It is noteworthy, however, that the BMPMP model is still capable to capture major changes in population-specific mortality trends through its tails, an important feature for the model's prediction quality, particularly when future outcomes are indeed unknown.

Figure 4.13: External validation for $\mathcal{K}$

Shown are fancharts of selected marginal posterior predictive distributions for future κ_t as time series in t for $1996, \ldots, 2014$, along with ML estimates for κ_t for $1956, \ldots, 1995$ for the same MCMC output as in Figure 4.10. The country-specific parameters are illustrated for the example of Spain. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) Intercept κ_t^0

(b) Gender effect for females κ_t^f

(c) Linear age effect κ_t^x

(d) Quadratic age effect $\kappa_t^{x^2}$

(e) Gender and linear age interactions κ_t^{xf}

(f) Gender and quadratic age interactions $\kappa_t^{x^2 f}$

(g) Main country effects κ_t^{ES}

(h) Gender and country interactions $\kappa_t^{\mathrm{ES},f}$

(i) Linear age and country interactions $\kappa_t^{\mathrm{ES},x}$

(j) Quadratic age and country interactions $\kappa_t^{\mathrm{ES},x^2}$

Figure 4.14: External validation for η_{xpgt} for the age of 60

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 60$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1996, \ldots, 2014$, along with crude estimates for η_{xpgt} for $1956, \ldots, 2009$ from observed data (green lines) for the same MCMC output as in Figure 4.11. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.

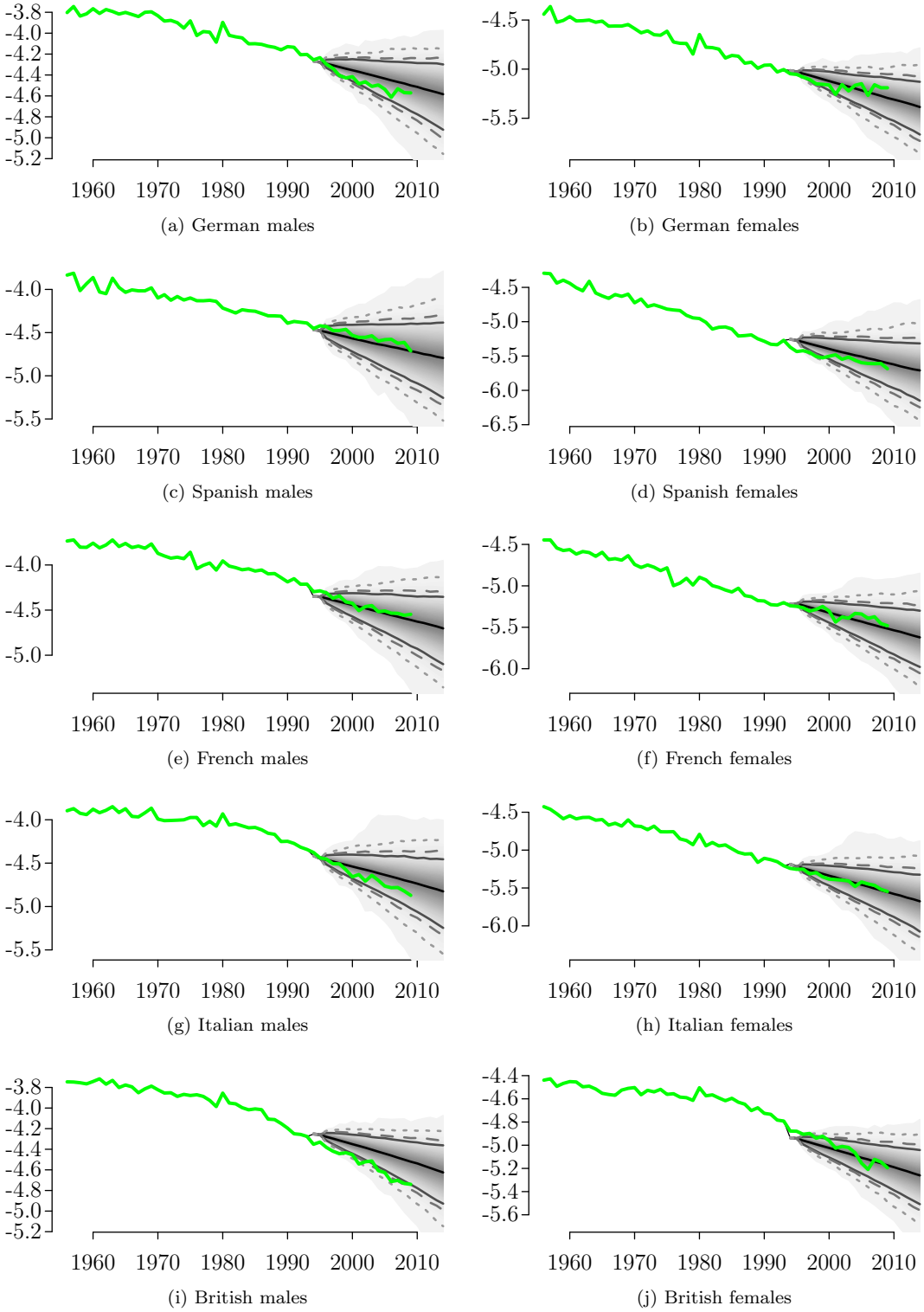
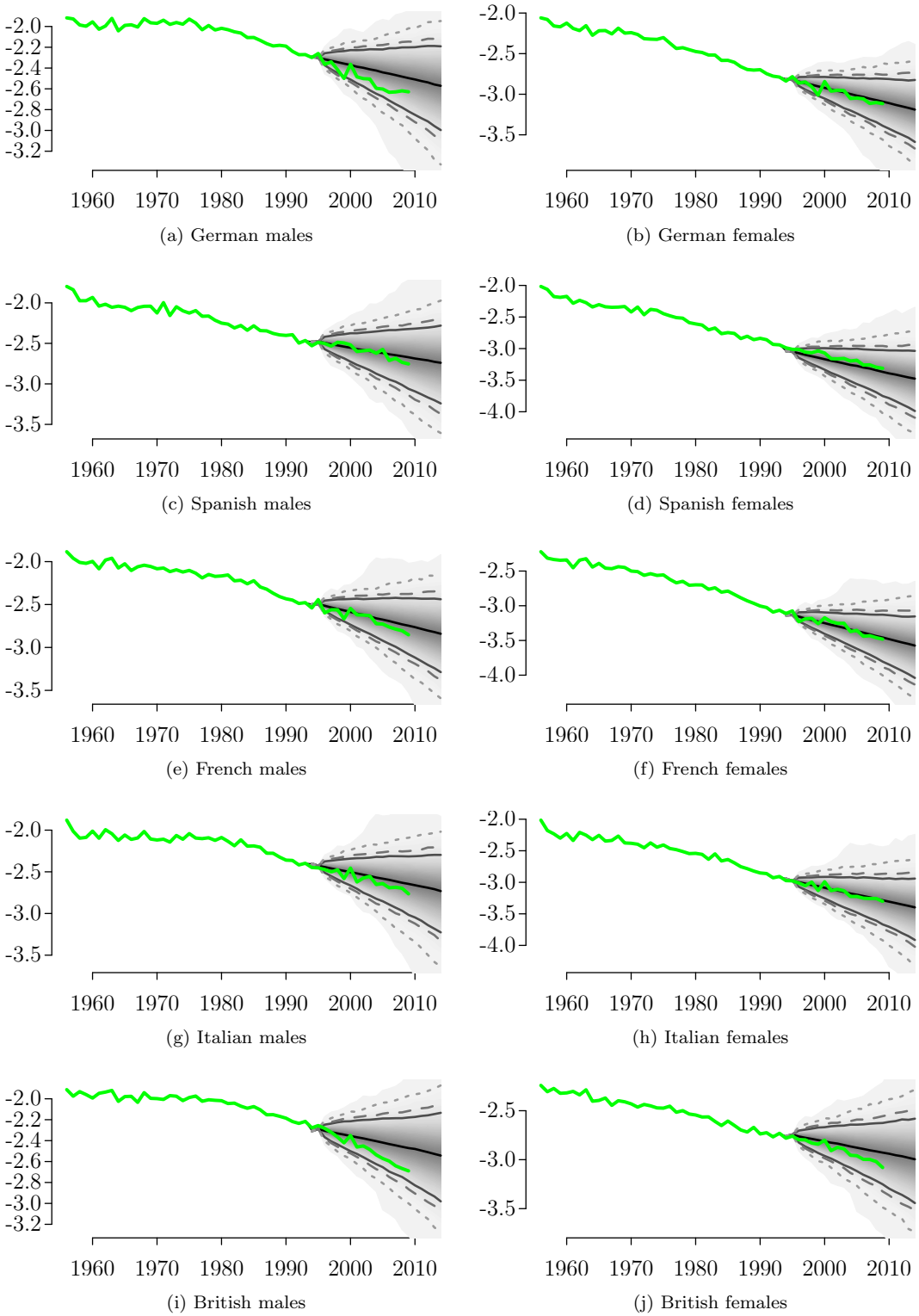


Figure 4.15: External validation for η_{xpgt} for the age of 80

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 80$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1996, \ldots, 2014$, along with crude estimates for η_{xpgt} for $1956, \ldots, 2009$ from observed data (green lines) for the same MCMC output as in Figure 4.12. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) German males (b) German females

(c) Spanish males (d) Spanish females

(e) French males (f) French females

(g) Italian males (h) Italian females

(i) British males (j) British females

4.1.7 Change of Calibration Period to 1981–2009

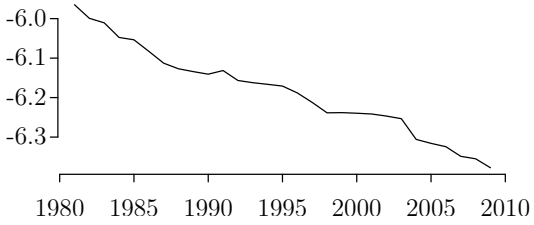
The previous sections have shown that the MCMC algorithm has not fully converged for the CBD parameters and, to minor extent, for the hyperparameters in $\mathcal{H}$, and that statistical tests lead to rejection of a good model fit. However, based on the posterior predictive distribution for simulated future mortality rates, the BMPMP model with rough posterior approximations for the hyperparameters produces good prediction results with the desired properties of quantifying a linearly increasing trend in uncertainty. As common in stochastic mortality prediction, sensitivity towards the choice of the calibration period was discovered, such that in the particular example above, it appears reasonable to restrict the calibration window to the years after 1980. For a sufficiently long history and for the most current application of mortality forecasts, the calibration window is chosen to be 1981–2009, i.e. including most recent data. Further details of the BMPMP model's outcome will be discussed in the course of the updated model. Prior to this, convergence and model diagnostics will be concisely re-evaluated.

Figures 4.16 to 4.29 contain the plots from the previous discussion replicated for the model in (4.1) with $t = 1, \ldots, 29$, representing the calibration window 1981–2009. All convergence plots depict the same hyperparameters from before, and for the CBD parameter level, the selected marginal MCMC output is made for κ_{15} and κ_{29}. Since data are only available up to the end of the calibration period, i.e. the calendar year 2009, external validation cannot be performed. However, plots of forecast until the year of 2100 are provided for a qualitative assessment of the model's future mortality predictions.

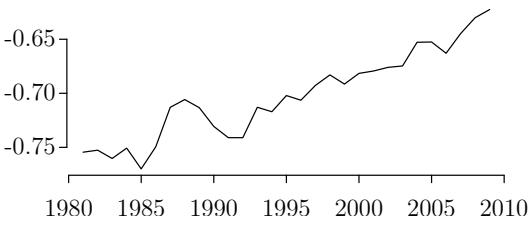
The starting values in Figure 4.16 reveal that, contrary to the previous discussion, no global regime change took place during the years 1981–2009. A point of reverse in the British and, to some minore extent, Spanish parameters at the early 1990s clearly explain why, in the earlier model, mortality predictions for these countries deviated from the best prediction given by the distribution's mean or median. More importantly, the quantification of population-specific deviations from the mean mortality development in the BMPMP model reveals that a convergent behaviour cannot be claimed for the period of 1981–2009. By contrast, differences remain rather stable or even diverge slightly. This result is in favour of the BMPMP model, which does not state any convergence hypotheses explicitly, but lets the data speak for themselves.

Figure 4.16: Starting values for $\mathcal{K}$

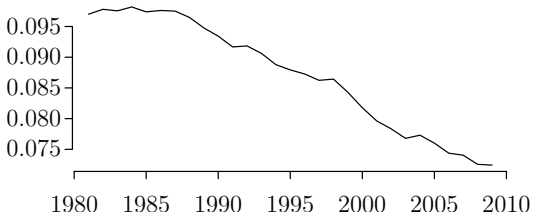
Shown are the ML estimates $\kappa_t^{(0)}$ of the CBD parameters as time series in t for $1981, \ldots, 2009$.



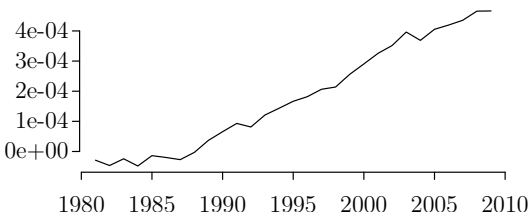
(a) Intercept $\kappa_t^{0(0)}$



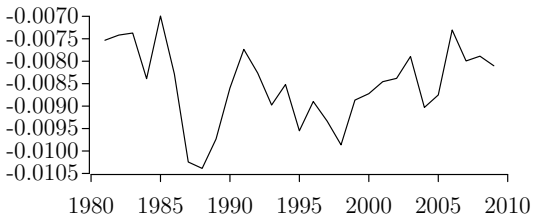
(b) Gender effect for females $\kappa_t^{f(0)}$



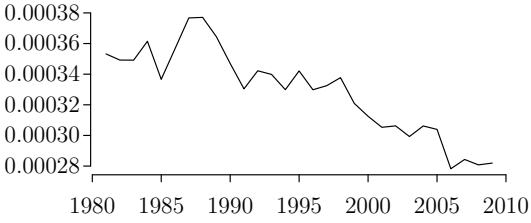
(c) Linear age effect $\kappa_t^{x(0)}$



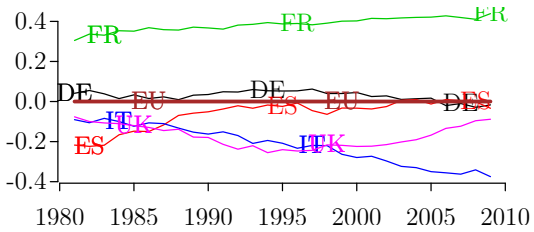
(d) Quadratic age effect $\kappa_t^{x^2(0)}$



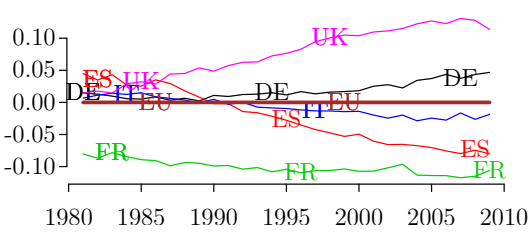
(e) Gender and linear age interactions $\kappa_t^{xf(0)}$



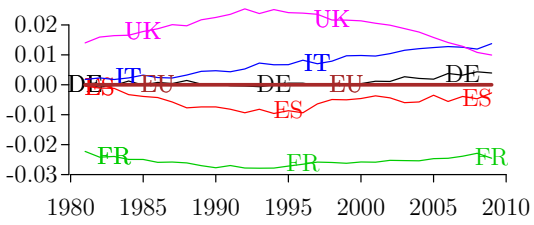
(f) Gender and quadratic age interactions $\kappa_t^{x^2f(0)}$



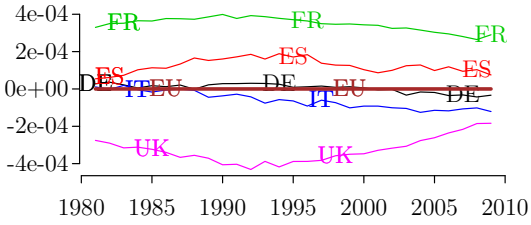
(g) Main country effects $\kappa_t^{p(0)}$



(h) Gender and country interactions $\kappa_t^{pf(0)}$



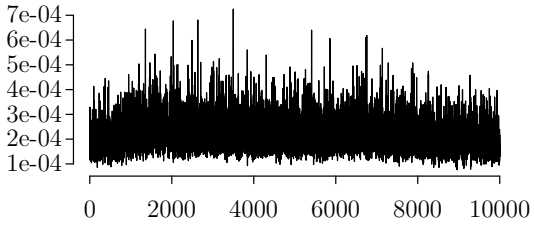
(i) Linear age and country interactions $\kappa_t^{px(0)}$



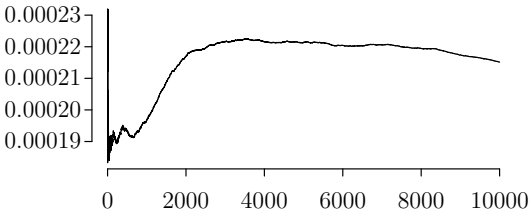
(j) Quadratic age and country interactions $\kappa_t^{px^2(0)}$

Figure 4.17: Convergence diagnostics for Ω

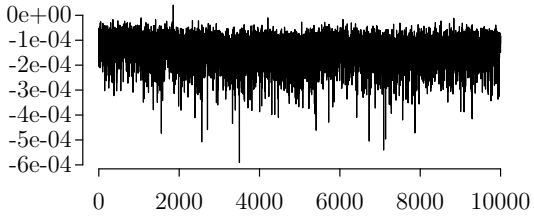
Left panel: Realised marginal paths of selected entries of Ω after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



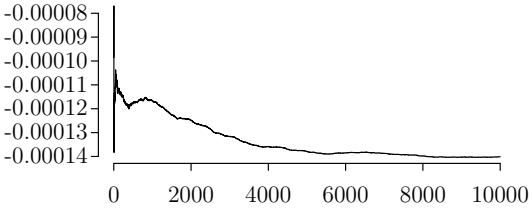
(a) MCMC path for Ω_{11}



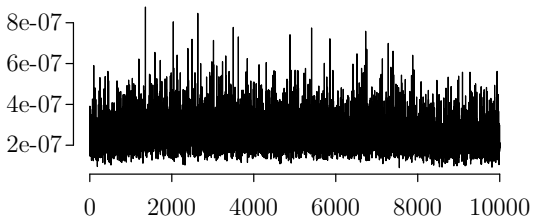
(b) Cumulative plot of ergodic mean for Ω_{11}



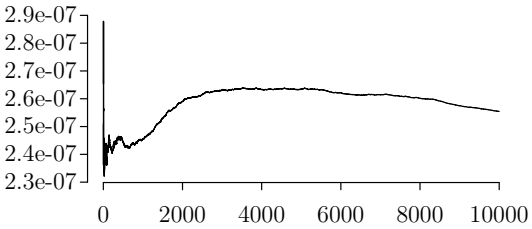
(c) MCMC path for Ω_{21}



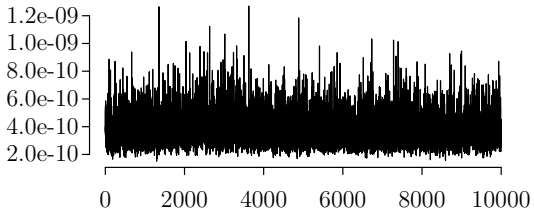
(d) Cumulative plot of ergodic mean for Ω_{21}



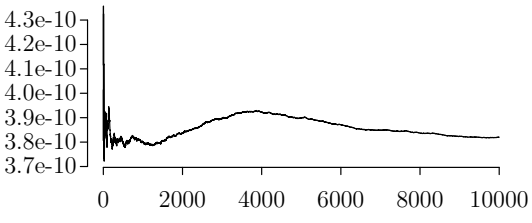
(e) MCMC path for Ω_{41}



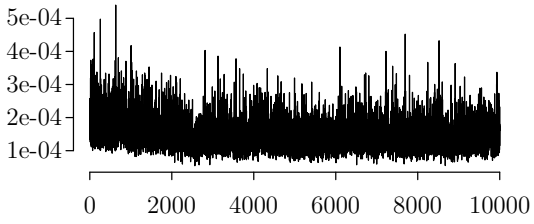
(f) Cumulative plot of ergodic mean for Ω_{41}



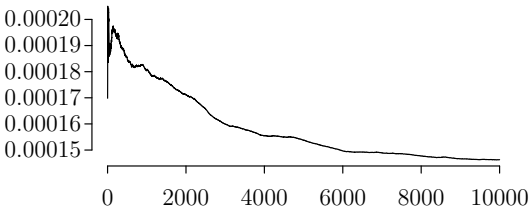
(g) MCMC path for Ω_{44}



(h) Cumulative plot of ergodic mean for Ω_{44}



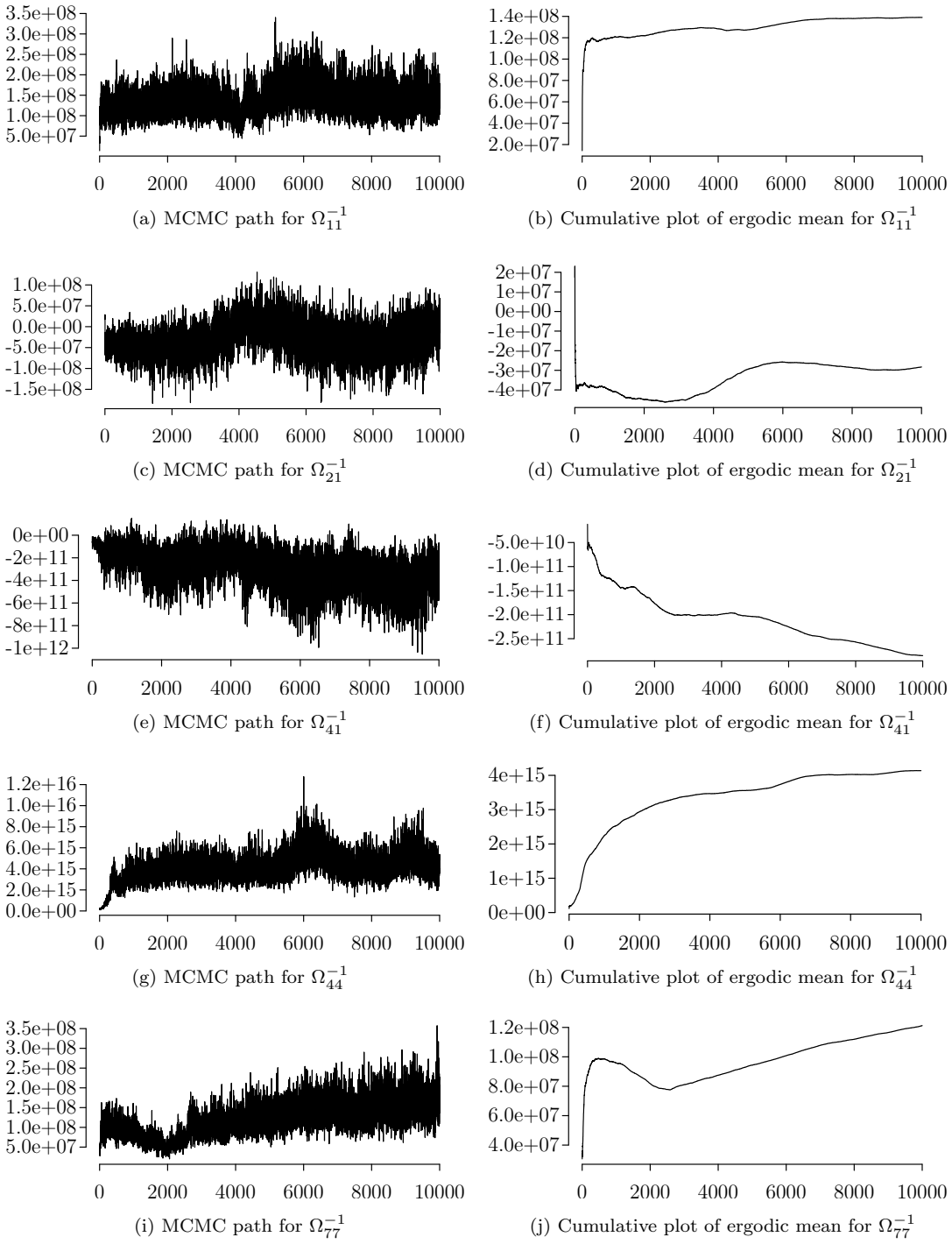
(i) MCMC path for Ω_{77}



(j) Cumulative plot of ergodic mean for Ω_{77}

Figure 4.18: Convergence diagnostics for Ω^{-1}

Left panel: Realised marginal paths of selected entries of Ω^{-1} after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



(a) MCMC path for Ω_{11}^{-1}

(b) Cumulative plot of ergodic mean for Ω_{11}^{-1}

(c) MCMC path for Ω_{21}^{-1}

(d) Cumulative plot of ergodic mean for Ω_{21}^{-1}

(e) MCMC path for Ω_{41}^{-1}

(f) Cumulative plot of ergodic mean for Ω_{41}^{-1}

(g) MCMC path for Ω_{44}^{-1}

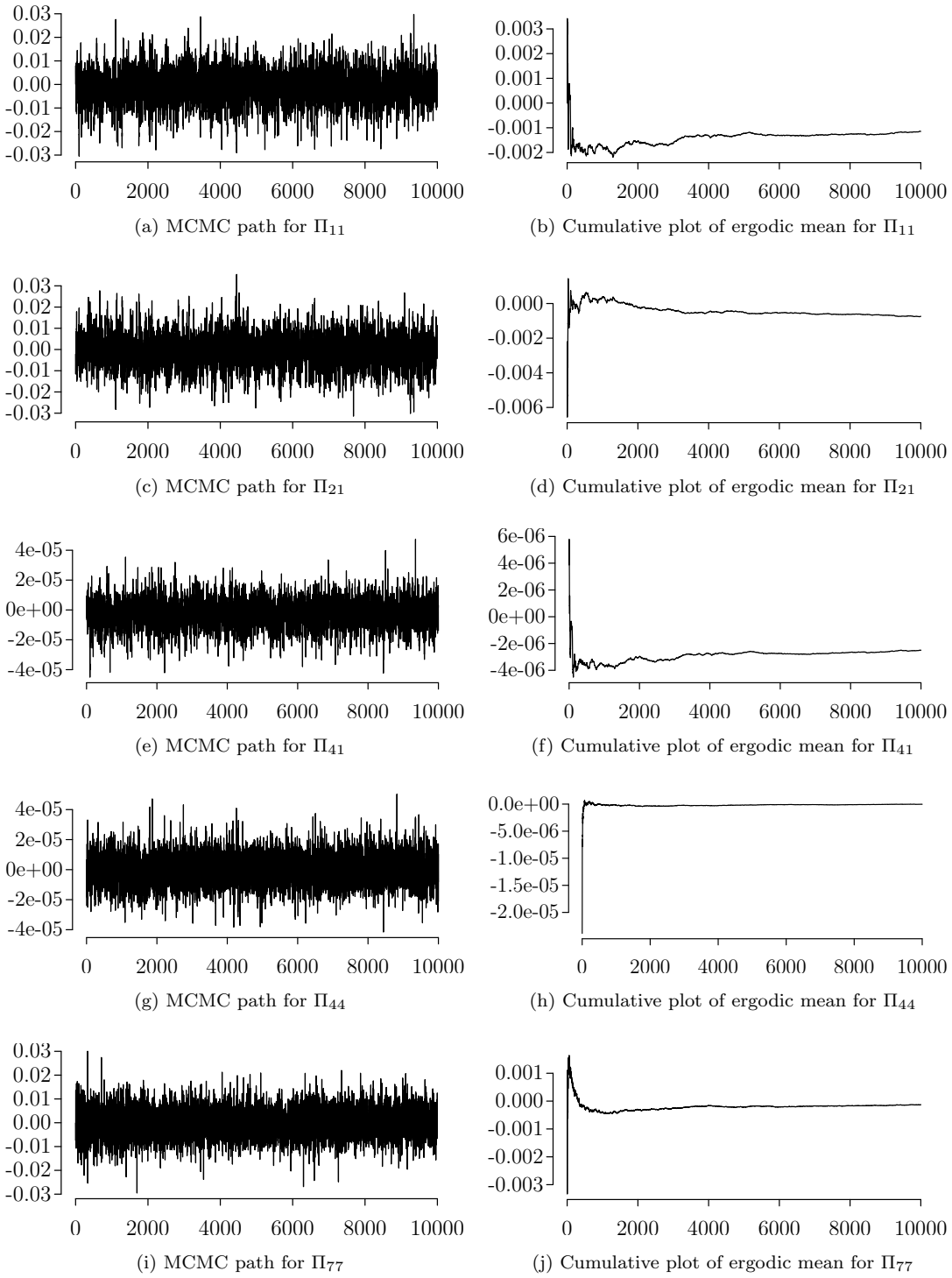
(h) Cumulative plot of ergodic mean for Ω_{44}^{-1}

(i) MCMC path for Ω_{77}^{-1}

(j) Cumulative plot of ergodic mean for Ω_{77}^{-1}

Figure 4.19: Convergence diagnostics for $\Pi = \alpha\beta'$

Left panel: Realised marginal paths of selected entries of Π after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



(a) MCMC path for Π_{11}

(b) Cumulative plot of ergodic mean for Π_{11}

(c) MCMC path for Π_{21}

(d) Cumulative plot of ergodic mean for Π_{21}

(e) MCMC path for Π_{41}

(f) Cumulative plot of ergodic mean for Π_{41}

(g) MCMC path for Π_{44}

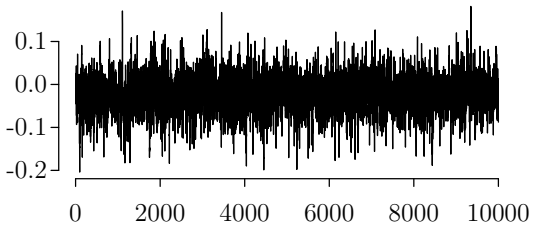
(h) Cumulative plot of ergodic mean for Π_{44}

(i) MCMC path for Π_{77}

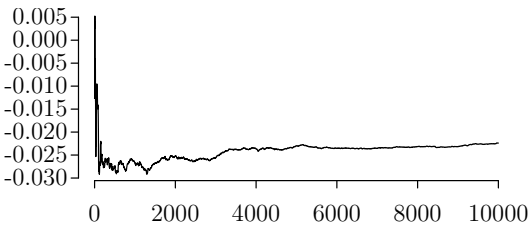
(j) Cumulative plot of ergodic mean for Π_{77}

Figure 4.20: Convergence diagnostics for ϕ

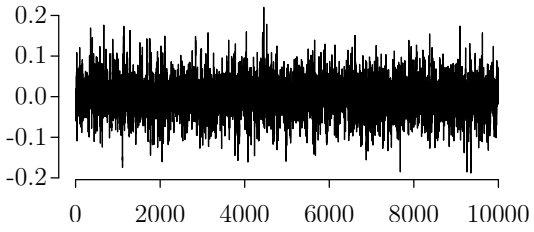
Left panel: Realised marginal paths of selected entries of ϕ after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$, $17 - \kappa_t^{x,\mathrm{DE}}$.



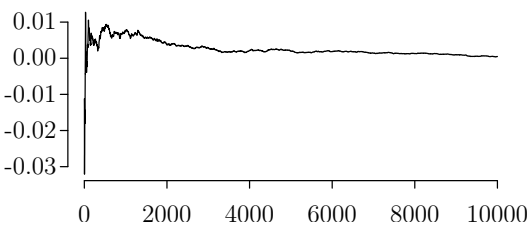
(a) MCMC path for ϕ_1



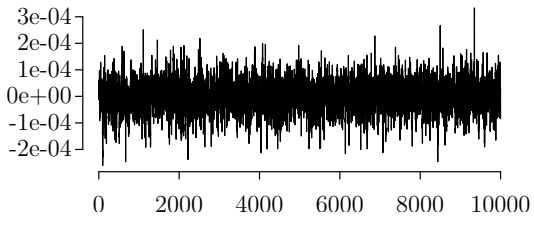
(b) Cumulative plot of ergodic mean for ϕ_1



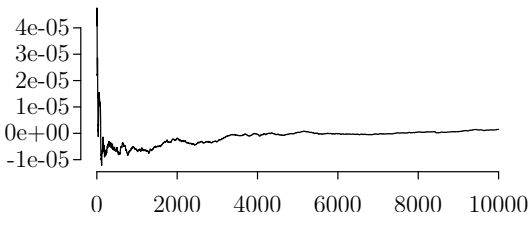
(c) MCMC path for ϕ_2



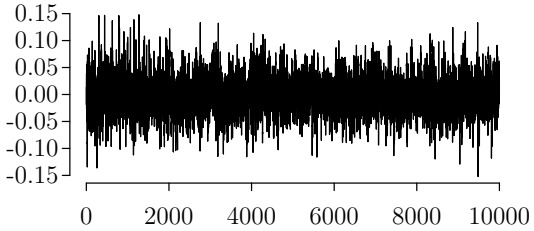
(d) Cumulative plot of ergodic mean for ϕ_2



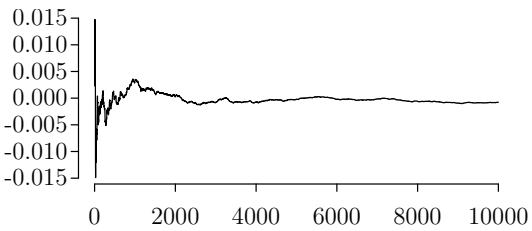
(e) MCMC path for ϕ_4



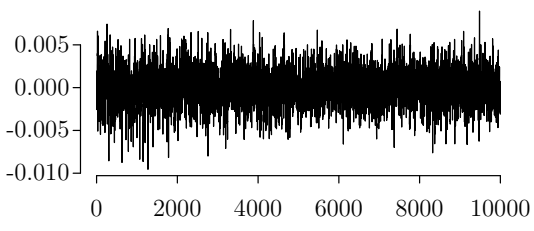
(f) Cumulative plot of ergodic mean for ϕ_4



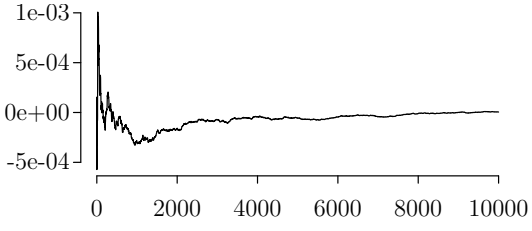
(g) MCMC path for ϕ_7



(h) Cumulative plot of ergodic mean for ϕ_7



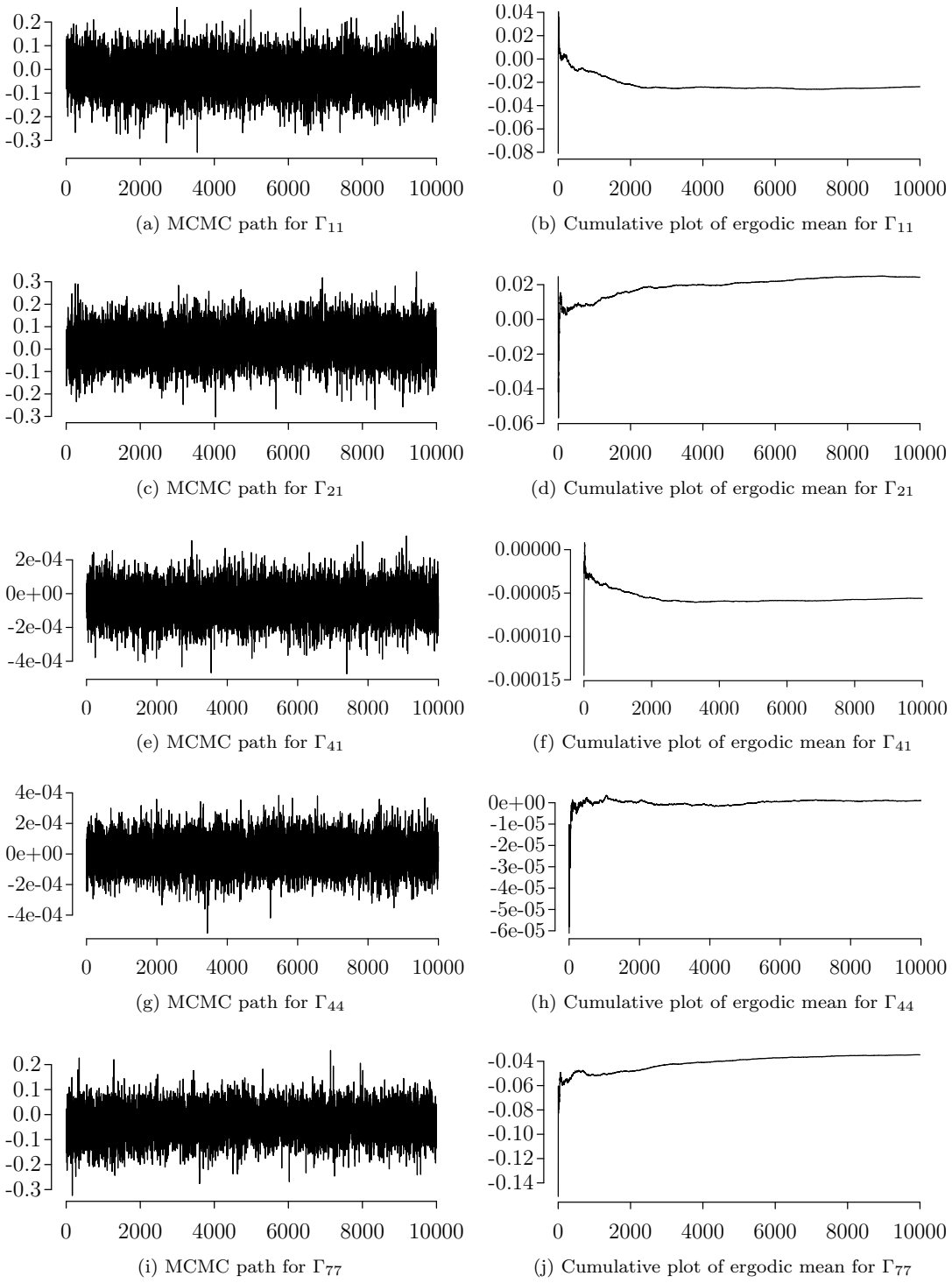
(i) MCMC path for ϕ_{17}



(j) Cumulative plot of ergodic mean for ϕ_{17}

Figure 4.21: Convergence diagnostics for $\Gamma = \Gamma_1$

Left panel: Realised marginal paths of selected entries of Γ after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means. *Indices:* Indices in VECM refer to the following marginal time series: $1 - \kappa_t^0$, $2 - \kappa_t^f$, $4 - \kappa_t^{x^2}$, $7 - \kappa_t^{\mathrm{DE}}$.



(a) MCMC path for Γ_{11}

(b) Cumulative plot of ergodic mean for Γ_{11}

(c) MCMC path for Γ_{21}

(d) Cumulative plot of ergodic mean for Γ_{21}

(e) MCMC path for Γ_{41}

(f) Cumulative plot of ergodic mean for Γ_{41}

(g) MCMC path for Γ_{44}

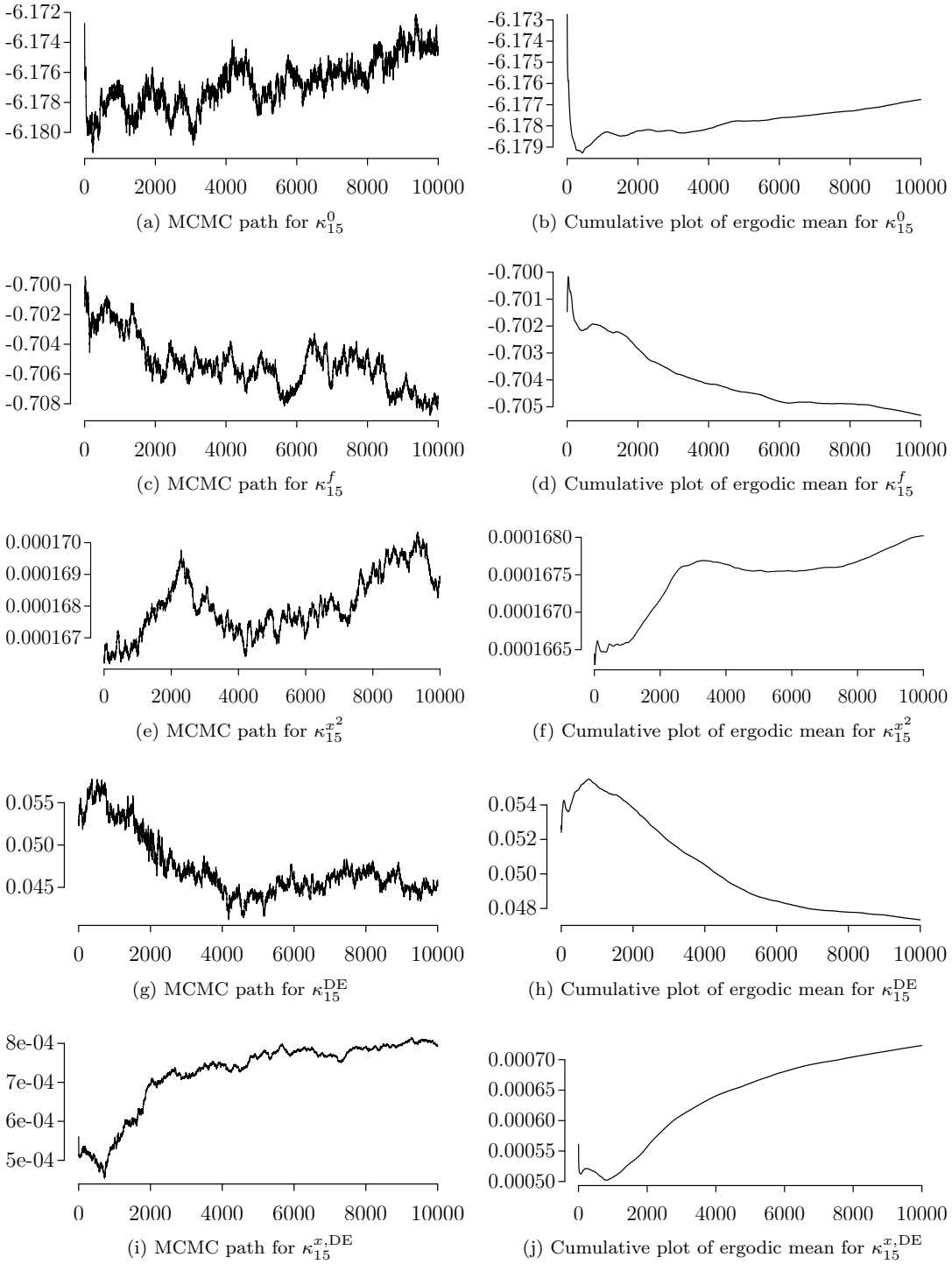
(h) Cumulative plot of ergodic mean for Γ_{44}

(i) MCMC path for Γ_{77}

(j) Cumulative plot of ergodic mean for Γ_{77}

Figure 4.22: Convergence diagnostics for κ_{15}

Left panel: Realised marginal paths of selected entries of κ_{15} after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means.



(a) MCMC path for κ_{15}^{0}

(b) Cumulative plot of ergodic mean for κ_{15}^{0}

(c) MCMC path for κ_{15}^{f}

(d) Cumulative plot of ergodic mean for κ_{15}^{f}

(e) MCMC path for $\kappa_{15}^{x^2}$

(f) Cumulative plot of ergodic mean for $\kappa_{15}^{x^2}$

(g) MCMC path for $\kappa_{15}^{\mathrm{DE}}$

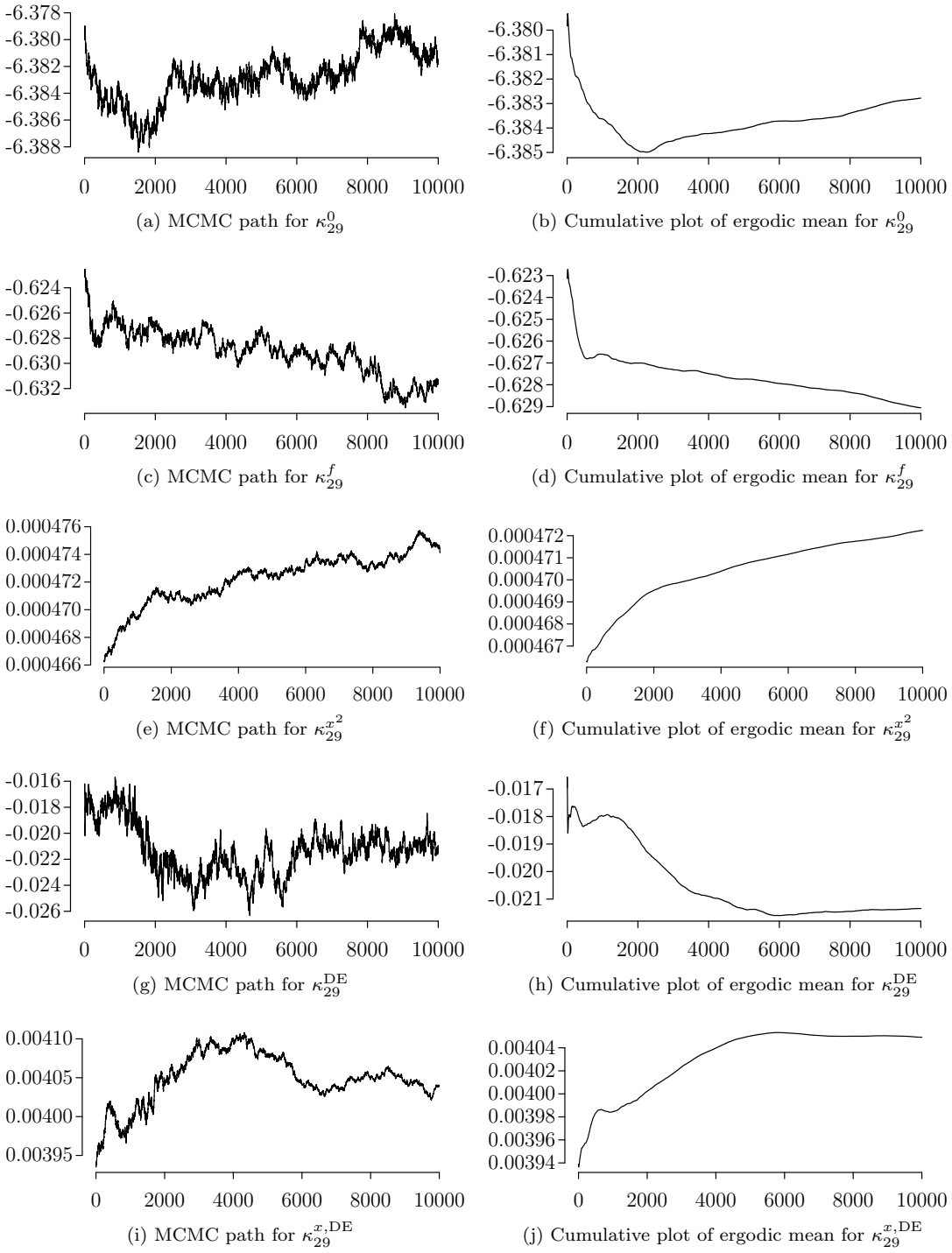
(h) Cumulative plot of ergodic mean for $\kappa_{15}^{\mathrm{DE}}$

(i) MCMC path for $\kappa_{15}^{x,\mathrm{DE}}$

(j) Cumulative plot of ergodic mean for $\kappa_{15}^{x,\mathrm{DE}}$

Figure 4.23: Convergence diagnostics for κ_{29}

Left panel: Realised marginal paths of selected entries of κ_{29} after running the MCMC algorithm with $N = 1{,}000{,}000$ iterations subject to thinning by a factor of 100. *Right panel:* Corresponding evolution of the ergodic means.



(a) MCMC path for κ_{29}^{0}

(b) Cumulative plot of ergodic mean for κ_{29}^{0}

(c) MCMC path for κ_{29}^{f}

(d) Cumulative plot of ergodic mean for κ_{29}^{f}

(e) MCMC path for $\kappa_{29}^{x^2}$

(f) Cumulative plot of ergodic mean for $\kappa_{29}^{x^2}$

(g) MCMC path for $\kappa_{29}^{\mathrm{DE}}$

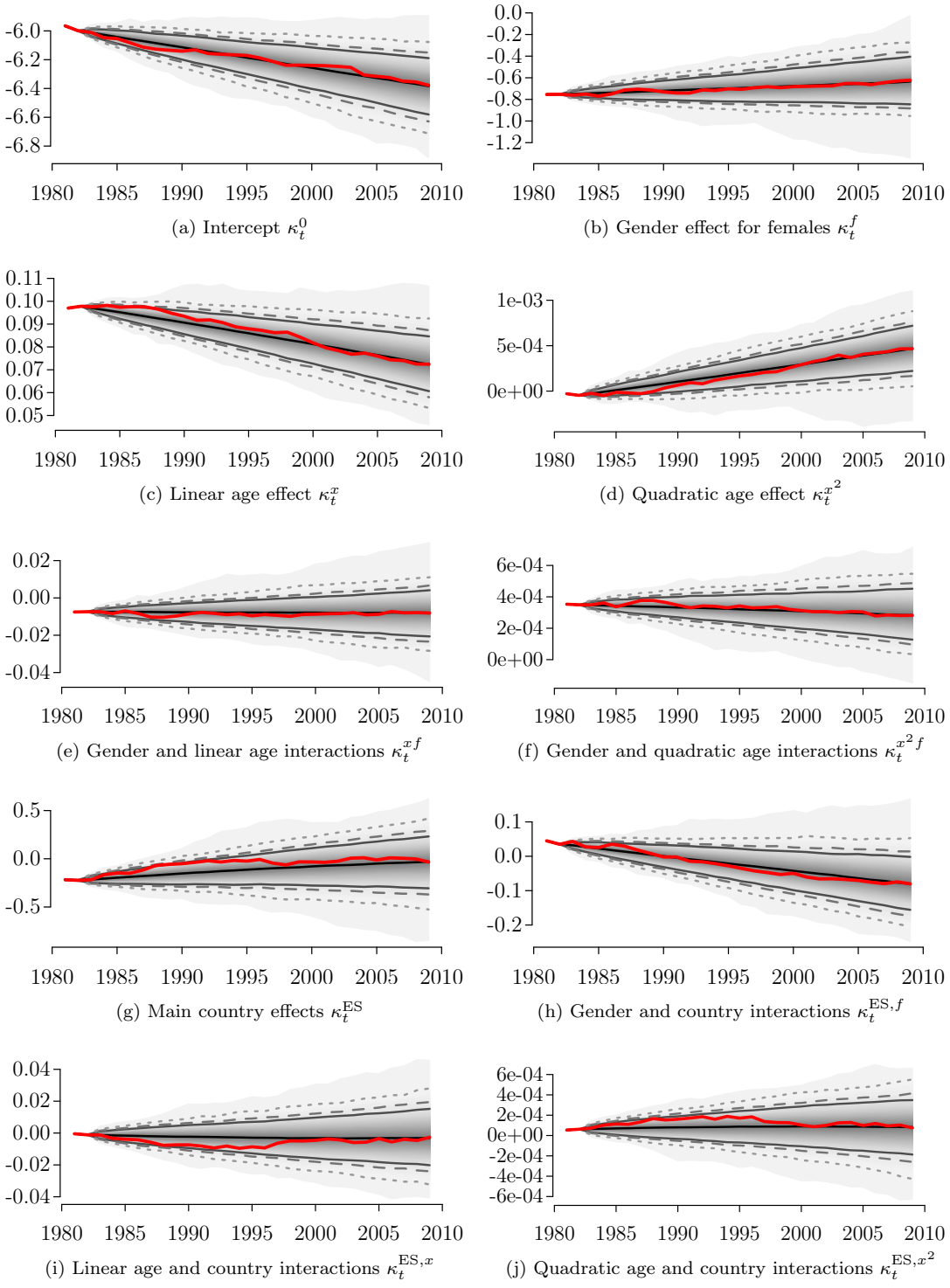
(h) Cumulative plot of ergodic mean for $\kappa_{29}^{\mathrm{DE}}$

(i) MCMC path for $\kappa_{29}^{x,\mathrm{DE}}$

(j) Cumulative plot of ergodic mean for $\kappa_{29}^{x,\mathrm{DE}}$

Figure 4.24: Posterior predictive checking for $\mathcal{K}$

Shown are fancharts of selected marginal posterior predictive distributions for κ_t as time series in t for $1981, \ldots, 2009$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. The country-specific parameters are illustrated for the example of Spain. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Red lines denote ML estimates (starting values).



(a) Intercept κ_t^0

(b) Gender effect for females κ_t^f

(c) Linear age effect κ_t^x

(d) Quadratic age effect $\kappa_t^{x^2}$

(e) Gender and linear age interactions κ_t^{xf}

(f) Gender and quadratic age interactions $\kappa_t^{x^2 f}$

(g) Main country effects κ_t^{ES}

(h) Gender and country interactions $\kappa_t^{\mathrm{ES},f}$

(i) Linear age and country interactions $\kappa_t^{\mathrm{ES},x}$

(j) Quadratic age and country interactions $\kappa_t^{\mathrm{ES},x^2}$

Figure 4.25: Posterior predictive checking for η_{xpgt} for the age of 60

Shown are fancharts of marginal posterior predictive distributions for η_{xpgt} with $x = 60$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1981, \ldots, 2009$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Green lines denote crude estimates from observed data.

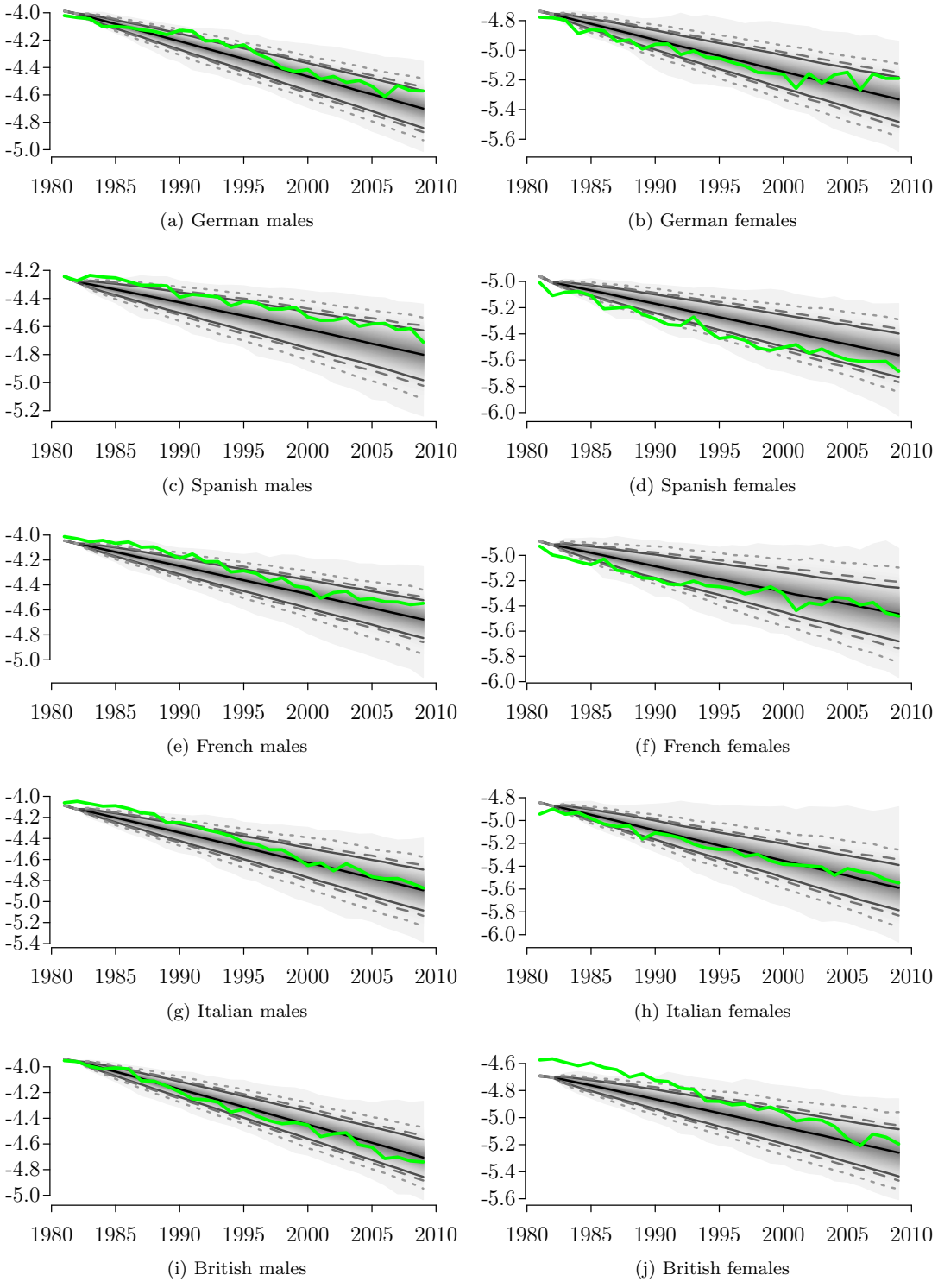


Figure 4.26: Posterior predictive checking for η_{xpgt} for the age of 80

Shown are fancharts of marginal posterior predictive distributions for η_{xpgt} with $x = 80$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1981, \ldots, 2009$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Green lines denote crude estimates from observed data.

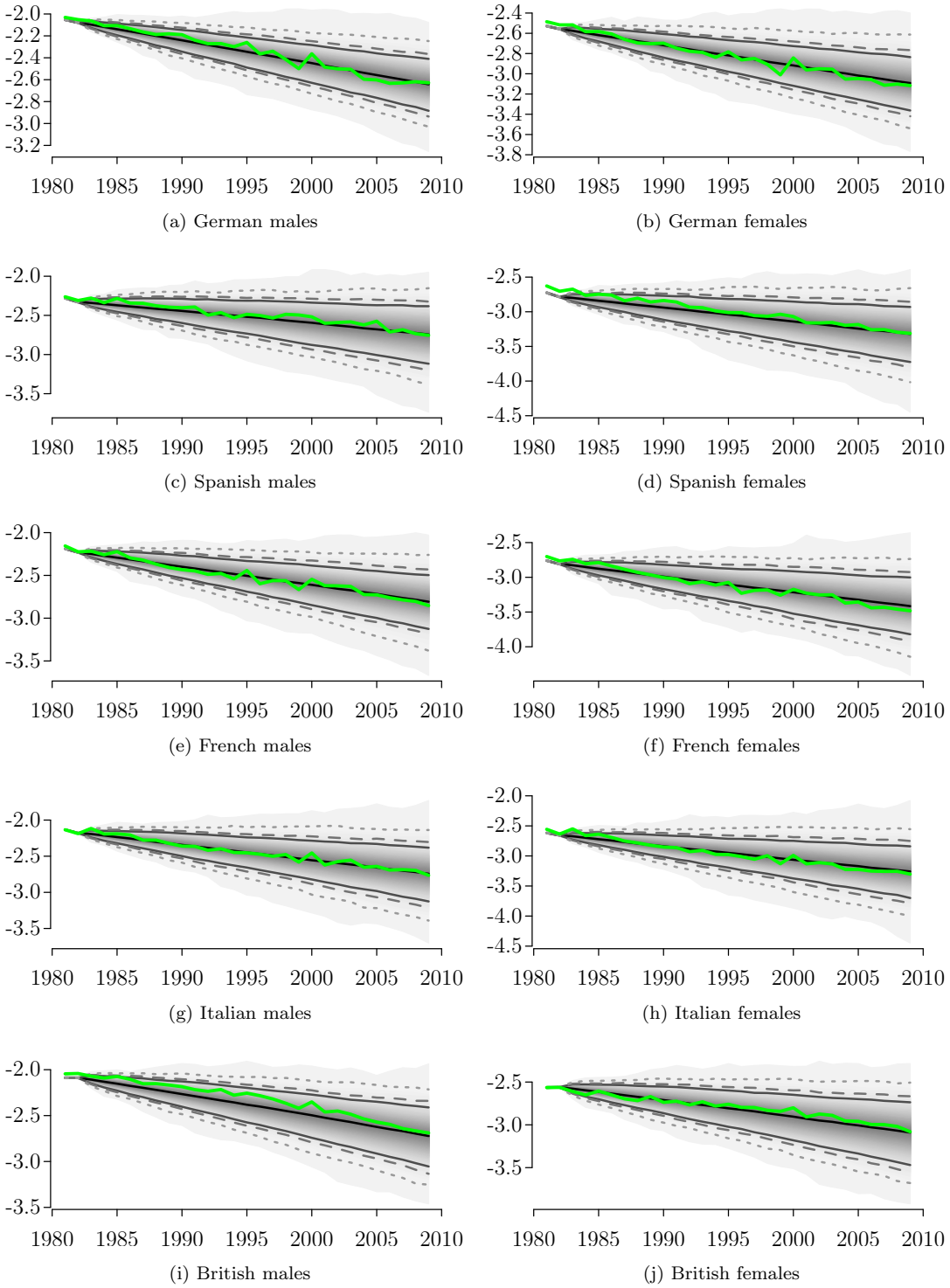
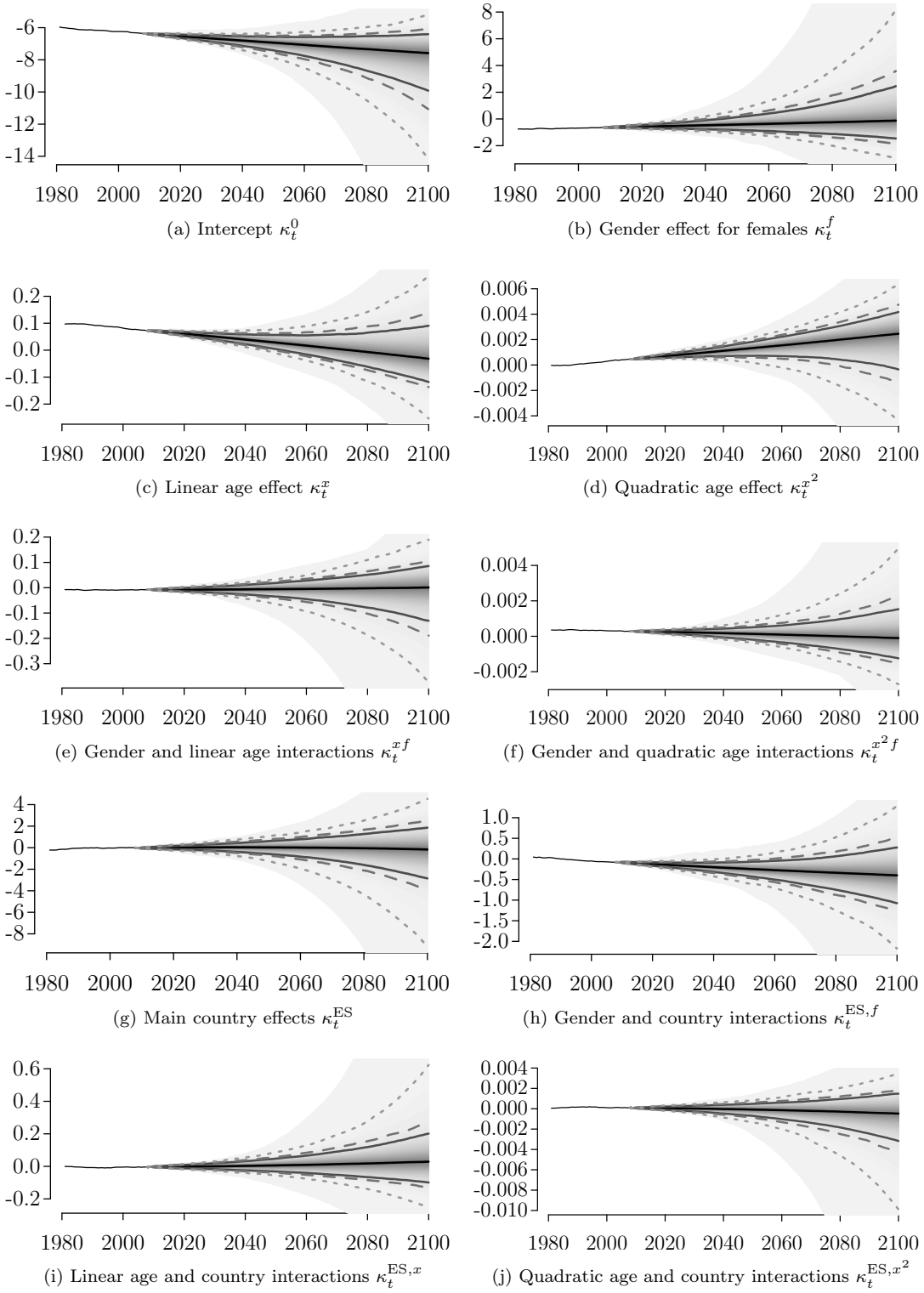


Figure 4.27: External validation for $\mathcal{K}$

Shown are fancharts of selected marginal posterior predictive distributions for future κ_t as time series in t for $2010, \ldots, 2100$, along with ML estimates for κ_t for $1981, \ldots, 2009$ for the same MCMC output as in Figure 4.24. The country-specific parameters are illustrated for the example of Spain. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) Intercept κ_t^0

(b) Gender effect for females κ_t^f

(c) Linear age effect κ_t^x

(d) Quadratic age effect $\kappa_t^{x^2}$

(e) Gender and linear age interactions κ_t^{xf}

(f) Gender and quadratic age interactions $\kappa_t^{x^2 f}$

(g) Main country effects κ_t^{ES}

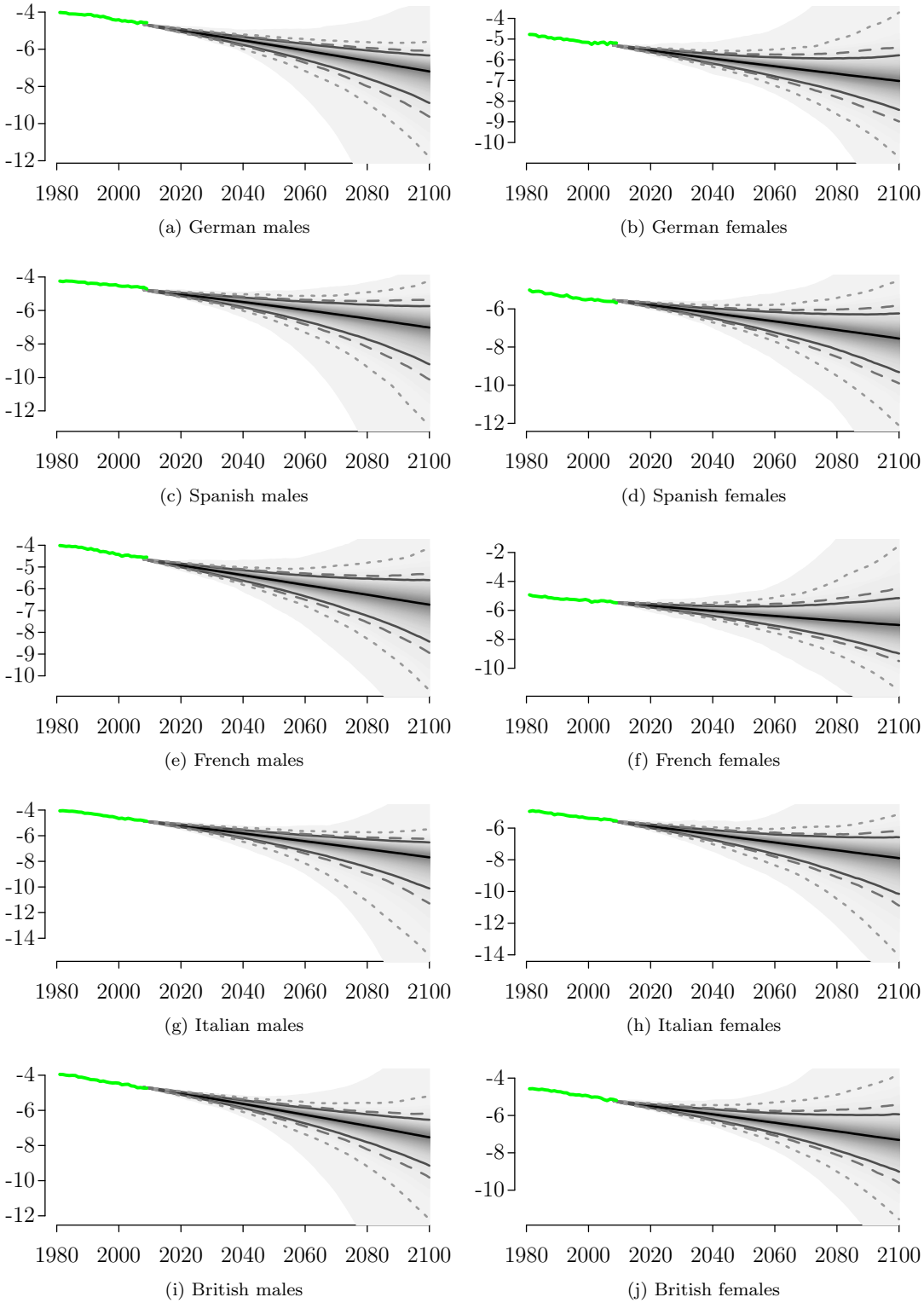
(h) Gender and country interactions $\kappa_t^{\mathrm{ES},f}$

(i) Linear age and country interactions $\kappa_t^{\mathrm{ES},x}$

(j) Quadratic age and country interactions $\kappa_t^{\mathrm{ES},x^2}$

Figure 4.28: External validation for η_{xpgt} for the age of 60

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 60$ for all countries p (by rows) and both genders g (by columns) as time series in t for $2010, \ldots, 2100$, along with crude estimates for η_{xpgt} for $1981, \ldots, 2009$ from observed data (green lines) for the same MCMC output as in Figure 4.25. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) German males

(b) German females

(c) Spanish males

(d) Spanish females

(e) French males

(f) French females

(g) Italian males

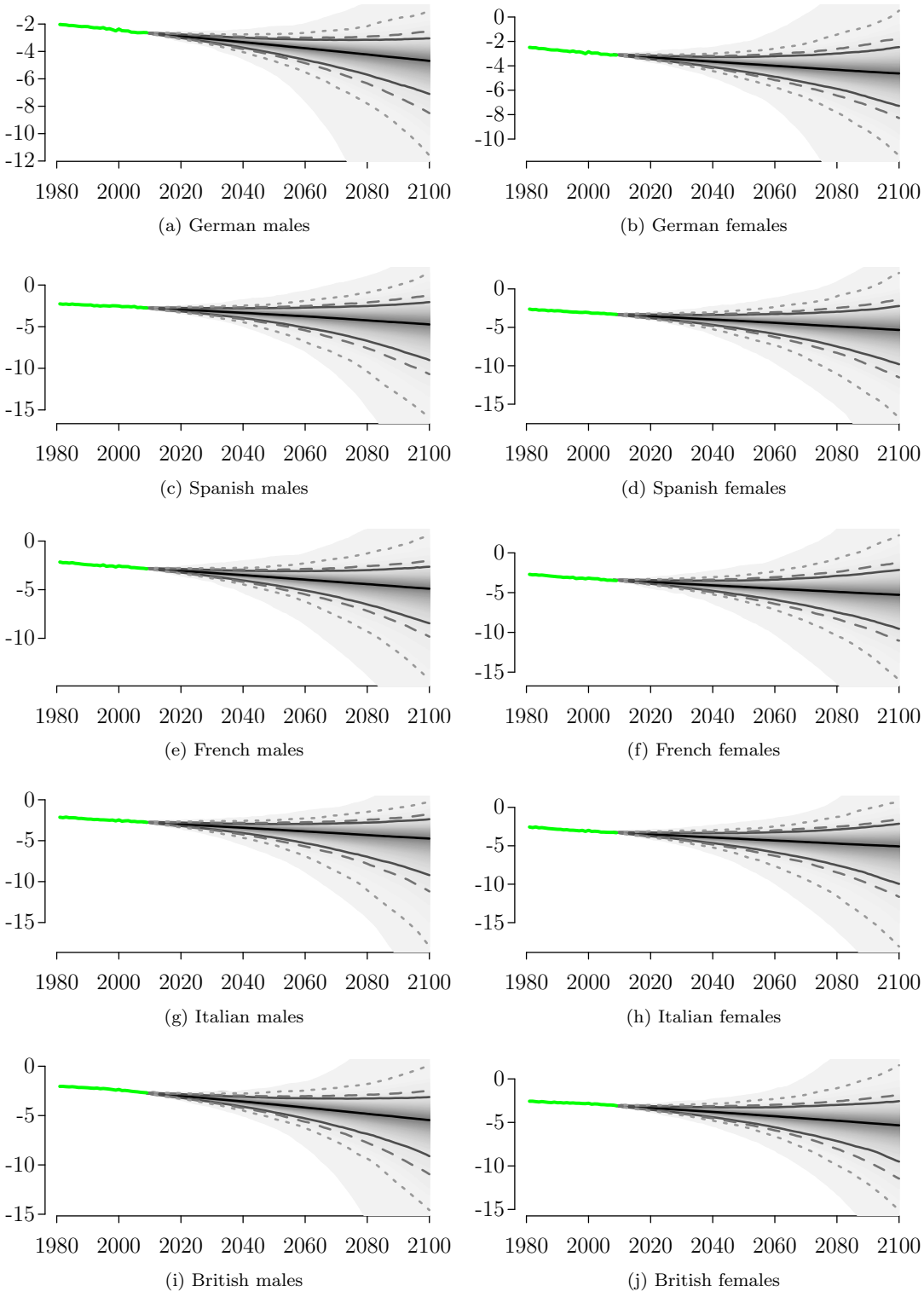
(h) Italian females

(i) British males

(j) British females

Figure 4.29: External validation for η_{xpgt} for the age of 80

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 80$ for all countries p (by rows) and both genders g (by columns) as time series in t for $2010, \ldots, 2100$, along with crude estimates for η_{xpgt} for $1981, \ldots, 2009$ from observed data (green lines) for the same MCMC output as in Figure 4.26. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) German males (b) German females

(c) Spanish males (d) Spanish females

(e) French males (f) French females

(g) Italian males (h) Italian females

(i) British males (j) British females

The convergence plots in Figures 4.17 to 4.21 for the hyperparameters, and in Figures 4.22 and 4.23 for the parameters, lead to the same conclusions as before. In the first case, an approximation of the posterior distribution based on the thinned sample after 750,000 burn-in iterations appears reasonable good, although the MCMC algorithm should be run for longer to reach ultimate convergence, notably for Ω and the undiagnosed joint distribution. Convergence to a stationary distribution has not yet been reached for the CBD parameters.

Internal validation is performed in Figures 4.24 to 4.26. The ML estimates are included in the 90% credibility intervals obtained from simulation of the time series over the calibration window, and generally line up with the medians of the posterior predictive distribution. While internal forecasts of the linear predictor at the age of 80 draw the overly optimistic picture as before, counterparts for the age of 60 are less sophisticating. Here, the width of the credibility intervals is rather realistic over the entire course of simulation, as obvious from the plots for, e.g., German males and females (a–b). More concerning, however, are results from plots of Spanish and British females, among others, which show that the posterior predictive distribution may not include crude estimates of η_{xpgt} based on observed data for the first twenty years. For these countries, changes in the trend are detected from the plots of country-specific age parameters in Figure 4.16, which cannot be captured by their corresponding posterior predictive medians in Figure 4.24. In fact, the bad performance is due to substantial differences between the crude values of the initial years 1981 and 1982 and their posterior predictive values of κ_1 and κ_2 with their ML estimates as prior means. An increase of the prior variance or even a non-informative prior might improve results, but the principal reliance of the forecasts on the posterior predictive distribution of $\mathcal{K}$, which was shown to suffer from the commonly observed lack of fit in mortality prediction models, seems to be a more substantial problem. For now, it is concluded that the BMPMP model may lack fit for early years of the forecasting period, and future work should be devoted to improve starting values of mortality predictions. However, from a biological point of view, the country-specific deviation parameters are expected to stay within some certain range. Future regime changes of such parameters should be captured by the growing long-term variance in the posterior predictive distribution, as indeed seen in all aforementioned cases.

Finally, Figures 4.24 to 4.26 depict predictions of selected marginal CBD parameters and their corresponding linear predictors for the unknown logits of mortality rates until 2100. The linear and symmetric increase of uncertainty over a short forecast period of only a couple of decades turns into a quadratic and skewed pattern in the long-run, notably after around 50 years. Based on the output in Figure 4.24, best estimates given by the medians of posterior predictive distributions prolong linear trends in the individual CBD parameters. For example, an ongoing linear decline of linear predictors for 40-year-old men in the reference population, known from the period 1981–2009, is anticipated. The difference between men and women for this age group is expected to remain constant or to diminish slowly. The projected medians for the age-related parameters suggest a sustainable transmission from the rather linear pattern of mortality rates versus age towards a merely and, in fact, pure quadratic effect in the year of 2100 without major differences between both genders. Medians for population-specific effects, exemplarily shown for Spain, stay closely around zero with a slightly declining trend for Spain over the entire course of the forecast period.

On the subject of the uncertainty in the parameter estimates, the posterior predictive distribution of the intercept allows the mortality improvements for a 40-year-old male to reduce in pace or to worsen to a level known from the early 1980s, at the year of 2100 on a credibility level of 95%. More probability mass is devoted to the lower tail of the distribution, i.e. deviations to a faster improvement in mortality are stronger. Here, 95% credibility intervals allow the speed in mortality improvements to increase by a factor of up to eight. As such, the Bayesian approach with its easily quantifiable uncertainty in long-term mortality improvements replaces the elliptical forecasts of confidence intervals in common frequentist approaches by a wider variety of scenarios. With regards to the rapid and sudden improvement in mortality rates observed for the 20^{th} century, it is a desired feature of the model's 100-year-span prediction to cover potentially biologically plausible changes due to unforeseen medical, social, and economic events. Interpretation of the remaining parameters in Figure 4.24 should not be done marginally, because the analogy of evolution in credibility bands suggests a strong correlation between future paths such that extreme outcomes affect each other. This is additionally supported by the consistency of individual population- and gender-specific fancharts in Figures 4.25 and 4.26, in which extreme outcomes for the parameters seem to coincide. Median forecasts extend the linear decline of the linear predictors over the calibration window into

the future. Posterior predictive distributions are generally skewed to the left, making faster improvements in mortality more likely than slower or even deteriorating changes. The similarity in future developments, e.g. the diminishing differences between both genders, and distributions for the year 2100, which principally cover the same range of possible values, indicate that population-specific mortality rates wander jointly rather than independently. A discussion of the joint distribution of future predictions is specifically undertaken in the following section.

4.1.8 Joint Posterior Predictive Distribution

In this section, the ability to jointly forecast mortality rates for an arbitrary number of populations is assessed for this case study. As outlined earlier, certain marginal time series in $\mathcal{K}$ show strong correlation. Hence, in Bayesian forecasting, this correlation should be observed for the future realisations of the CBD parameters. Figures 4.30 and 4.31 visualise the correlation matrices for the posterior predictive distribution of the vector κ_t for the years 2050 and 2100, respectively. Strong or almost perfect correlation exists between the six main population-independent parameters for intercept, gender, and age effects. For example, negative correlation is strong for the pairs intercept and gender, linear age and quadratic age, and the linear and quadratic age interactions with gender. Such dependencies are easily interpretable as cancellation effects, if one gender or age group benefits from major mortality improvements only. Similar conclusions are drawn for the many other correlations between these parameters. For the population-specific parameters, high correlation is rather sparse, indicating that shocks within national mortality data are connected to a limited extent. Whereas the population-specific main and gender effects do not exhibit much correlation among themselves, the analogous correlation structure between main and age effects is apparent. A notable feature is the consistent joint behaviour for predicted $\mathcal{K}$ when going from 2050 to 2100 with a slightly increasing magnitude in correlations, particularly for British parameters. Some other strong covariances are scattered, whereas remaining correlation is less distinct. The latter finding suggests that parameter matrices in the VECM could be thinned out in future work to reduce complexity in such high-dimensional models.

Figure 4.30: Correlation matrix for κ_t at the year of 2050

Shown are the bivariate correlations of all marginal posterior predictive distributions for future κ_t with t for 2050 for the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. Blue and red circles denote positive and negative correlation, respectively. The magnitude of the absolute correlation is given by both the circle size and colour intensity.

Order of parameters: Following the six population-independent parameters $\kappa_t^0, \kappa_t^f, \kappa_t^x, \kappa_t^{x^2}, \kappa_t^{x,f}, \kappa_t^{x^2,f}$, country-specific parameters are given blockwise for their main effects, gender interactions, linear age interactions, and quadratic age interactions, respectively, each of which ordered by DE, ES, FR, IT, UK.

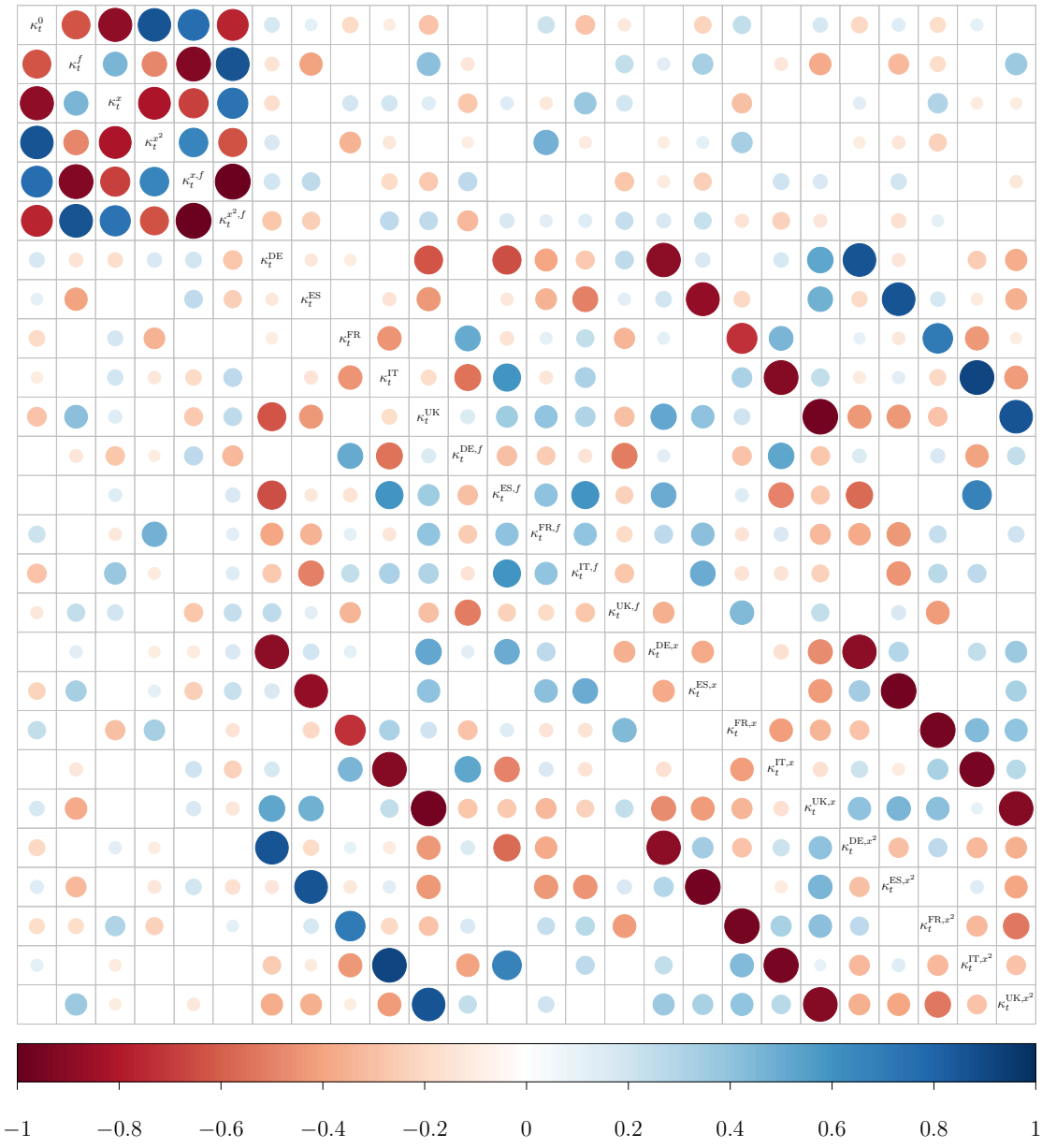


Figure 4.31: Correlation matrix for κ_t at the year of 2100

Shown are the bivariate correlations of all marginal posterior predictive distributions for future κ_t with t for 2100 for the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. Blue and red circles denote positive and negative correlation, respectively. The magnitude of the absolute correlation is given by both the circle size and colour intensity.

Order of parameters: Following the six population-independent parameters $\kappa_t^0, \kappa_t^f, \kappa_t^x, \kappa_t^{x^2}, \kappa_t^{x,f}, \kappa_t^{x^2,f}$, country-specific parameters are given blockwise for their main effects, gender interactions, linear age interactions, and quadratic age interactions, respectively, each of which ordered by DE, ES, FR, IT, UK.

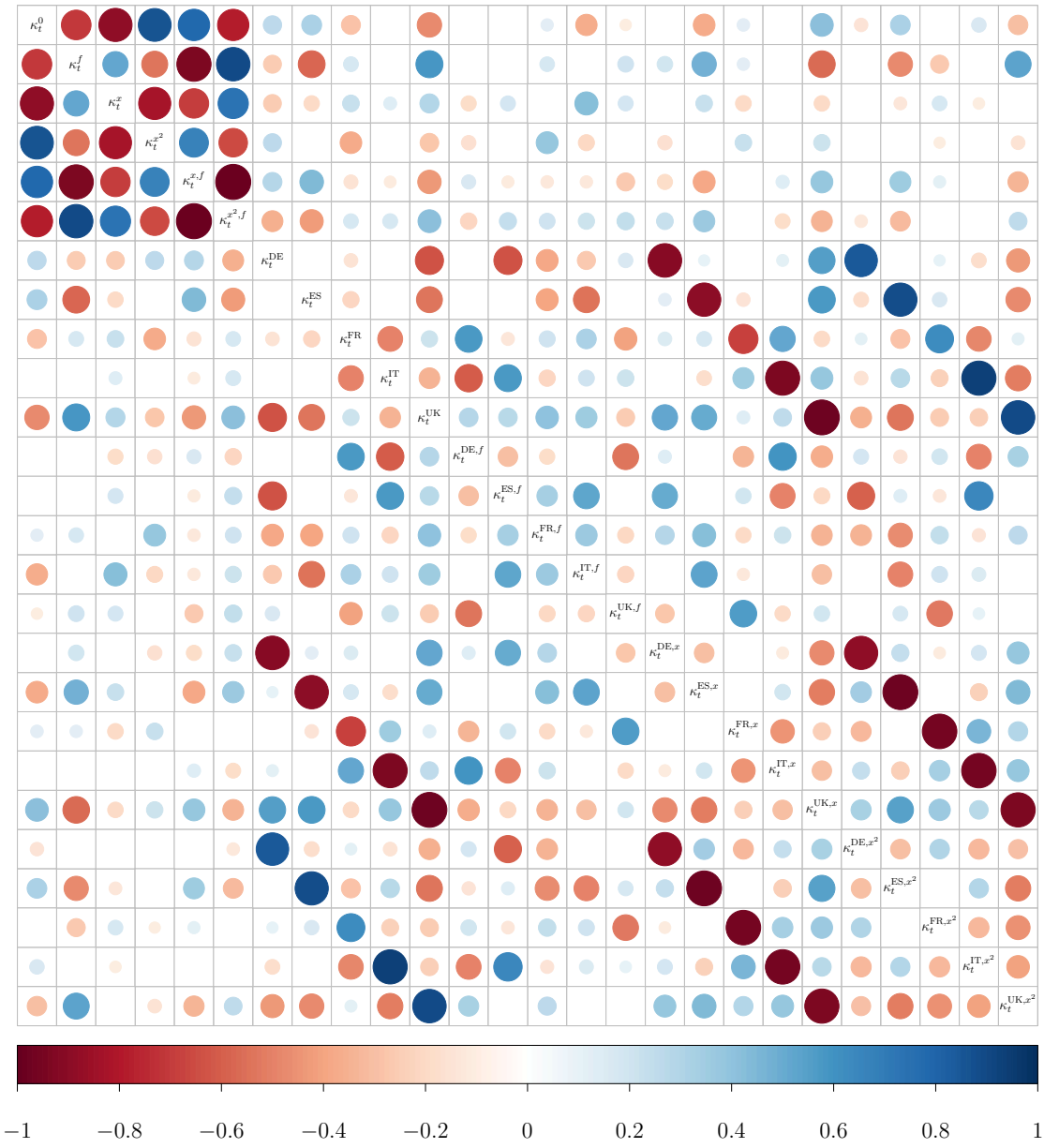
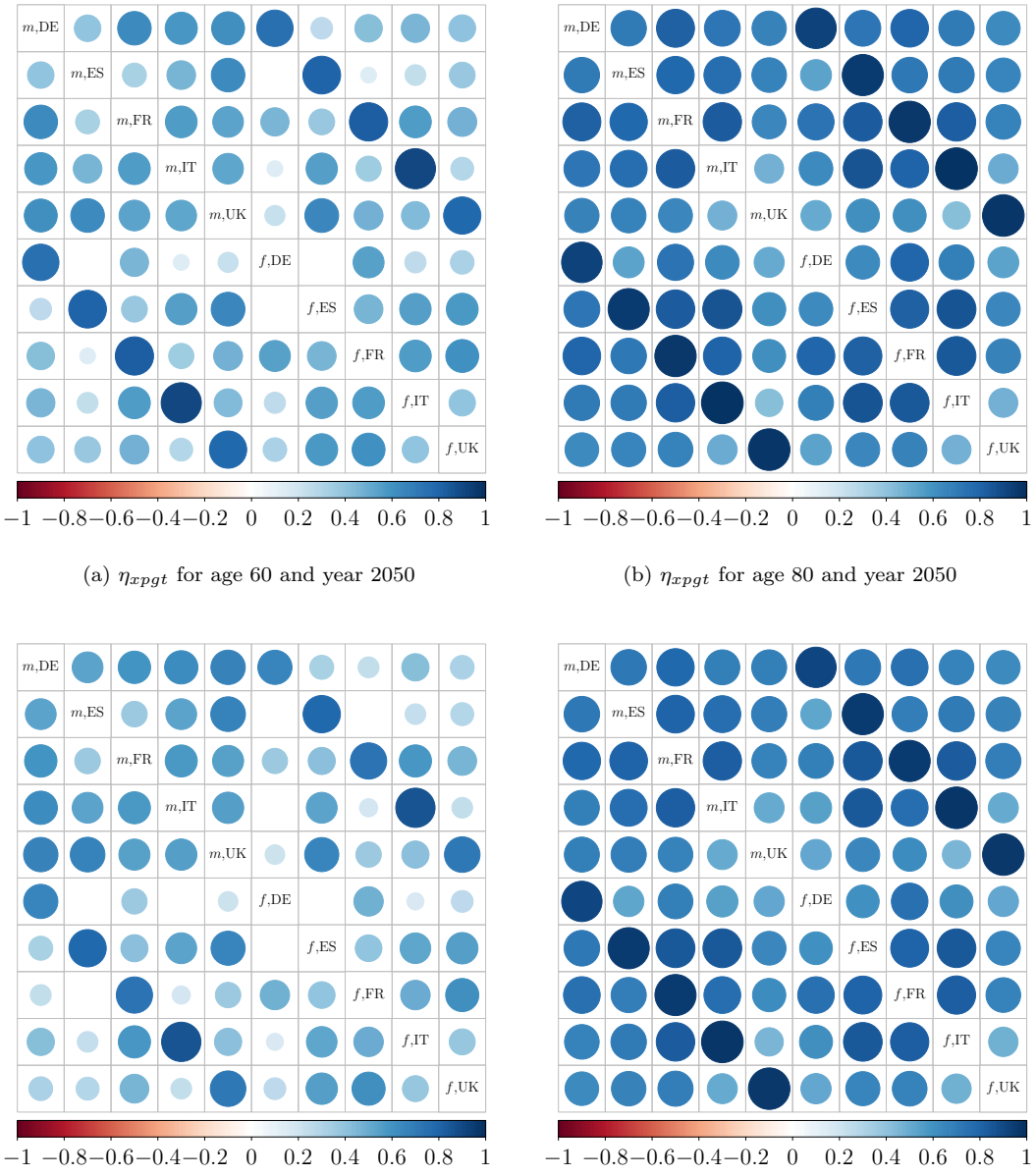


Figure 4.32: External validation for η_{xpgt} for the ages of 60 and 80 at years 2050 and 2100

Shown are the bivariate correlations of all population-specific posterior predictive distributions for future η_{xpgt} with $x = 60$ and $x = 80$ (by columns) and t for 2050 and 2100 (by rows) for the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. Blue circles denote positive correlation, and its magnitude is given by both the circle size and colour intensity.
Order of parameters: Male and female populations are given in two blocks, each of which ordered by DE, ES, FR, IT, UK.



(a) η_{xpgt} for age 60 and year 2050

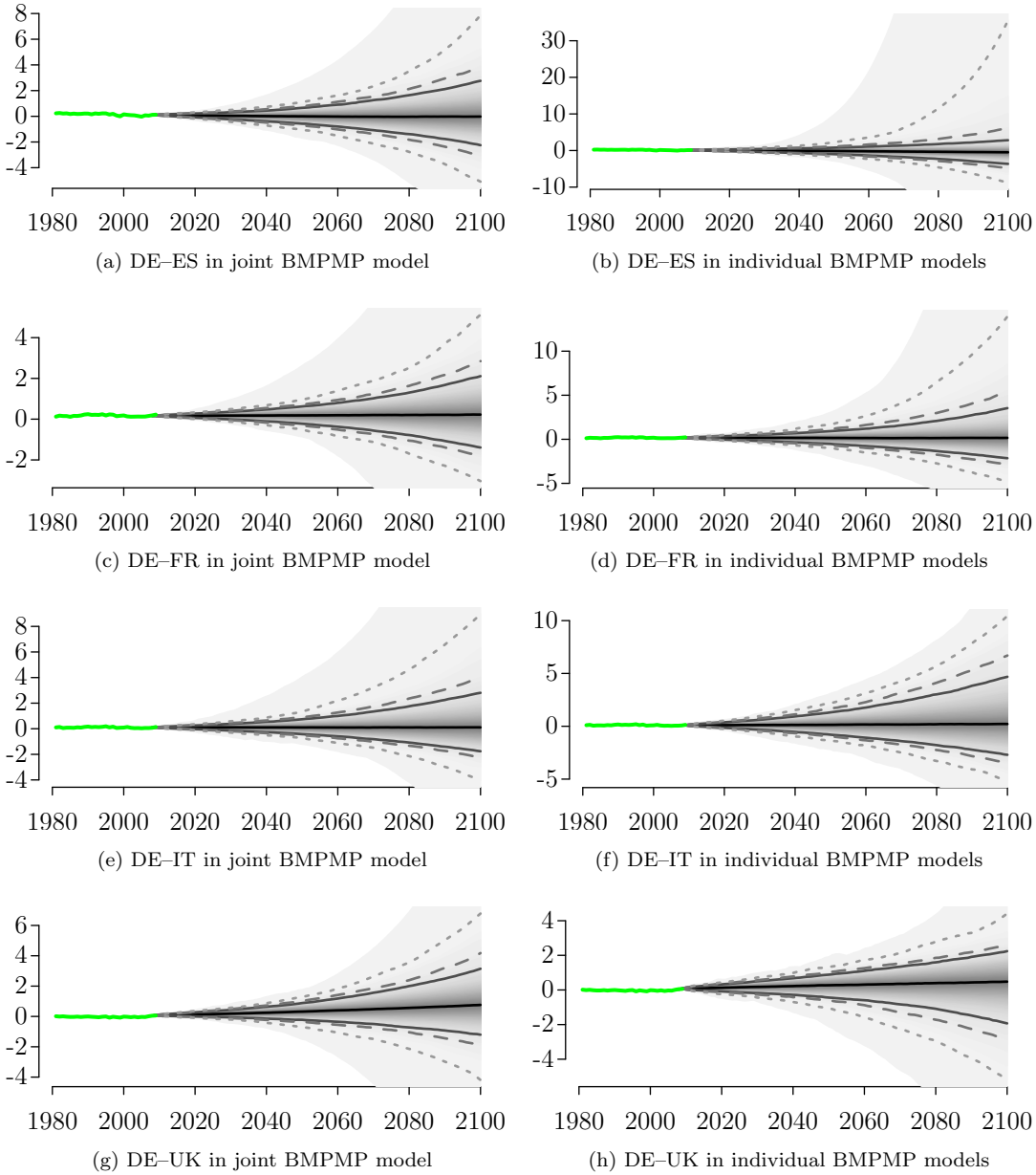
(b) η_{xpgt} for age 80 and year 2050

(c) η_{xpgt} for age 60 and year 2100

(d) η_{xpgt} for age 80 and year 2100

Figure 4.33: Differences in η_{xpgt} for different populations p of males for the age of 80

Shown are fancharts of marginal posterior predictive distributions for future differences in η_{xpgt} with $g = m$ and $x = 80$ for selected pairs of countries p (by rows) in both joint and individual BMPMP models (by columns) as time series in t for $2010, \ldots, 2100$, along with crude estimates for the differences in η_{xpgt} for $1981, \ldots, 2009$ from observed data (green lines) for the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.
Left panel: The joint BMPMP model for all five countries DE, ES, IT, FR, UK with $k = 2$ and $r = 5$. *Right panel:* Combined output from five individual BMPMP models for each country with the six main CBD parameters only and $k = 2$ and $r = 1$, respectively. The MCMC algorithm is run with the same length, burn-in period, and thinning.



(a) DE–ES in joint BMPMP model

(b) DE–ES in individual BMPMP models

(c) DE–FR in joint BMPMP model

(d) DE–FR in individual BMPMP models

(e) DE–IT in joint BMPMP model

(f) DE–IT in individual BMPMP models

(g) DE–UK in joint BMPMP model

(h) DE–UK in individual BMPMP models

Figure 4.32 shows visualisations of the correlation matrices for η_{xpgt} at all combinations of fixed ages 60 and 80 and fixed years 2050 and 2100. Correlations are assessed between all ten strata composed by the five countries and both genders. Note that the dependencies between different age groups need not be assessed due to the design of the BMPMP model. The plots show that the model indeed exhibits the desired property of projecting highly positive correlations between the individual populations into the future. As expected, the magnitude is highest for both genders within the same country. In this case study, mortality rates further appear to have a stronger correlation between the countries for higher ages. While any combination of populations has a minimum correlation coefficient of 0.5 for the age of 80, these dependencies may even vanish for the age of 60; see German and Spanish females, for instance. Surprisingly, differences between 2050 and 2100 are negligible.

Despite the discussion of the aforementioned plots, the BMPMP model's outcome needs to be compared to corresponding results from independent univariate projection models in order to fully assess the distinct feature of joint mortality rate projections for an arbitrary number of populations. If the BMPMP model is able to quantify a biologically anticipated stability in mortality patterns, or to describe dependencies between the individual populations, then the result of the joint modelling approach should generally differ from combined results of independent marginal models. For a fair comparison, the principal framework of the univariate models must coincide with the BMPMP model, apart from the change in the number of populations. Consequently, each individual population is modelled via the CBD approach as in (4.1) without any population-specific effects, i.e. the linear predictor only consists of the intercept, the gender-related main effect, and the linear and quadratic effects for age. The number of parameters decreases to six including interaction terms. The parameters are again forecast through a VECM, which is now far less high-dimensional than in the multi-population case. The lag order is kept constant at $k = 2$, and the cointegration rank is chosen to be $r = 1$ for each independent model. Apart from the latter specification, Bayesian estimation with the same prior assumptions and MCMC approximation of the posterior distribution is applied to avoid any other methodological differences between the univariate and multivariate cases.

Since the individual models for all five countries are independent, replicated plots as in Figure 4.32 (not shown) only exhibit correlation between both genders within each

nation. However, to quantify the extent to which the joint modelling approach leads to more biological plausibility, the predictions must be analysed for convergent or divergent behaviour rather than simple correlation. Fancharts of posterior predictive distributions for differences in selected linear predictors for males are plotted over the forecasting period for both the joint and the independent modelling approaches in Figure 4.33. Although credibility bands remain narrow for several decades, it is noteworthy that even for the joint modelling approach, the stochastic model design cannot prevent a considerable spread between linear predictors of two different countries. Median predictions, however, do forecast anticipated non-divergence of marginal mortality rates. A comparison of results from both models reveals that even the independent models do not seem principally implausible, as their best estimates suggest a stable and non-divergent future, too. However, except for the difference between German and British males, credibility intervals are usually much wider – up to a doubling in size. Thus, the joint modelling approach of the newly introduced BMPMP model leads to substantial reduction in overly conservative and unreasonable credibility regions from combined single-population models. The case study greatly supports the usage of the joint forecasts.

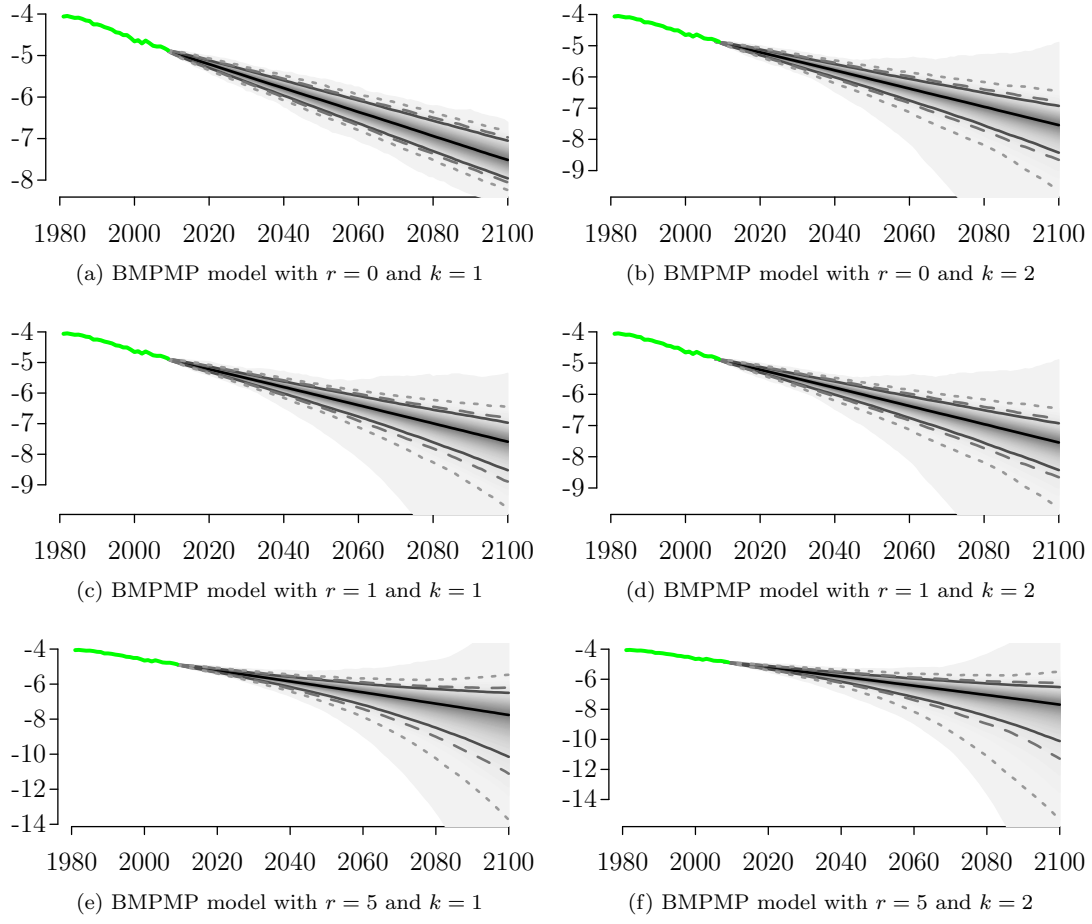
4.1.9 Comparison of Different Model Set-ups

In this section, the BMPMP model of mortality projections for the Big Five over the calibration period 1981–2009 is compared for different choices in the VECM specifications. The six model set-ups under consideration are comprised of all combinations of the values $k = 1, 2$ and $r = 0, 1, 5$ for the lag and cointegration order, respectively. Each model's posterior distribution is approximated by the MCMC output of $N = 1{,}000{,}000$ iterations subject to a thinning factor of 100, and after discarding a burn-in period of 750,000 iterations. Convergence and model diagnostics are all analogous to previous findings. Most notably, a rough approximation can only be achieved for the hyperparameters, and convergence for the CBD parameters is not satisfying. A detailed discussion on such diagnostic measurements is omitted here for convenience. Note that a detailed sensitivity analysis w.r.t. the choice of priors is not provided within this discussion and left open for future work.

A comparison of all six model specifications is conducted via qualitative assessment of differences in posterior predictive forecasts over the window 2010–2100. Figure 4.34

Figure 4.34: Comparison of posterior predictive η_{xpgt} for 60-year-old Italian males for different choices of cointegration and lag orders

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 60, p = \text{IT}, g = m$ for BMPMP models specified through different cointegration orders $r = 0, 1, 5$ (by rows) and lag orders $k = 1, 2$ (by columns) as time series in t for $2010, \ldots, 2100$, along with crude estimates for η_{xpgt} for $1981, \ldots, 2009$ from observed data (green lines) for the respective MCMC outputs with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) BMPMP model with $r = 0$ and $k = 1$

(b) BMPMP model with $r = 0$ and $k = 2$

(c) BMPMP model with $r = 1$ and $k = 1$

(d) BMPMP model with $r = 1$ and $k = 2$

(e) BMPMP model with $r = 5$ and $k = 1$

(f) BMPMP model with $r = 5$ and $k = 2$

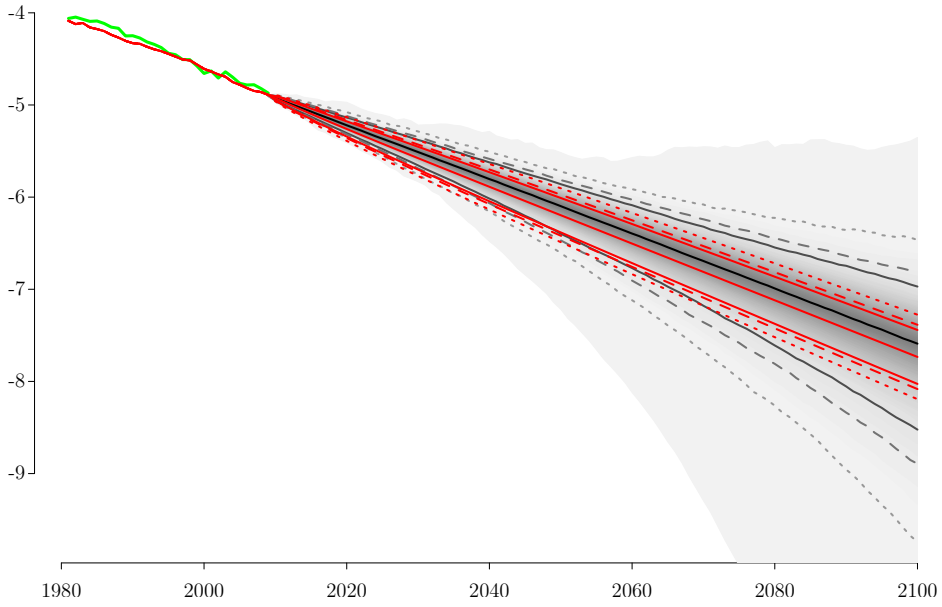
shows posterior mortality projections at the example of 60-year-old Italian males. The cointegration order r has the significant effect of increasing both the variability and left-skewness in the posterior predictive distributions of the long-term forecasts for the linear predictors. Whereas best estimates given through median values are stable throughout all set-ups, the choice of $r = 0$ or $r = 1$ yields an elliptical uncertainty pattern, known from standard frequentist and single-population models in the literature on stochastic mortality predictions. Credibility intervals remain narrow and symmetric over the entire course of the prediction window. Notably, the posterior predictive distribution at the year of 2100 is of similar nature than what is observed for the year 2050. Therefore, increasing uncertainty about future developments in medicine, economy, and society over such long time spans is neglected. In contrast, the plots for $r = 5$ reveal that the inclusion of more cointegrating relationships in the VECM leads to increasing variability over time, with a shift of probability mass towards a possible slowdown and, to a larger extent, fast-pace mortality improvements. Such results not only underline the distinct feature of a more complex and principally more realistic uncertainty structure, but also the analyst's ability to express prior beliefs through the choice of r. Apart from effects on the 1% tails when $r = 0$, the lag order k, however, appears to have no visible effect on the BMPMP model's performance, so that the joint modelling approach without AR features appears adequate. Summarising, the robustness for $r > 0$ suggests the usage of the BMPMP model with $k = 1$ and, implicitly, a large reduction in computation time.

4.1.10 Comparison of Bayesian and Maximum-Likelihood Estimation

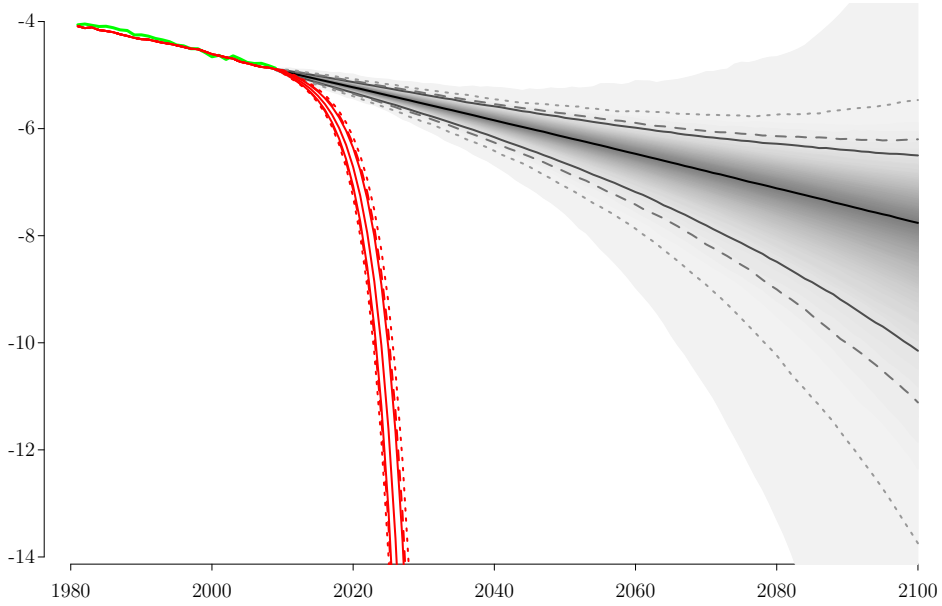
In this case study, in which mortality rates of five different countries stratified by two genders are to be forecast based on a calibration window of around 30 years, the Bayesian methodology is essential, when $k = 2$, to overcome the problem of over-parametrisation. Flexibility in model specification is therefore strongly limited. Due to the vanishing importance of AR terms in the VECM, as seen in the previous section, this case study only allows a comparison of Bayesian and frequentist estimation procedures with the choice of $k = 1$. Note, however, that with increasing number of populations, even the Markovian version of the BMPMP model becomes over-parametrised and frequentist approaches cannot be established.

Figure 4.35: Comparison of Bayesian posterior predictive and maximum-likelihood estimates for η_{xpgt} for 60-year-old Italian males

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 60, p = \text{IT}, g = m$ for BMPMP models with lag order $k = 1$ and both cointegration ranks $r = 1, 5$ as time series in t for $2010, \ldots, 2100$, along with ML forecasts (red lines) and crude estimates for η_{xpgt} for $1981, \ldots, 2009$ from observed data (green lines) for the respective MCMC outputs with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals for the Bayesian estimates are given by solid, dashed, dotted, and limiting lines, respectively. Analogous line types are used for the 90%, 95%, and 99% confidence intervals for the frequentist estimates.



(a) BMPMP model with $r = 1$



(b) BMPMP model with $r = 5$

In the following, frequentist forecasts are computed via subsequent ML estimation of the CBD model, i.e. the previously used starting values, and the VECM with Gaussian white noise, as outlined in Appendix C.4. Least-squares estimation, which is additionally described in the appendix, will not be considered here. Figure 4.35 shows the forecasts for the BMPMP models with $k = 1$ and $r = 1, 5$, known from Figure 4.34, with superimposed confidence intervals from ML estimation. Obviously, the frequentist approach fails for the model with five cointegration relationships in that the time series and, consequently, the linear predictors η_{xpgt} explode for all populations. A natural explanation is that with 612 VECM parameters for 728 latent observations, ML estimation – although mathematically possible – becomes highly unstable with an ill-conditioned maximisation problem. Valid ML results for the case of $r = 1$ and for other case studies (not shown) support this theory. As also discovered for other examples, the first plot shows that when frequentist estimation is doable and well-conditioned, best estimates are close to the median values of the posterior predictive distributions from Bayesian estimation. Confidence intervals are narrower than their credibility counterparts and closest for the choice of $r = 0$. In general, they are completely contained in the corresponding credibility bands and less sophisticating than the Bayesian outcome w.r.t. uncertainty patterns. In conclusion, ML estimation is not available for the BMPMP model in most but the smallest applications, and, if it is, problems with its robustness and plausibility still speak for the Bayesian approach.

4.2 Case Study 2: Central European Countries

This case study applies the gender-specific model in (4.1) for the age interval [40,100] and calibration window of 1981–2009 to the five Central European countries of Austria (AT), the Czech Republic (CZ), Germany[2] (DE), Hungary (HU), and Poland (PL), as shown in Figure 4.36. These countries are selected to combine populations which in the augmented common factor model by Li and Lee (2005) could either be modelled – both as members of the so-called *low-mortality group* (i.e. Austria and Germany) or as members of the remaining out-of-group sample (i.e. the Czech Republic) – or had to be abandoned due to an explosive behaviour in the marginal first-order AR time series model (i.e. Hungary). The remaining country Poland was not analysed in this study.

[2] As before, German data are restricted to the territory of former West Germany for consistency purposes.

Figure 4.36: Map of the five Central European countries in case study 2

Shown in green are the five Central European countries of the second case study: Austria (AT), the Czech Republic (CZ), Germany (DE, data only for West Germany in this case study), Hungary (HU), and Poland (PL), within Europe. For details on territorial coverage for these five countries, see the list of countries in the preface.

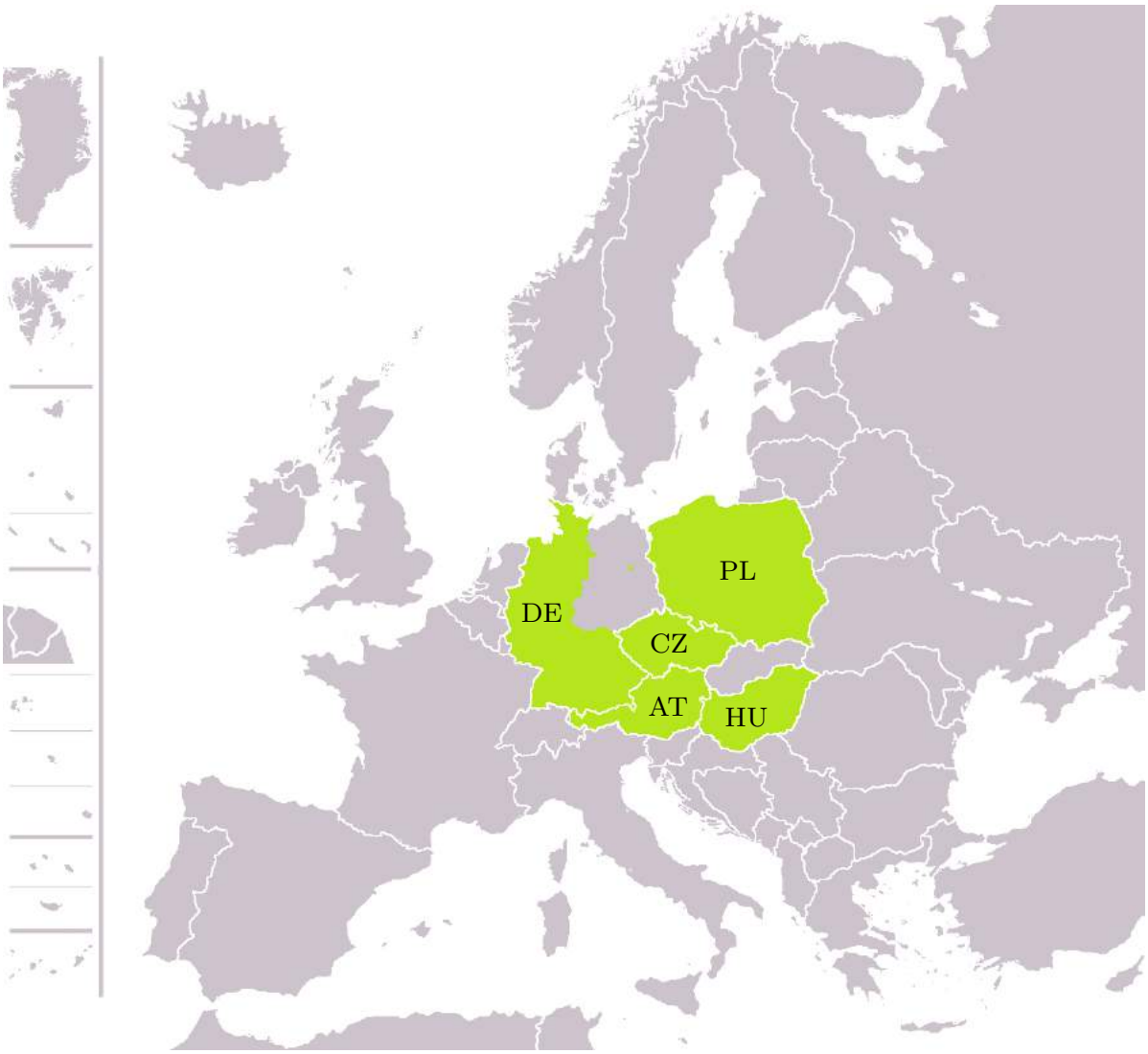
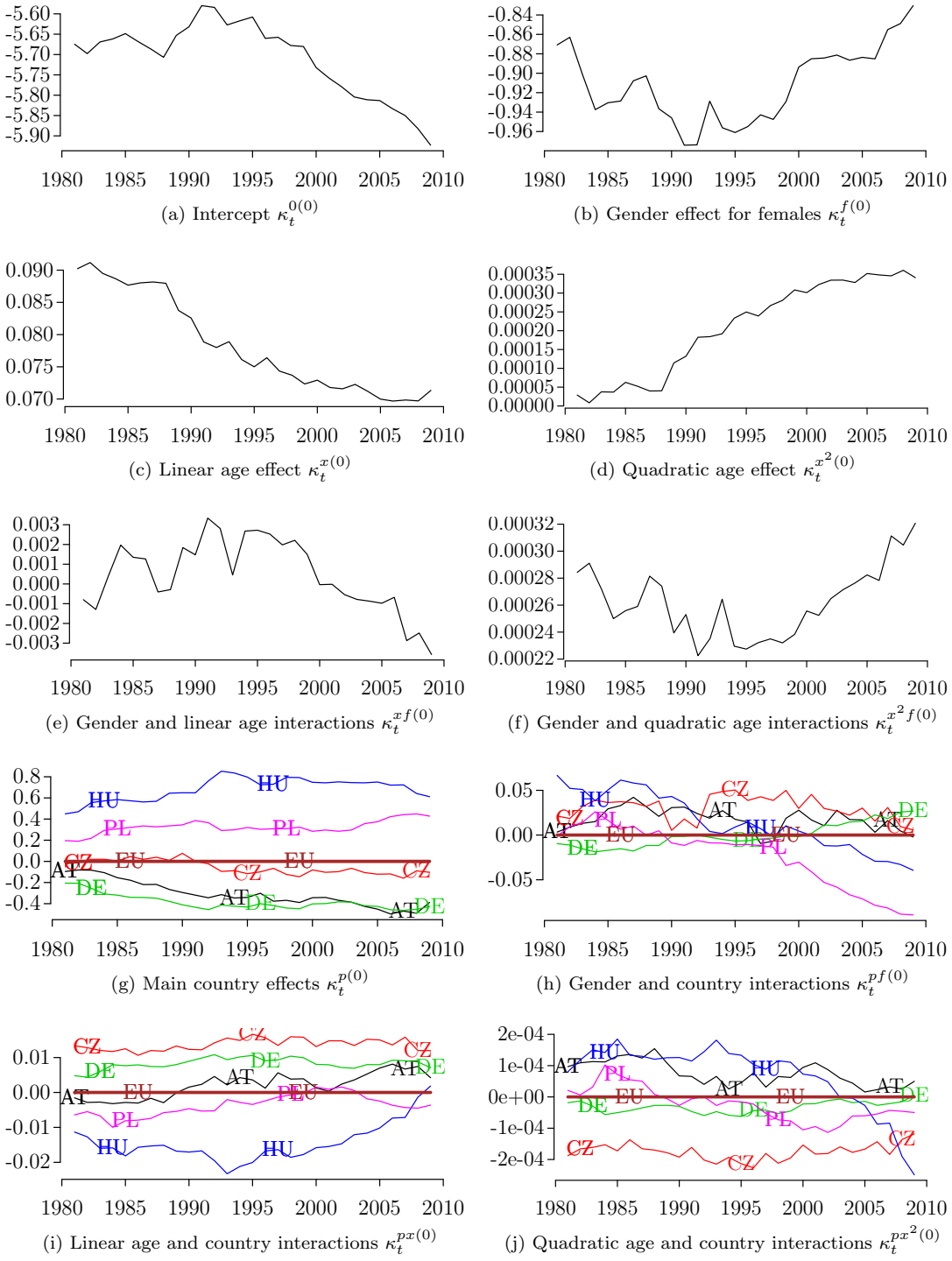


Figure 4.37: Starting values for $\mathcal{K}$

Shown are the ML estimates $\kappa_t^{(0)}$ of the CBD parameters as time series in t for $1981, \ldots, 2009$.



(a) Intercept $\kappa_t^{0(0)}$

(b) Gender effect for females $\kappa_t^{f(0)}$

(c) Linear age effect $\kappa_t^{x(0)}$

(d) Quadratic age effect $\kappa_t^{x^2(0)}$

(e) Gender and linear age interactions $\kappa_t^{xf(0)}$

(f) Gender and quadratic age interactions $\kappa_t^{x^2f(0)}$

(g) Main country effects $\kappa_t^{p(0)}$

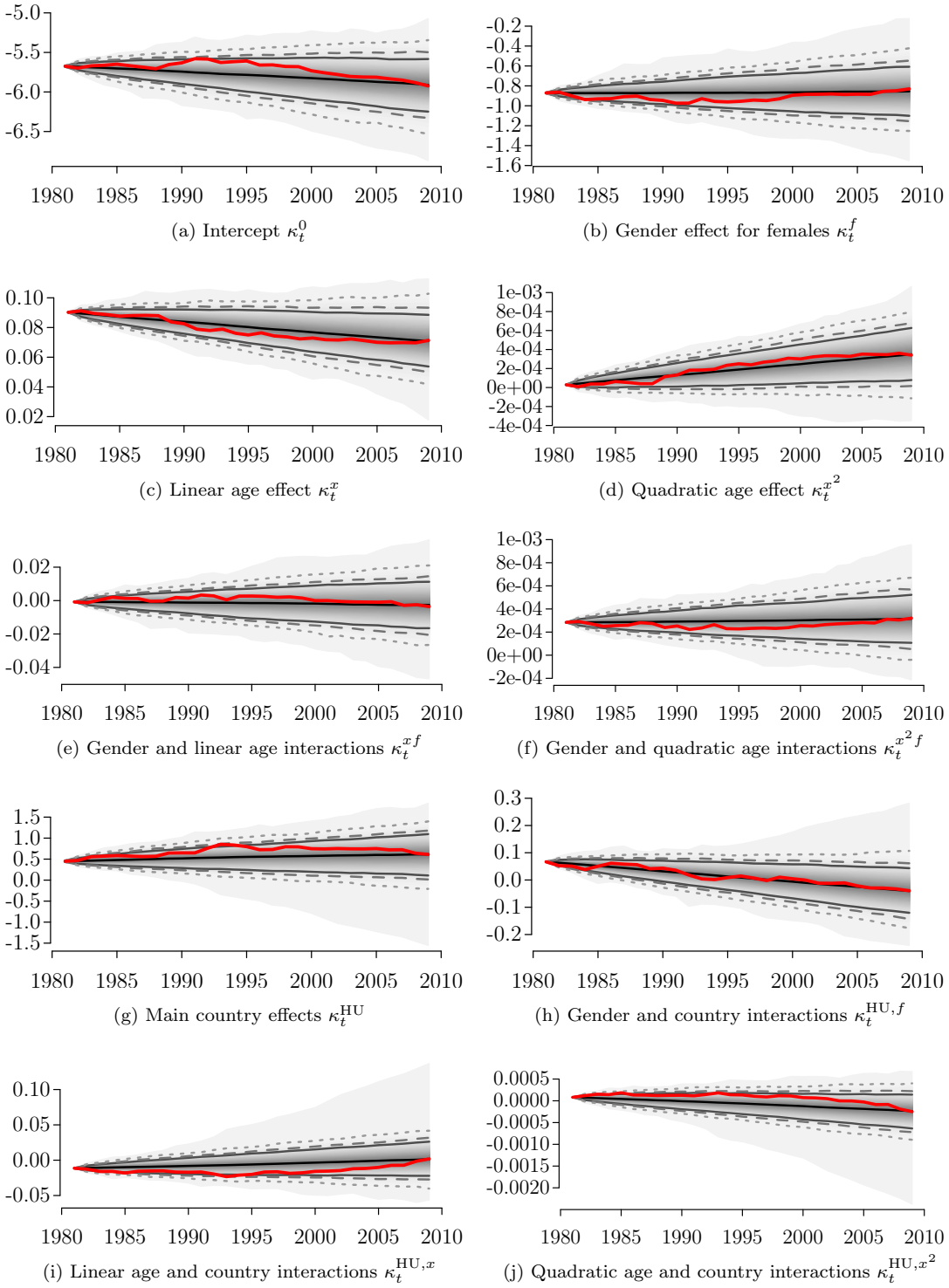
(h) Gender and country interactions $\kappa_t^{pf(0)}$

(i) Linear age and country interactions $\kappa_t^{px(0)}$

(j) Quadratic age and country interactions $\kappa_t^{px^2(0)}$

Figure 4.38: Posterior predictive checking for $\mathcal{K}$

Shown are fancharts of selected marginal posterior predictive distributions for κ_t as time series in t for $1981, \ldots, 2009$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. The country-specific parameters are illustrated for the example of Hungary. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Red lines denote ML estimates (starting values).



(a) Intercept κ_t^0

(b) Gender effect for females κ_t^f

(c) Linear age effect κ_t^x

(d) Quadratic age effect $\kappa_t^{x^2}$

(e) Gender and linear age interactions κ_t^{xf}

(f) Gender and quadratic age interactions $\kappa_t^{x^2 f}$

(g) Main country effects κ_t^{HU}

(h) Gender and country interactions $\kappa_t^{\text{HU},f}$

(i) Linear age and country interactions $\kappa_t^{\text{HU},x}$

(j) Quadratic age and country interactions κ_t^{HU,x^2}

Figure 4.39: Posterior predictive checking for η_{xpgt} for the age of 60

Shown are fancharts of marginal posterior predictive distributions for η_{xpgt} with $x = 60$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1981, \ldots, 2009$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Green lines denote crude estimates from observed data.

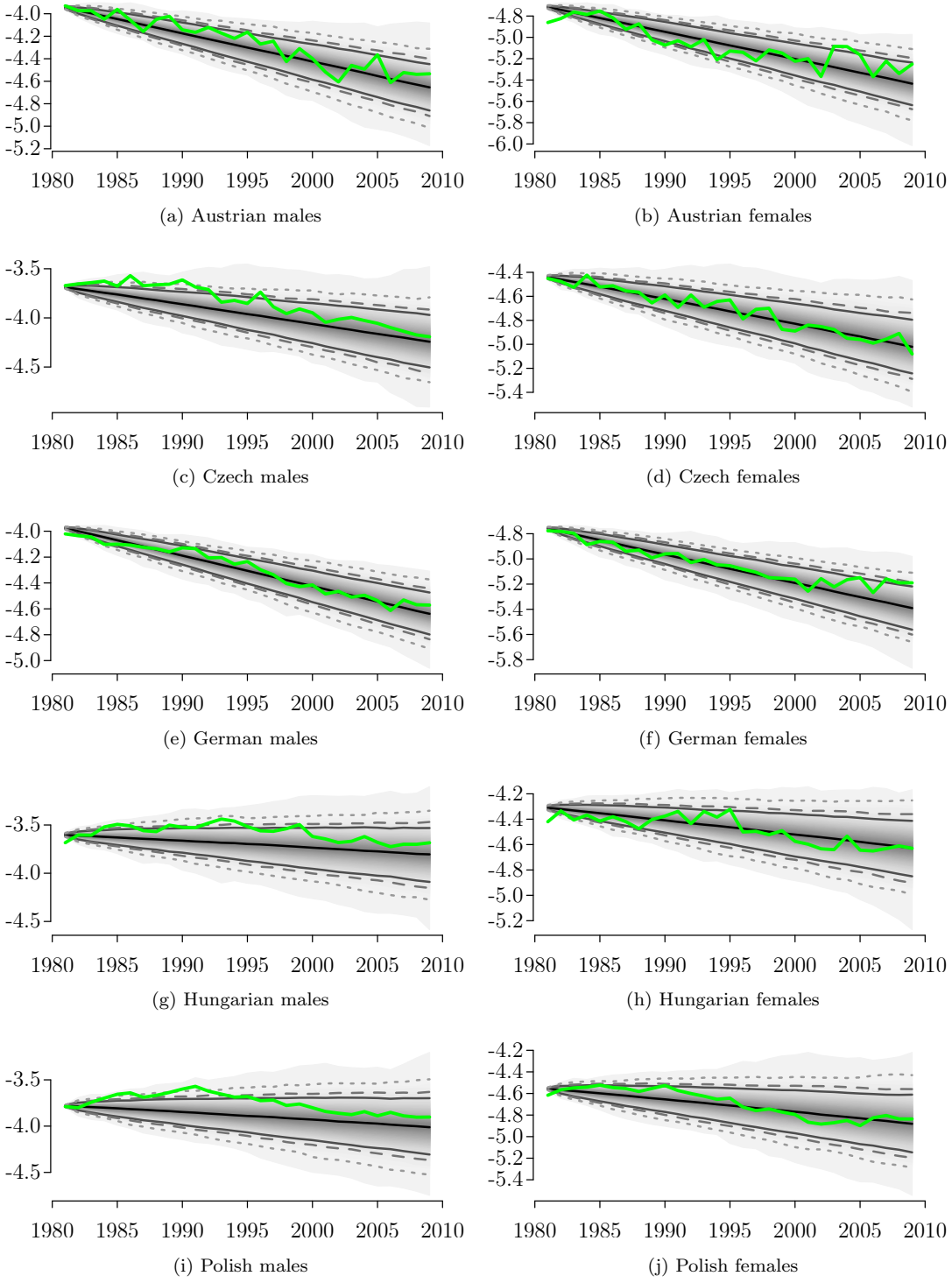


Figure 4.40: Posterior predictive checking for η_{xpgt} for the age of 80

Shown are fancharts of marginal posterior predictive distributions for η_{xpgt} with $x = 80$ for all countries p (by rows) and both genders g (by columns) as time series in t for $1981, \ldots, 2009$ based on the MCMC output with $N = 1{,}000{,}000$, burn-in length of 750,000, and thinning factor 100. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively. Green lines denote crude estimates from observed data.

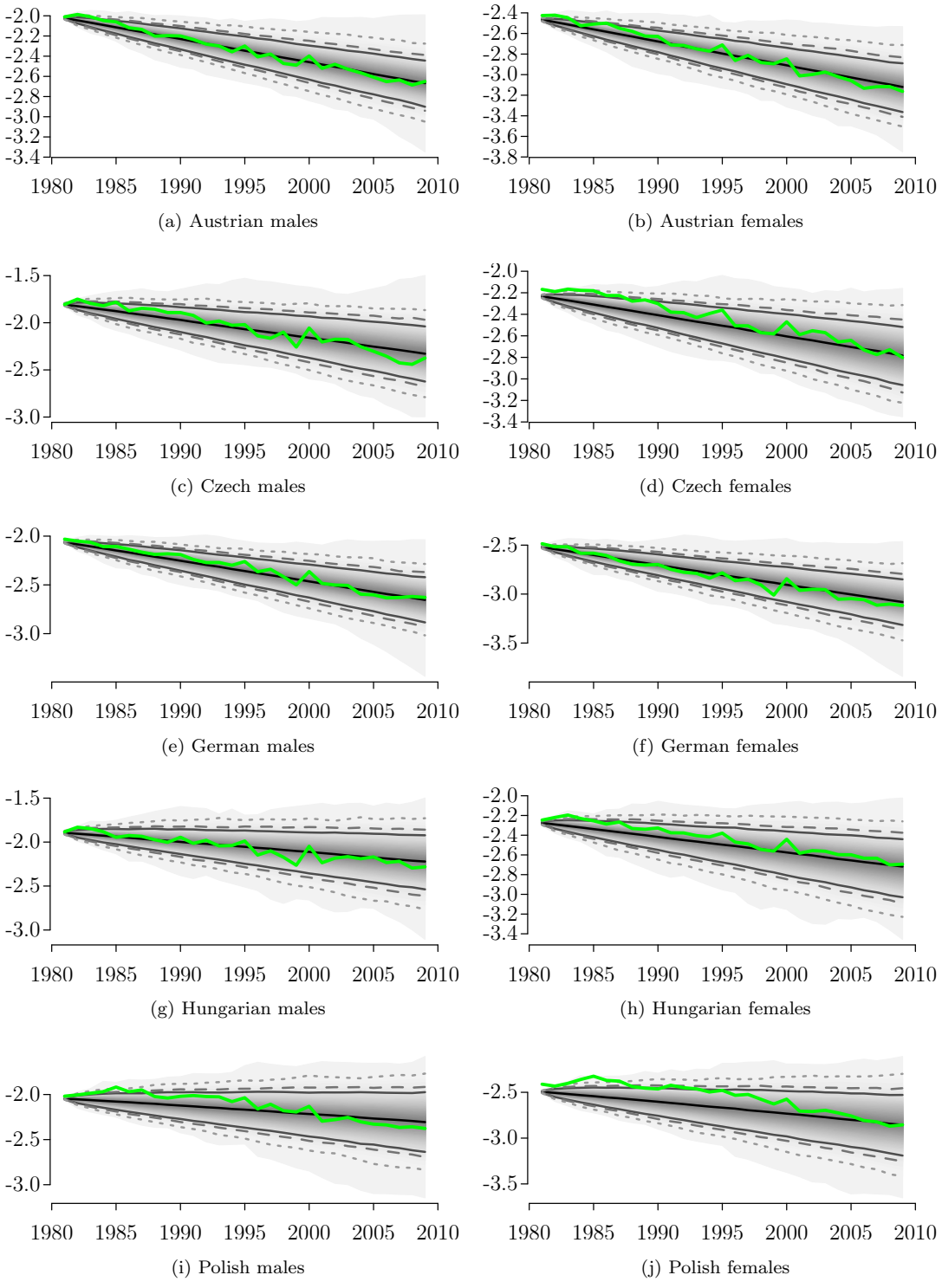
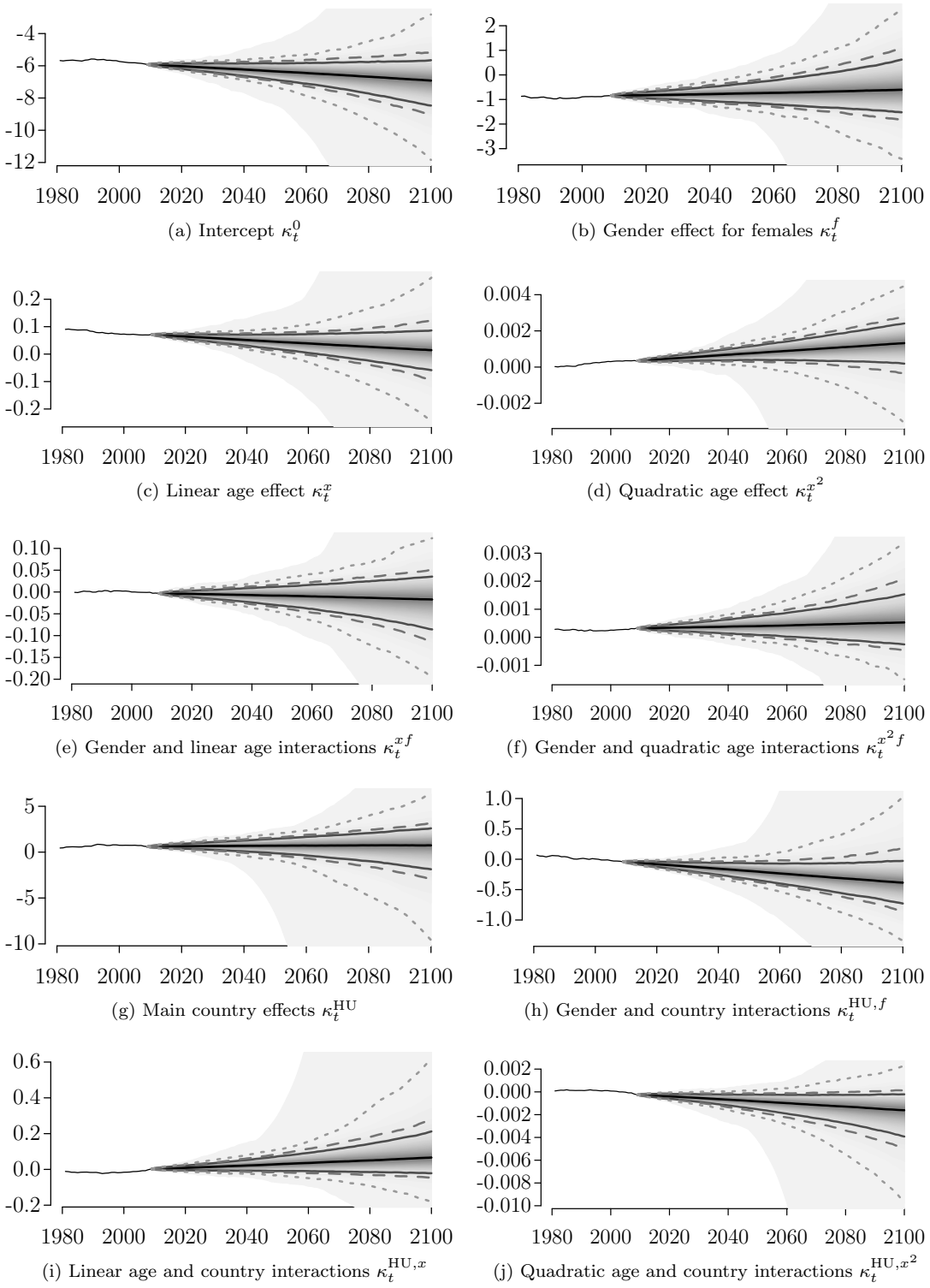


Figure 4.41: External validation for $\mathcal{K}$

Shown are fancharts of selected marginal posterior predictive distributions for future κ_t as time series in t for $2010, \ldots, 2100$, along with ML estimates for κ_t for $1981, \ldots, 2009$ for the same MCMC output as in Figure 4.38. The country-specific parameters are illustrated for the example of Hungary. 90%, 95%, 99%, and 100% credibility intervals are given by solid, dashed, dotted, and limiting lines, respectively.



(a) Intercept κ_t^0

(b) Gender effect for females κ_t^f

(c) Linear age effect κ_t^x

(d) Quadratic age effect $\kappa_t^{x^2}$

(e) Gender and linear age interactions κ_t^{xf}

(f) Gender and quadratic age interactions $\kappa_t^{x^2 f}$

(g) Main country effects κ_t^{HU}

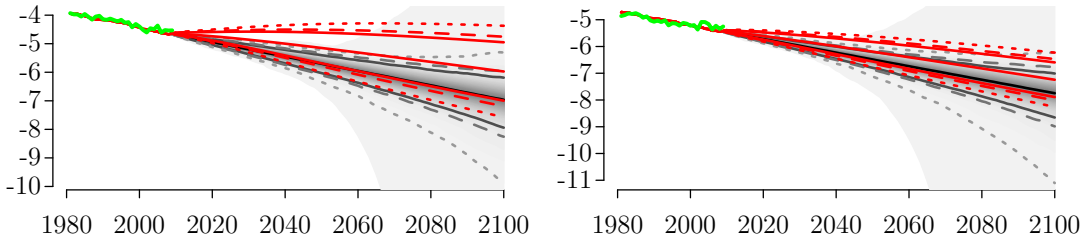
(h) Gender and country interactions $\kappa_t^{\mathrm{HU},f}$

(i) Linear age and country interactions $\kappa_t^{\mathrm{HU},x}$

(j) Quadratic age and country interactions $\kappa_t^{\mathrm{HU},x^2}$

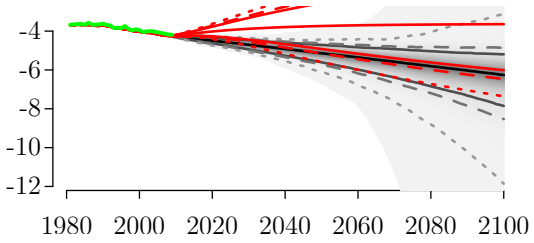
Figure 4.42: External validation for η_{xpgt} for the age of 60

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 60$ for all countries p (by rows) and both genders g (by columns) as time series in t for $2010, \ldots, 2100$, along with ML forecasts (red lines) and crude estimates from observed data (green lines) for the same MCMC output as in Figure 4.39. 90%, 95%, 99%, and 100% credibility and confidence intervals are given by solid, dashed, dotted, and limiting lines, respectively.

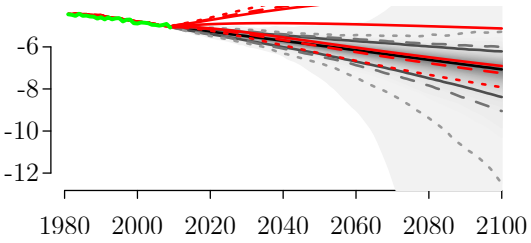


(a) Austrian males

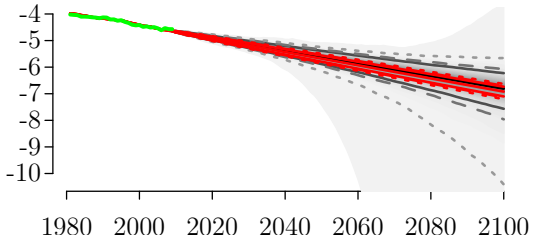
(b) Austrian females



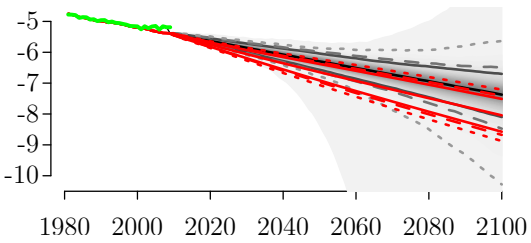
(c) Czech males



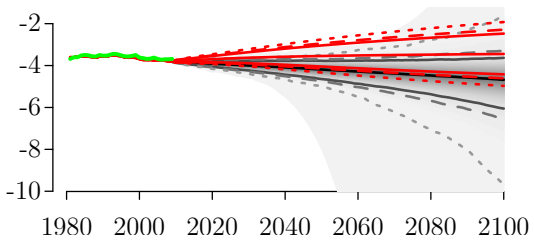
(d) Czech females



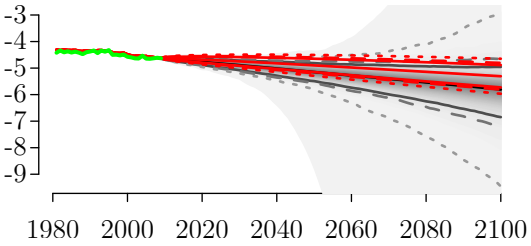
(e) German males



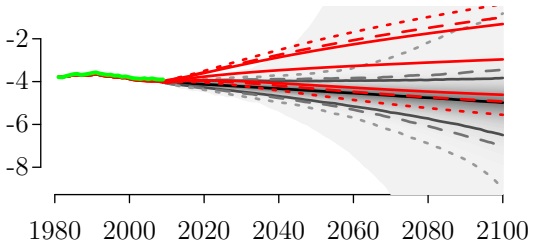
(f) German females



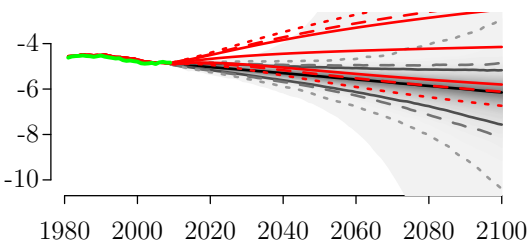
(g) Hungarian males



(h) Hungarian females



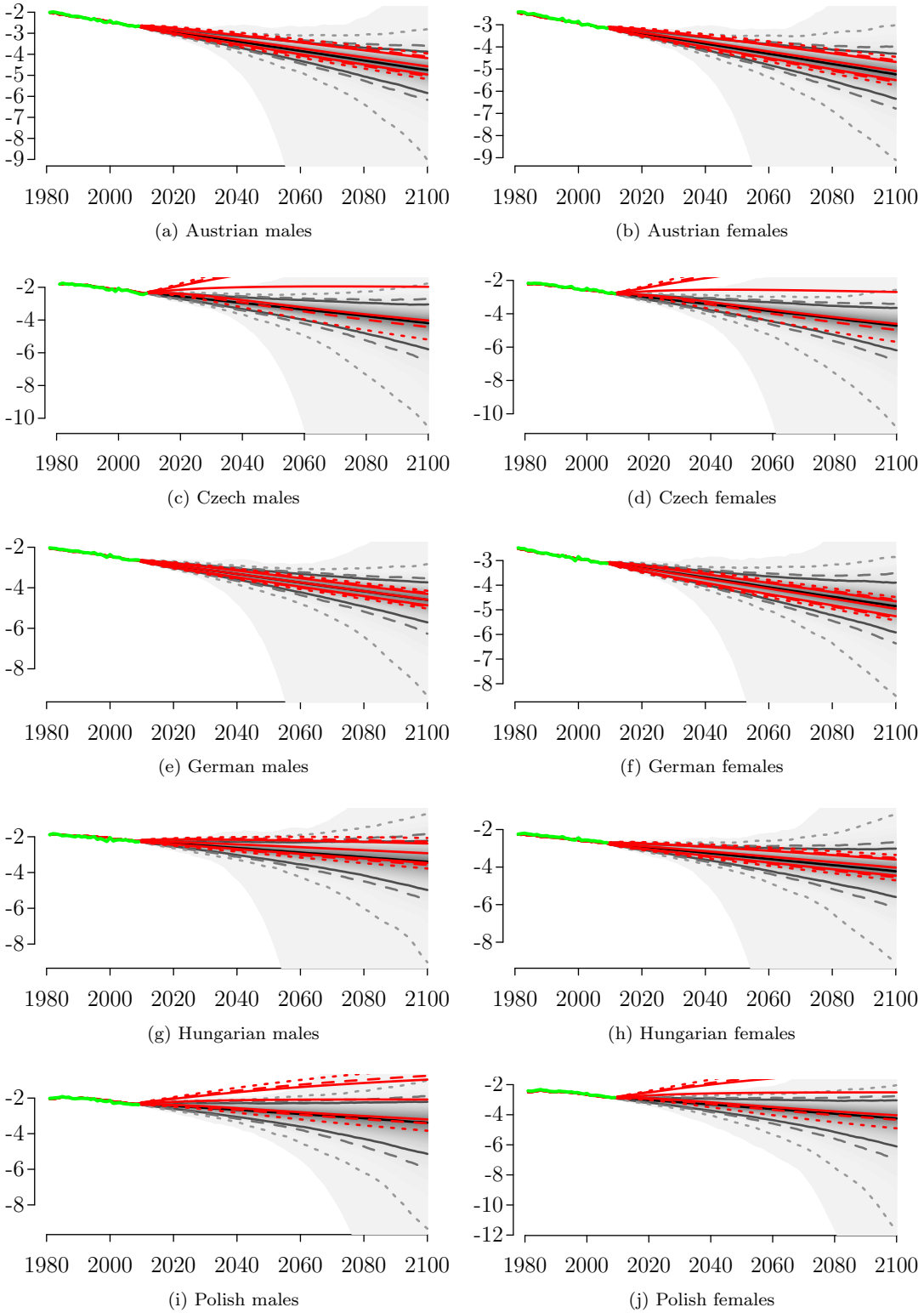
(i) Polish males



(j) Polish females

Figure 4.43: External validation for η_{xpgt} for the age of 80

Shown are fancharts of marginal posterior predictive distributions for future η_{xpgt} with $x = 80$ for all countries p (by rows) and both genders g (by columns) as time series in t for $2010, \ldots, 2100$, along with ML forecasts (red lines) and crude estimates from observed data (green lines) for the same MCMC output as in Figure 4.40. 90%, 95%, 99%, and 100% credibility and confidence intervals are given by solid, dashed, dotted, and limiting lines, respectively.



With all priors chosen as in Section 4.1.2 and the algorithm run as outlined in Section 4.1.3, different specifications for k and r are compared, and the choice of one for both the lag and cointegration orders is concluded to yield a parsimonious and well-fitting model. Figures 4.37 to 4.43 show the starting values and posterior predictive distribution plots for both internal and external validation, including comparison with ML estimation. The outcomes will be assessed in the remainder of this section. Convergence plots depict a very similar behaviour as those investigated in the previous case study, and are therefore omitted here for convenience.

The starting values in Figure 4.37 somewhat differ from the last case study, particularly for the intercept and main gender effect. When interpreting the combined evolution, the shift from five Western to five Central European countries leads to a negative shock in mortality improvement for men around the year of 1990. This bump, i.e. a sudden increase in logits of mortality rates around 1990 before turning to a fast-paced negative trend, is observed for most Central and Eastern European countries, and is often argued to be an effect of the fall of the Iron Curtain. While population-specific parameters for Austria, the Czech Republic, and Germany remain merely stable, the time series for Hungary and Poland indicate some regime changes and should hence be treated carefully. Internal validation in Figures 4.38 to 4.40 is as satisfying as before and does not indicate additional problems with mixing low- and high-mortality countries. ML estimates and crude mortality rates are generally included in the plotted credibility intervals. Strong deviations from the posterior predictive distribution for 60-year-old males of the Eastern European countries in the first half of the calibration window were visible and discussed in the previous case study and no unique feature in this application. Moreover, comparison of forecasts for German data, which are particularly modelled in both case studies, reveals no substantial differences in their quality and, if any, then a slight improvement. Hence the model appears robust w.r.t. the homogeneity in underlying data.

Rather surprisingly, external validation in Figures 4.41 to 4.43 lead to narrower credibility intervals for German mortality predictions than what was previously observed. An explanation might be that, in the first case study, German data were implicitly trained on more similar countries and that their characteristics, e.g. in variability, were correctly incorporated into German forecasts. The current case study combines rather different mortality patterns and effectively leads to a minor extent of information loss. This ex-

planation would motivate the increase in the number of population in joint forecasts using the BMPMP model.

Apart from the comparison of German data, the external validation plots clearly show reasonable fancharts for the country-specific forecasts. A qualitative difference exists between Austrian, Czech, and German data on the one hand, and Hungarian and Polish data on the other. The first-mentioned countries depict a strong negative trend, whereas the remaining countries devote more probability mass to a slower pace or even deteriorating outcomes. A comparison to individual models is not conducted for this case study, but it can be gathered from the analysis in Section 4.1 that the joint model approach is naturally superior to isolated forecasts. What is still done, however, is the comparison of the Bayesian forecasts with their ML counterparts, given by the superimposed red lines in the last to figures. Frequentist confidence intervals can either be much narrower than Bayesian credibility intervals or, what is observed in most instances, depict best estimates that substantially differ from the posterior predictive median values by generally being too conservative, which seems far less plausible.

To summarise, the BMPMP model yields a flexible and robust approach to model even unconventional combinations of populations in an easily interpretable way.

5 Conclusion

In light of modern challenges in stochastic mortality forecasting, the BMPMP model, a Bayesian approach in the multi-population framework, was derived, defined, and applied in this work. This concluding chapter summarises the model's characteristics, advantages, and limitations based on theoretical and empirical findings in Chapters 3 and 4 and addresses potential or even necessary future work.

5.1 Summary and Outcome

In this thesis, the BMPMP model was established to allow for mortality projections of high ages in a globalised world with particular focus on modelling inter-population dependencies and associated uncertainties. The detailed literature review in Chapter 2 revealed that the interest in such approaches has already led to a variety of attempts. However, due to the complexity in high-dimensional forecasting models, an ultimate state-of-the-art technique does not seem to have evolved yet. In fact, the literature review could indeed be used to postulate an extensive list of limitations in existing models.

The BMPMP model was designed to address such problems to provide a concrete tool for improved joint mortality forecasts and a substantial contribution to other future work. It has a two-level hierarchical structure, which is known from most other models and has been successfully established in modern stochastic mortality forecasting. Here, the CBD model is used as framework for observed mortality rates and the VECM is the data-driven technique to project today's patterns into the future. Most notably, the approach differs from existing model by combining such a flexible, easily applicable, and interpretable multi-population model with the non-frequentist paradigm of Bayesian statistics to assure a well-conditioned estimation procedure, biological plausibility, and quantification of uncertainty – as specifically expressed via the model's name. The key outcomes of the BMPMP model are:

- The BMPMP model is a flexible multi-population mortality projection model, which is not generally restricted to a certain number or type of populations.
- Its hierarchical structure with forecasts of CBD parameters and corresponding linear predictors as immediate output lead to an easy, but also detailed, interpretation and analysis. It can be further extended or adjusted for specific needs.
- The Bayesian approach has two important consequences. On the one hand, in contrast to frequentist estimation, it guarantees well-conditioned long-term mortality projections based on comparably short calibration periods. On the other hand, it results in posterior distributions of future model parameters and mortality rates, which are fruitful for coherent quantification of risks and dependencies.
- The empirical analyses of the BMPMP model point towards success in biologically plausible quantification of the interaction between different populations through completely data-driven techniques rather than postulated convergence assumptions. As a consequence, unforeseen possibilities of acceleration or deceleration in mortality improvements are intentionally included.
- Comparison of different model specifications indicated robustness of the model w.r.t. the lag order in the VECM and the possibility of adjusting the uncertainty in mortality predictions via the cointegration rank.
- The empirical case studies for the BMPMP model finally stressed the growing importance of quadratic age effects and diminishing country-specific differences between both genders in general mortality forecasting.

It can therefore be claimed that the BMPMP model generally fulfils the desired outcomes, which initially motivated its derivation. However, the empirical case studies could only give limited insight into the model's characteristics. In particular, analyses of biological plausibility and dependencies were restricted to assessment of a few graphical outputs and cannot draw a conclusive picture. Moreover, despite the apparently valid forecasts in both case studies, the MCMC algorithm for the Bayesian estimation of the model parameters did not fully converge. Although approximations of the posteriors for the hyperparameters by the marginal output yielded satisfying results in this work, they should not be considered to be ultimately sufficient. As a drawback, full convergence

is numerically expensive and requires unwanted computation time, but the model's robustness w.r.t. the lag order indicates that calibration for the cost-efficient case of $k = 1$ suffices. It must further be kept in mind that the CBD model approach does not give the option to model mortality rates for ages below 40. As common in stochastic mortality forecasting, the BMPMP model was also detected to be sensitive towards the calibration window.

However, considering all of the challenges that the model addresses, it still appears to be a strong tool for future applications with high ages, whose flexibility possibly outperforms many of the cited alternatives. Its universal set-up further yields a promising framework for future extensions. The benefits through joint forecasts and the Bayesian paradigm can be expected to be applied with other approaches to provide reasonable mortality forecasts. Most importantly, due to the implementation in the statistical software R and its inherent flexibility, the BMPMP model is immediately available to the interested reader and ready for own applications.

5.2 Future Work

This work defined a completely new model and thoroughly discussed its application in two different case studies. Although much insight could be provided, this work is limited in what had to be analysed for a full assessment of the BMPMP model. For an improved understanding of the model, empirical outcomes need to be studied for a larger variety of applications and longer-lasting MCMC algorithms that have indeed fully converged. Here, based on findings in Chapter 4, focus can be laid on model specifications with lag order $k = 1$, because the model appears to be robust against the lack of AR components while reducing its dimensionality significantly through omission of m-dimensional parameter square matrices. Computation time decreases quadratically and might guarantee this framework to become more convenient.

Furthermore, no sensitivity analysis of the prior choices for the constants introduced in Section 3.3.2 has so far been conducted. It is necessary to infer the effects of changes in such values to further assess the model's robustness and to get a clearer picture of how the analyst can set-up the model a priori based on their needs. Similarly, a discussion of

the performance with given deterministic scenarios as outlined in Section 3.4 would give deeper insight into the model's advantages. The qualitative assessment of the model's ability to project plausible results in a globalised world needs to be extended to a more quantitative and general analysis. The sensitivity towards the calibration period can be addressed specifically and used for future development of this approach. Based on the findings in the empirical case studies, future work can further be undertaken to widen the model's short-term credibility bands and to reduce its high-dimensionality by thinning out unnecessary parameters in the parameter matrices. Of course, to bend the bow to the very introduction to this work, an application of the mortality outcomes in an actuarial context, as common in the cited literature, needs to be done for assessment of the model's applicability to quantify benefits of the improved knowledge of risks and dependencies when pricing of life insurances and pension funds. Summarising, with the positive results in this work, the analysis of the BMPMP model has just begun.

APPENDICES

A Bayesian Statistics

Through substantial growth in computer efficiency during the last decades, computer-intensive numerical techniques have become more and more popular. *Bayesian statistics*, named after Thomas Bayes (1702–1761), is a prominent example among these methods and provides a notable alternative to standard methodologies in statistical data analysis. In the presence of Bayesian statistics, the well-established approaches for drawing conclusions about unknown quantities from numerical data, which build upon hypothesis testing and confidence intervals, are commonly referred to as *frequentist inference*. It emphasises the interpretation of results as probabilistic statements about infinite sequences of the experiment under consideration. Since Bayesian and frequentist inferences differ in their basic philosophies, the core features of both paradigms are reviewed in Section A.1. Details on the actual inference conducted in a Bayesian framework are presented in Section A.2. Sections A.3 and A.4 are particularly devoted to hierarchical approaches and model diagnostics in the Bayesian context, respectively. The discussion in the entire section is mainly based on the excellent standard textbook on Bayesian statistics by Gelman et al. (2013), which is recommended for a thorough review on this topic.

A.1 Pragmatic Comparison of Frequentist and Bayesian Statistics

Given a probabilistic model for all observed quantities in an underlying scientific problem, in frequentist estimation any unknown model parameter θ is generally assumed constant. The rationale is that even if a parameter cannot be observed, there exists one true value and randomness stems from natural deviations of anything unknown when experiments are repeated. The fundamental measure of such uncertainty is captured by probability, which in frequentist's terms is thought of as the relative frequency of an

event in a very long, theoretically infinite, sequence of the same experiment, conducted independently of each other. Probabilistic statements in this classical framework have to be interpreted in terms of future experiments and not, as usually but falsely done, in terms of the currently observed data. For example, a $100p\%$ confidence interval includes the true but unknown value θ with confidence $p \in (0, 1)$, as it is a realisation of a random interval containing the true value with probability p in the limit of repeated experiments. However, the realised interval cannot be stated as a fixed range in which the unknown parameter lies with probability p, although it generally serves as good approximation and is hence interpreted as such for convenience. Similarly, given a statistic to test a null hypothesis H_0 related to the problem, the corresponding frequentist p-value is not the probability that H_0 is true, as it is commonly said to be, but the probability of observing a result at least as extreme for the outcome under the null distribution in a sequence of similar inferences. It is natural to use the p-value as a measure of inconsistency between the data and the hypothesis, but it does not tell anything about the likelihood of H_0 being true. A more detailed review on frequentist interpretations in light of Bayesian statistics can be found in Dobson and Barnett (2008).

Conversely, the Bayesian approach makes use of an alternative paradigm, in which probability statements are applied with *both* observed and unobserved quantities, i.e. the sampled data and the parameters of interest. A consequence is that uncertainty is quantified explicitly through probability, which is the key philosophy in Bayesian statistics. Gelman et al. (2013) describe this methodology in three steps. First, a joint probability model for the data y and parameters θ is postulated based on scientific knowledge of the underlying problem, eventually including dependence on additional explanatory variables. After collecting sample data, the conditional probability distribution of the parameters of interest given the observed data is derived. Inference is based on this *posterior distribution* as it combines the general probabilistic assumption on θ, referred to as the *prior distribution*, and what is learned from the data under the full model. The third step comprises tools for analysing the model fit and sensitivity towards model assumptions. Since the model is set up via a full probability approach, any probabilistic statements can be immediately interpreted as such in a common sense without relating it to a sequence of independent repetitions. A $100p\%$ probability interval then expresses a range for the quantity of interest with coverage probability p and a p-value is interpreted as the probability of replicated data being more extreme than observed data evaluated

under a specified test statistic. Besides the advantages of common-sense interpretation, Bayesian statistics is said to be appealing due to a reduced impact of overparametrisation, in particular when using hierarchical models, in which the number of parameters may even exceed the number of data points. Generally, the freedom of a full probability model for all quantities enables the analyst to model complex problems by models which are not restricted to be too simplistic. More details on hierarchical models are given in Section A.3, whereas for a general overview on the usage of Bayesian inference the reader is referred to Gelman et al. (2013).

A.2 Bayesian Inference

Given the probability distribution for the parameters of interest and one set of realised data, the consequence of Bayesian statistics is that the posterior distribution for θ, on which all inference is based on, depends on the observed values, which reverses the conditioning of probabilistic statements known from frequentist approaches. Conditional on the given data, the posterior is given by $p(\theta \mid y) = p(\theta, y)/p(y)$, which depends on θ only through the postulated full probability model $p(\theta, y)$. For the sake of simplification in notation, any dependence on covariates is dropped in this chapter. Noting that $p(\theta, y) = p(y \mid \theta)p(\theta)$ is the product of the sampling distribution $p(y \mid \theta)$ and the prior distribution $p(\theta)$, the posterior distribution is ultimately obtained via the well-known Bayes' rule, i.e.

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}.$$

The normalising constant $p(y) = \int p(y \mid \theta)p(\theta)d\theta$, which is evaluated as a sum in case of discrete θ, is considered constant for the given realisation of the data and Bayesians usually limit their attention to the non-normalised posterior

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta),$$

which summarises the key tasks in Bayesian inference: formulating an appropriate probability model, usually done by breaking down the problem into finding suitable sampling and prior distributions, and computation of the posterior distribution. Since the data are considered the fixed outcome of one experiment, $p(y \mid \theta)$ is read as a function in θ for

fixed y and called the *likelihood*. Bayesian inference hence follows the so-called *likelihood principle* such that identical inference is obtained for θ if the underlying probability models share the same likelihood functions for given y.

If statistical conclusions about the unobservable parameters are the aim of Bayesian data analysis, posterior distributions are the main tools to use. Of course, Bayesian methodology is also useful for *predictive inference*, i.e. probabilistic statements about observable quantities, for which the sampled data y constitute one realisation. Before this sample is collected, the probability distribution for the unknown data is given by the normalising constant $p(y)$, which can be computed by integrating over both the likelihood and the prior, as stated above. Gelman et al. (2013) call this marginal distribution the *prior predictive distribution* to emphasise the independence of previous observations and the fact that y is observable. After data y have been collected, the additional information leads to the *posterior predictive distribution* for the random quantity $\tilde{y}$, which stands for further observable but yet unknown outcomes of the same quantity that generated y. It is computed as

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y}, \theta \mid y) d\theta \\ &= \int p(\tilde{y} \mid \theta, y) p(\theta \mid y) d\theta \\ &= \int p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta, \end{aligned}$$

i.e. the average of the likelihood weighted by the posterior distribution for θ. The last equality follows from the conditional independence of $\tilde{y}$ and y given θ. Apart from the purpose of predictive inference, the posterior predictive distribution is a fundamental tool in model diagnostics, as further discussed in Section A.4.

Critics in favour of frequentist statistics often state the dependence on the rather subjective choice of a prior distribution and the comparably difficult derivation of the posterior distribution. Whereas the latter problem has significantly decreased over the last decades with availability of computer-intensive numerical methods, mainly *MCMC* procedures which are presented in Appendix B, Gelman et al. (2013) point out that scientific reasoning must be applied even for specification of the likelihood, i.e. also in

frequentist methodology. In particular, assumptions such as a specific error distribution in a regression model can be similarly subjective as the choice of the prior for underlying unknown quantities. Statistical inference relies to large extent on the chosen model, and the specification of a prior distribution can be seen as part of the process of scientific judgement on which model to choose. Moreover, accuracy of prior assumptions and the corresponding sensitivity of posterior results can be similarly examined as it has to be done in frequentist models. In practice, prior distributions are specified using information which is already at hand, e.g. opinions from experts or results from previous studies. In this case, the prior distribution is called *informative* and, analogously, the prior is called *non-informative* if it does not contain any information based on prior beliefs. Prior distributions are referred to as *conjugate*, if the posterior distribution remains in the same family, i.e. the observed data only update the underlying parameters of the distribution. Conjugate priors are convenient in that they offer analytic solutions for the posterior distributions and are often chosen for convenience, but their necessity has decreased through improvements in the numerical computation of posteriors. It is worth mentioning that prior distributions need not be proper per se, i.e. they may not integrate to one. Such *diffuse* priors, e.g. an unbounded uniform prior, which are useful options for non-informative priors, can still lead to proper posteriors and Bayesian inference is hence possible.

A.3 Hierarchical Models

Due to inclusion of a full probability model for both data and parameters in the Bayesian framework, this methodology makes it particularly easy to use models whose parameters reflect a high level of dependence on each other. With decompositions of joint distributions by $p(\theta_1, \theta_2) = p(\theta_1 \mid \theta_2)p(\theta_2)$, it is natural to express such complex models in a hierarchical fashion. Depending on the scientific context, the parameters θ may be modelled to follow a distribution with another set of unobservable parameters ϕ, the so-called *hyperparameters*, which are of own interest to the modeller. The data y are considered independent of ϕ given the parameters θ. The joint probability distribution can then be simply written as

$$p(y, \theta, \phi) = p(y \mid \theta, \phi)p(\theta, \phi) = p(y \mid \theta)p(\theta \mid \phi)p(\phi).$$

One sees that the joint prior distribution for θ and ϕ is now given by $p(\theta, \phi) = p(\theta \mid \phi)p(\phi)$ and the according non-normalised posterior is

$$p(\theta, \phi \mid y) \propto p(\theta, \phi)p(y \mid \phi, \theta) = p(\theta, \phi)p(y \mid \theta).$$

The formulation of the more complex model results in specification of a prior for ϕ rather than θ. With an increasing number of levels in the model, prior information on the hyperparameters becomes less available, and non-informative prior distributions must be employed. Common choices are diffuse priors, which allow for maximum flexibility, but the posterior distribution must be carefully analysed, either analytically or by inspection of the resulting variation, in order to verify that it is a proper distribution. Predictive inference and model diagnostics can now be based on different versions of the posterior predictive distribution. Assume that $\theta = (\theta_1, \ldots, \theta_J)'$ comprises several parameters for mutually exclusive blocks of data, and the individual θ_j's are realisations from a superpopulation whose distribution is governed by ϕ. The posterior predictive distribution should be computed solely based on the inference made for a single θ_j, if further outcomes $\tilde{y}$ are to be predicted for the j-th block. Predictions of future values $\tilde{y}$ for an entirely different block refer to a posterior predictive distribution depending on future values θ_j, which are themselves first drawn from the superpopulation based on the posterior for ϕ.

Hierarchical models are useful when latent values in θ should be analysed in more detail or when the nature of the scientific problem calls for such complex dependencies. For example, count data are often modelled using the Poisson distribution for ease in calculation and interpretation. However, data may exhibit overdispersion, i.e. the observed variance exceeds the theoretical variance of the model, which in this case would coincide with the mean. Modelling the Poisson parameter via, say, the two-parametric Gamma distribution allows mean and variance to disagree. In particular, Gelman et al. (2013) describe how hierarchical models with more parameters than actual data points can be used for valid inference in Bayesian statistics. They also stress that ignorance of hierarchical structure that is seen in the data, generally leads to failure in fitting large datasets, when there are only few parameters in the non-hierarchical model, or to overfitting, if the number of parameters is too large.

A.4 Model Diagnostics

As mentioned earlier, Gelman et al. (2013) characterise Bayesian data analysis in three steps. The first two steps, choice of a suitable probability model for all quantities and computation of the posterior, were outlined in Section A.2. The third step encompasses assessment of model fit, which will be discussed in what follows. Determination of the underlying likelihood and prior distribution is based on scientific knowledge, which may be limited so that assumptions are not accurate and the resulting poor model will yield biased inference. Therefore, it is crucial for any statistical analysis to check the outcome of a model for adequacy in fit and plausibility w.r.t. the purpose of the analysis.

Conceptually, model diagnostics in Bayesian statistics are very similar to standard methods in the frequentist framework. The main approach is to compare theoretical values under the resulting model with observed data. *External validation* is one of the most useful tools, especially for the purpose of prediction: fresh data on y are collected through further sampling or experiments and their characteristics are compared to the theoretical counterparts from the posterior predictive distribution. This is usually done via quantitative summaries, e.g. empirical averages and theoretical posterior means or nominal and true coverage probabilities of Bayesian probability intervals should agree, respectively. However, collection of new data is often not feasible or even impossible, and one has to turn to *internal validation*, also called *posterior predictive checking*, instead. Here, one checks whether the already observed data y appear plausible under the posterior predictive distribution

$$p(y^{rep} \mid y) = \int p(y^{rep} \mid \theta) p(\theta \mid y) d\theta.$$

In practice, one simulates sufficiently many realisations of the parameter θ from $p(\theta \mid y)$ and for each outcome, a sample y^{rep} of the same size as the original dataset is generated from $p(y^{rep} \mid \theta)$. The notation y^{rep} is used to indicate that these data are obtained under the very same conditions as y, whereas the previously used $\tilde{y}$ may also denote values from other experiments or depending on other underlying covariates. Note that for hierarchical models, an appropriate choice for the posterior predictive distribution should be made, as outlined in the previous section. Various numerical and graphical techniques for comparison of observed and simulated values are possible. One may display all data

for several replications if the dataset is rather small. Computation and plotting of summary statistics such as mean values or quantiles over all replicates and the observed data, respectively, is more suitable for large datasets. An alternative is the computation and visualisation of *Bayesian residuals*, which will not be further discussed in this work. For more details on graphical tools for posterior predictive checks, the reader is again referred to Gelman et al. (2013). It is worth mentioning that for prediction purposes, advanced internal diagnostics such as cross-validation can be easily adopted, of course.

Although qualitative examination of predicted and observed data gives much insight into the accuracy of the fit, it is desirable and scientific practice to base the final conclusion of whether a model is appropriate or not, on well-defined quantitative decision rules, such as hypothesis test statistics in frequentist analysis. Gelman et al. (2013) define the Bayesian equivalent $T(y, \theta)$, which can be any suitable scalar summary of parameters and data, as *discrepancy measure*. If such a measure only depends on the data, i.e. $T(y, \theta) = T(y)$ for all θ, one speaks of a *test statistic*. Decision rules are quantified in terms of p-values, i.e. tail-area probabilities of the discrepancy measure. In analogy to frequentist statistics, the *classical p-value* for a test statistic $T(y)$ with a fixed value for the parameter θ is defined as

$$p_C = P(T(y^{rep}) \geq T(y) \mid \theta),$$

i.e. the probability of observing replicated data at least as extreme as the sampled data given the specified value θ. The definition makes use of the fact that $T(y^{rep} \mid y, \theta) = T(y^{rep} \mid \theta)$. The value θ is either chosen to be a null value, which is tested for plausibility, or a substituted point estimate, e.g. an ML estimate in frequentist analysis, when the model accuracy is to be assessed. In contrast to the classical approach, the *posterior predictive p-value* for model fit, going back to the influential work by Rubin (1984), uses the inferred posterior distribution for θ to compare replicated data under the posterior predictive distribution with observed data. Since the randomness in θ is inherently accounted for, there is no need for pre-specified values for θ to be kept fixed such that the discrepancy measure $T(y, \theta)$ can indeed be a function of both the data and the parameters. The Bayesian p-value for $T(y, \theta)$ is given by the probability under (θ, y^{rep}) given y, i.e. the joint posterior and posterior predictive distribution given the data, that the test quantity evaluated at the posterior pair (θ, y^{rep}) exceeds its outcome

for θ combined with the observed sample y, i.e.

$$\begin{aligned} p_B &= P(T(y^{rep}, \theta) \geq T(y, \theta) \mid y) \\ &= \int \int I_{T(y^{rep},\theta) \geq T(y,\theta)} \; p(y^{rep} \mid \theta) \; p(\theta \mid y) \; dy^{rep} \; d\theta \end{aligned}$$

with indicator function I, again noting that $p(y^{rep} \mid \theta, y) = p(y^{rep} \mid \theta)$. The formula implies that Bayesian p-values can be conveniently computed by first simulating from the posterior distribution for θ and subsequent generation of corresponding predictive values using $p(y^{rep} \mid \theta)$. The resulting realisations are draws from the joint distribution $p(y^{rep}, \theta \mid y)$ and the p-value is estimated by the fraction of simulations for which the inequality holds. Graphical visualisation of p-value computation is of additional help in model assessment compared to a single numerical output. In case of test statistics, histograms for the posterior distribution of $T(y^{rep})$ depict the tail-area probability w.r.t. the observed threshold $T(y)$. For discrepancy measures that also depend on θ, histograms for the posterior difference $T(y^{rep}, \theta) - T(y, \theta)$ should include 0 if the model provides a good fit, and scatterplots of posterior $T(y^{rep}, \theta)$ versus posterior $T(y, \theta)$ show the p-value as proportion of points in the upper half of the first quadrant. Indeed, Gelman et al. (2013) recommend to base any conclusions about the fit of a model not only on the single p-value but also on assessment of the magnitude of discrepancy detected in the plots. Generally, a p-value close to 0 or 1 indicates possibly severe discrepancy between the model and the observed data. It can be directly interpreted as the posterior probability of seeing the observed sample in replicated data and is as such a measure of statistical significance, however even in Bayesian analysis it *cannot* be stated as the probability of the model being true given the data.

There is no general guideline for the choice of the discrepancy measure, because it depends on the scientific problem how well a summarising scalar can measure discrepancy between observed and simulated data. Common practice is to account for features which are not primarily addressed by the full probability model in order to avoid optimistic results. Therefore, in most examples one would rather use summary statistics of less interest such as the rank of a sample rather than, say, the mean value or, by their very definition, analysis of the residuals to account for remaining uncertainty. Generally, finding meaningful discrepancy quantities is far easier in Bayesian statistics compared

to the construction of pivotal frequentist test statistics. Besides such particular tests, Gelman et al. (2013) suggest *omnibus tests*, i.e. tests on comparison of explained and unexplained variance, as additional model checks. For an observed sample with values $y_i, i = 1, \ldots, n$, important examples are the χ^2 *discrepancy quantity*

$$T(y, \theta) = \sum_{i=1}^{n} \frac{(y_i - \mathrm{E}(y_i \mid \theta))^2}{\mathrm{Var}(y_i \mid \theta)} \tag{A.1}$$

or the *deviance* given by $T(y, \theta) = -2 \log(p(y \mid \theta))$. In frequentist analysis, plugging in null values θ_0 for certain hypotheses yields hypothesis tests with statistics $T(y) = T(y, \theta_0)$. It is possible to use such test statistics for the replicated data y^{rep} in posterior predictive model checks, too. Similarly, by plugging in a point estimate $\widehat{\theta}(y)$ into (A.1), e.g. an ML estimate, one obtains a classical χ^2 *goodness-of-fit* test $T(y) = T(y, \widehat{\theta}(y))$. Although in Bayesian model diagnostics one could compute $T(y^{rep})$ based on the replicated data and their corresponding estimate $\widehat{\theta}(y^{rep})$, Gelman et al. (2013) recommend applying the discrepancy measure $T(y, \theta)$ with the posterior outcomes for θ directly, because this type of inference does not require any computational burden in parameter estimation. The reference distribution for the Bayesian omnibus test is immediately given through the distribution of $T(y^{rep}, \theta)$ based on the posterior predictive simulations y^{rep}.

Note that the previous discussion was solely devoted to goodness-of-fit analyses for one given model. An elaborate theory has been developed on the important diagnostic tool of model comparisons and sensitivity analysis. In particular, the *deviance information criterion* as Bayesian extension of the Akaike and Bayesian information criteria, developed by Spiegelhalter et al. (2002), is a strong tool to test different models against each other by taking into account both the model's fit and complexity. Since model comparisons are not discussed in this work, the reader is referred to the aforementioned literature for further information.

B Markov Chain Monte Carlo

Although Bayesian inference is traced back to work by Thomas Bayes (1702–1761) and Pierre-Simon Laplace (1749–1821), see e.g. Stigler (1986), its usage has been denied to all but the simplest problems for which the theoretical posterior distribution for the parameters could be derived. In more complex problems, the normalising constant in the product of prior and likelihood, which is necessary for a well-defined density or probability mass function, could not be analytically computed. As a result, prior distributions had to be chosen to be conjugate w.r.t. the likelihood in order to meet computational feasibility rather than any scientific reasoning. With the fast increase in computational power, however, Bayesian statistics has seen a dramatic growth. The innovations in numerical mathematics now allow for feasible approximations of posterior distributions for general choices of priors, even in high-dimensional and hierarchical problems. In particular, history of modern Bayesian statistics is closely connected to the development of MCMC. Rather surprisingly, MCMC algorithms solve the difficult problem of simulation from a complex distribution, which is only known up to a normalising factor and not suitable to generate from by plain Monte Carlo methods, by another complex tool: the construction of a correspondingly high-dimensional discrete-time Markov chain with continuous state space for the parameters of interest. Well-known algorithms have been developed, for which the resulting Markov chain has the theoretical posterior distribution as a stationary distribution. Although the existence of a unique limiting distribution, which in this case would be the posterior, cannot be proven in most applications due to the implicit complexity – which in fact requires the use of MCMC –, this procedure has become undoubtedly successful. This appendix provides a review on general Markov chain theory in Section B.1 and introduces the two main sampling algorithms, the Gibbs and Metropolis-Hastings samplers, in Sections B.2 and B.3. Section B.4 concludes with comments on convergence diagnostics. The review builds upon the excellent discussions by Robert and Casella (2004) and Gamerman and Lopes (2006) on MCMC and by Meyn and Tweedie (2009) on the theory of Markov chains with continuous state spaces.

B.1 Markov Chain Theory

The purpose of MCMC in Bayesian statistics is to construct a Markov chain, whose state space equals the parameter space, and which converges to a unique limiting distribution, which coincides with the posterior distribution. Since parameter spaces are generally of continuous nature, the well-known theory on discrete-space Markov chains must be expanded correspondingly. It is worth mentioning that the index set, which represents the iteration steps in the algorithm, is still discrete such that the term Markov *chain* is indeed appropriate.

Let $S \subseteq \mathbb{R}^d$ be the possibly continuous parameter space for the d-dimensional parameter vector of interest in a probability model, and let $\mathcal{B}(S)$ be the set of all Borel sets on S. A *transition kernel* is a function $P\colon S \times \mathcal{B}(S) \to [0,1]$ for which $P(x,\cdot)$ is a probability measure on S for all $x \in S$, and $P(\cdot, A)$ is measurable for all $A \in \mathcal{B}(S)$. For such P, the corresponding *Markov chain* with state space S is defined as the discrete stochastic process $X := \{X^{(n)}\colon n \in \mathbb{N}_0\}$ with the Markov property, i.e. the *transition probabilities* fulfil

$$\begin{aligned} &P\left(X^{(n+1)} \in A \mid X^{(n)} = x_n, \ldots, X^{(1)} = x_1, X^{(0)} = x_0\right) \\ &= P\left(X^{(n+1)} \in A \mid X^{(n)} = x_n\right) = \int_A P(x_n, dx) \end{aligned}$$

for all $n \in \mathbb{N}_0, x_0, x_1, \ldots, x_n \in S, A \in \mathcal{B}(S)$. The Markov chain is *homogeneous* if the distribution of $X^{(n_1)}, \ldots, X^{(n_k)} \mid X^{(n_0)}$ equals the distribution of $X^{(n_1-n_0)}, \ldots, X^{(n_k-n_0)} \mid X^{(0)}$ for all $k \in \mathbb{N}$ and $n_0, n_1, \ldots, n_k \in \mathbb{N}_0$ with $n_0 \leq n_1 \leq \cdots \leq n_k$, i.e. it is invariant w.r.t. shifts in the index. For $n \in \mathbb{N}$, the *n-step transition kernel* $P^{(n)}\colon S \times \mathcal{B}(S) \to [0,1]$ is then recursively given by $P^{(n)} = P$ for $n = 1$ and

$$P^{(n)}(x, A) = \int_S P^{(n-1)}(y, A) P(x, dy), \quad x \in S, A \in \mathcal{B}(S)$$

for $n > 1$. If $\pi^{(0)}$ denotes the probability measure on S for the initial state, one obtains $\pi^{(n)} = \int_S P^{(n)}(x,\cdot)\pi^{(0)}(dx)$ as the distribution of the state of $X^{(n)}$.

In order to define stationary and limiting distributions for Markov chains with continuous state spaces, let π be a probability measure on S. Then π is called a *stationary*

distribution for the Markov chain X with transition kernel P if

$$\pi(A) = \int_S P(x, A)\pi(dx)$$

for all $A \in \mathcal{B}(S)$. The definition shows that once the current state's distribution of X is equal to π, all future states are also distributed according to π, and X has reached an equilibrium. However, a Markov chain may have several stationary distributions. In MCMC, one is interested whether X has a unique stationary distribution to which the Markov chain will ultimately converge. Denoting the *total variation norm* between two probability measures ξ_1 and ξ_2 as a measure of dissimilarity by $\|\xi_1 - \xi_2\| := \sup_{A \in \mathcal{B}(S)} |\xi_1(A) - \xi_2(A)|$, the *limiting distribution* of X with transition kernel P and initial state distribution $\pi^{(0)}$ is the unique probability measure π which fulfils

$$\lim_{n \to \infty} \left\| \pi^{(n)} - \pi \right\| = 0,$$

if this limit exists.

The following attributes of a Markov chain are important for the broad theory on necessary and sufficient characteristics for the existence of stationary and limiting distributions. A Markov chain is said to be π-*irreducible* if π is a probability measure on S and for all $A \in \mathcal{B}(S)$ with $\pi(A) > 0$, it holds that there is a non-zero probability of X reaching A in finitely many steps for any initial state[1]. In an informal way, irreducibility assures that it does not matter how the Markov chain is initialised, to guarantee that all non-trivial sets under π can be reached. It can be shown that a π-irreducible Markov chain has the unique stationary distribution π. A stronger assumption is the Harris-recurrence named after the American mathematician Theodore Harris (1919–2005), which extends irreducibility to infinitely many visits of non-trivial sets under π given any starting value. A π-irreducible Markov chain is *Harris-recurrent* if for every $A \in \mathcal{B}(S)$ with $\pi(A) > 0$, the number of visits $\sum_{n=1}^{\infty} I(X^{(n)} \in A \mid X^{(0)} = x)$ given any $x \in S$ equals infinity, π-

[1] To be precise, Markov chain theory w.r.t. continuous state spaces defines π-irreducibility for general measures on S rather than probability measures. For this review, it suffices to focus on the interesting case of probability measures; however, the terminology is directly restricted to this special case and looses some of its generality. For the very exact definitions of irreducibility and other terms, the reader is referred to the literature referenced in this section.

almost surely. The third important property is aperiodicity, i.e. the non-existence of any deterministic transition pattern which, for example, could allow an alternating sequence of stationary distributions without a well-defined limit. A π-irreducible Markov chain is called *periodic* if there is a $d \in \mathbb{N}$ with $d > 1$ and a sequence of non-empty and disjoint sets $E_0, E_1, \ldots, E_{d-1} \in \mathcal{B}(S)$ with $\pi(\cup_{i=0}^{d-1} E_i) = 1$ such that for all $i = 0, 1, \ldots, d-1$, it holds that $P(x, E_{(i+1) \bmod d}) = 1$ for all $x \in E_i$. The Markov chain is *aperiodic* if it is not periodic. The previous conditions are conveniently summarised under the the term of ergodicity. A Markov chain X with stationary distribution π is *ergodic* if it is π-irreducibly, Harris-recurrent, and aperiodic. It can be shown that for an ergodic Markov chain with stationary distribution π and arbitrary initial state, the limiting distribution exists and is given by π. It is worth mentioning that in contrast to discrete-space Markov chains, the existence of a stationary distribution is a necessary condition rather than the result of irreducibility, recurrence and aperiodicity. Note also that no statements are made here regarding the pace of convergence, for which the reader is referred to Robert and Casella (2004), for instance.

For a stochastic process, ergodicity can be informally stated as the property that after convergence, the distribution of state occupancy *over time* agrees with the state distribution *at a fixed point in time*. This is the desired attribute of a Markov chain in MCMC, because the sample of realised values should approximate the unknown limiting distribution, which in Bayesian inference represents the posterior. The interpretation is due to the *Ergodic Theorem*, a version of the law of large numbers for dependent realisations of a Markov chain. For a Markov chain X and a function $t\colon S \to \mathbb{R}$, define the *ergodic mean* of $t(X)$ after $n \in \mathbb{N}$ steps via $\bar{t}_n := \sum_{i=1}^{n} t\left(X^{(i)}\right)/n$ and the *expected value* of $t(X)$ under π as $E_\pi(t(X)) := \int_S t(x)\pi(dx)$. The Ergodic Theorem states that if X is π-irreducible and Harris-recurrent with stationary distribution π and $E_\pi(t(X)) < \infty$, then

$$\lim_{n\to\infty} \bar{t}_n = E_\pi(t(X)).$$

There are also central limit theorems applied with Markov chains for the quantity $\bar{t}_n$, which require stronger assumptions on the Markov chain, for which the reader is referred to Robert and Casella (2004).

B.2 Gibbs Sampling

While named after the American physicist Josiah Willard Gibbs (1839–1903) due to an application with the Gibbs distribution in mechanical statistics, the *Gibbs sampler* was developed by Geman and Geman (1984) to become one of the two main sampling approaches to construct a Markov chain in MCMC. Let π be the posterior distribution of the parameter $\theta = (\theta_1, \ldots, \theta_d)' \in S \subseteq \mathbb{R}^d$, which is to be estimated. For simplicity in notification, the general dependence on the data y in the Bayesian context is dropped. The necessity of MCMC implies that π is unknown or its simulation is intractable. The Gibbs sampler assumes that the so-called *full conditionals* of π are known and easy to simulate from, i.e. for each $j = 1, \ldots, d$, the full conditional probability $\pi_j(\theta_j \mid \theta_{-j}) =: \pi_j(\theta_{-j})$ of θ_j given all other parameters $\theta_{-j} := \{\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d\}$ is available. Beginning with a starting value $\theta^{(0)}$ for the parameter θ, in each iteration step n, all entries θ_j are successively visited and simulated according to the full conditionals, where for all other parameters current realisations are used. More specifically, in iteration step n, the following algorithm is run:

$$
\begin{aligned}
\theta_1^{(n)} &\sim \pi_1\left(\theta_2^{(n-1)}, \ldots, \theta_d^{(n-1)}\right), \\
\theta_2^{(n)} &\sim \pi_2\left(\theta_1^{(n)}, \theta_3^{(n-1)}, \ldots, \theta_d^{(n-1)}\right), \\
&\vdots \\
\theta_j^{(n)} &\sim \pi_j\left(\theta_1^{(n)}, \ldots, \theta_{j-1}^{(n)}, \theta_{j+1}^{(n-1)}, \ldots, \theta_d^{(n-1)}\right), \\
&\vdots \\
\theta_d^{(n)} &\sim \pi_d\left(\theta_1^{(n)}, \ldots, \theta_{d-1}^{(n)}\right).
\end{aligned}
$$

The distribution of θ_j is hence determined via the realisations $\theta_1^{(n)}, \ldots, \theta_{j-1}^{(n)}$, which have already been sampled in the current iteration, and the realisations $\theta_{j+1}^{(n-1)}, \ldots, \theta_g^{(n-1)}$ from the previous step for those parameters, which have not yet been visited. This entire procedure is repeated until a sufficiently large sample of values for θ is established.

It is easy to see that the ultimate outcome $\{\theta^{(n)}\}$ yields the path of a homogeneous Markov chain with the parameter space S as state space. Theory on the Gibbs sampler reveals that the distribution π for θ is indeed a stationary distribution for the underly-

ing Markov chain. This statement is the theoretical foundation for the Gibbs sampler, because if a limiting distribution for $\{\theta^{(n)}\}$ exists, it must agree with the then unique stationary equilibrium, and the Markov chain will converge to the posterior. The sufficient condition of ergodicity for the existence of a limiting distribution, however, need not be true for the resulting Markov chain and must be proven individually for each π. Since MCMC is principally applied to simulate from complex distributions, establishing ergodicity is far from trivial. The interested reader is referred to Robert and Casella (2004) for more details on sufficient conditions for ergodicity of a Markov chain obtained through Gibbs sampling. In practice, modellers simply apply the algorithm and empirically check for convergence based on the realised path, see Section B.4. If an equilibrium has been reached, it is accepted as the posterior distribution, and a large number of simulations from the chain provides a discrete pseudo-independent approximation of the theoretical distribution.

Due to its computational efficiency, the Gibbs sampler is a favourable MCMC technique with typically fast convergence. The knowledge of the full conditionals can be further used for consistent continuous estimators of the density for π. Several amendments of the standard algorithm above have been proposed in the literature to improve computational efficiency, speed of convergence, or methods for statistical inference on π. Details are discussed by Robert and Casella (2004). The drawbacks of Gibbs sampling lie in the assumption on the knowledge of all full conditionals, which requires possibly difficult analytical groundwork and may not be met in many applications.

B.3 Metropolis-Hastings Sampling

If full conditional distributions are not available or not feasible to simulate from, application of the Gibbs sampler is not possible. In this case, the most useful alternative is the *Metropolis-Hastings sampler*, a proposal-and-rejection type algorithm named after Nicholas Constantine Metropolis (1915–1999) and W. Keith Hastings (*1930). A special case of this sampling procedure, the *Metropolis sampler*, was first described by Metropolis et al. (1953) to simulate molecules in chemical liquids via the Boltzmann distribution. Hastings (1970) extended this algorithm, which resulted in the more general Metropolis-Hastings sampler. Since the full conditionals need not be known, the mini-

mal requirements on π make this sampling algorithm widely applicable, but convergence is typically slower.

In the Metropolis-Hastings algorithm, the ultimate Markov chain $\{\theta^{(n)}\}$, which should converge against π, is constructed to have a transition kernel of the form $P(x, dy) = \alpha(x, y)q(x, dy)$ for $x \neq y$, where $\alpha\colon S\times S \to [0, 1], q\colon S\times\mathcal{B}(S) \to \mathbb{R}_+$, and $q(x, \cdot)$ is a probability measure for all $x \in S$. If the current state of the Markov chain is x, the transition probability function is decomposed into a proposal density $q(x, \cdot)$ and the probability $\alpha(x, \cdot)$, which can be thought of as the acceptance rate for the suggested new state. For the Metropolis-Hastings sampler, the analyst defines *proposal distributions* $q(x, \cdot)$, which are known up to a constant or at least symmetric, i.e. $q(x, dy) = q(y, dx)$, and which are easy to simulate from. The algorithm only requires the minimal assumption on π that for all $x, y \in S$, the ratio $\pi(dy)/q(x, dy)$ is known up to a constant independent of x, which is particularly useful when π has a normalising constant that cannot be computed. With this ratio, the so-called *Metropolis-Hastings acceptance probability*

$$\alpha(x, y) = \min\left\{1, \frac{\pi(dy)q(y, dx)}{\pi(dx)q(x, dy)}\right\} \tag{B.1}$$

for all $x, y \in S$ with $\pi(dx)q(x, dy) > 0$ is well-defined. With an initial value $\theta^{(0)}$ in the support of π, the n-th step of the Metropolis-Hastings algorithm starts with the simulation of a proposal value $\theta^* \sim q(\theta^{(n-1)}, d\theta)$ based on the current state $\theta^{(n-1)}$. With probability $\alpha\left(\theta^{(n-1)}, \theta^*\right)$, it is accepted as new state $\theta^{(n)}$; otherwise the Markov chain remains at $\theta^{(n-1)}$, i.e.

$$\theta^{(n)} := \begin{cases} \theta^*, & U \leq \alpha\left(\theta^{(n-1)}, \theta^*\right) \\ \theta^{(n-1)}, & U > \alpha\left(\theta^{(n-1)}, \theta^*\right) \end{cases},$$

where $U \sim U([0, 1))$ is independent of θ^*, and $\alpha(\theta^{(n-1)}, \theta^*)$ is given by (B.1). The proposal is always accepted if $\pi(d\theta^*)/q(\theta^{(n-1)}, \theta^*)$ increases w.r.t. $\pi(d\theta^{(n-1)})/q(\theta^*, \theta^{(n-1)})$ and, interestingly, may be still adopted with positive probability in the other case. In many applications, q is chosen to be symmetric such that the Metropolis-Hastings acceptance probability reduces to the likelihood ratio $\alpha(x, y) = \min\{1, \pi(dy)/\pi(dx)\}$.

The resulting Markov chain is homogeneous with transition kernel given by $P(x, dy) =$

$\alpha(x, y)q(x, dy)$ and stationary distribution π. As with the Gibbs sampler, this result expresses the main motivation behind this algorithm, because the Markov chain will converge against π if a limiting distribution exists. Again, sufficient conditions for the existence of such a limit cannot be claimed in general due to the dependence on π and q; however, ergodicity is easier to establish at the design stage through the analyst's freedom in the choice of the proposal distribution, see Robert and Casella (2004). Detailed discussions are also given for special cases of the universal algorithm above, for example when the proposal distribution $q(x, dy) = q(dy)$ is independent of the current state, or – as in the original Metropolis sampler – it is a symmetric distribution centred at the current value. While the Gibbs sampler requires more groundwork on analytical expressions of the full conditionals, the Metropolis-Hastings sampler requires so-called *tuning*, a careful analysis of the outcome to assess the performance of the chosen proposal distribution q. If the range of proposals is chosen to be very close to the current value, innovations in the path of the Markov chain are small and many iterations are necessary to achieve a good level of mixing w.r.t. π. On the other hand, the opposite extreme of a rather uninformative q may lead to many proposals outside the range of π and the number of accepted proposals can be vanishingly small so that the chain stays in the current state for a long time. Both scenarios imply slow convergence and, after reaching stationarity, low mixing such that many steps of the algorithm are required. In these cases, the proposal distribution must be tuned to achieve a good balance between the acceptance rate and the variability in the path of the Markov chain.

B.4 Convergence Diagnostics

Once a sample path of a Markov chain $\{\theta^{(n)}\}$ has been created via any of the above MCMC procedures, convergence diagnostics must be conducted to assess the validity of the ergodic sample as a pseudo-independent sample of the posterior distribution π. Such diagnostics aim to justify the existence of a unique limiting distribution (if this could not be established a priori via theoretical arguments) and, if so, whether and when the actual equilibrium has been reached. The latter implies detection of the *burn-in period*, i.e. the time which is required for the sample path to leave its initial set-up and reach the stable support of π. The iterations within the burn-in period must be discarded for the ergodic sample to be meaningful. Moreover, convergence diagnostics tools play the

central role in tuning of Metropolis-Hastings algorithms.

There exist both theoretical and empirical approaches for convergence diagnosis in the literature on MCMC. However, theoretical tools suffer from the reliance on the true distribution π, which in MCMC applications is often not fully known or too complex to work with analytically. As a consequence, convergence diagnostics are generally only based on empirical approaches for realised paths of the Markov chain, although statistical errors prevent them from being correct with absolute certainty. Most commonly, an informal graphical analysis is conducted, which includes some of the following:

- Plots of the paths of the marginal components $\theta_j^{(n)}$ versus the iterations n to check for an equilibrium after a possible burn-in period.
- Scatterplots of the empirical bivariate distribution for θ_i and θ_j with $i \neq j$, taken over time, to check for an equilibrium with extreme observations only stemming from a burn-in period.
- Plots of the paths of the marginal ergodic means and variances for $\theta_j^{(0)}, \ldots, \theta_j^{(n)}$ versus the iterations n to check for convergence against a constant value.
- If available, plots of the marginal distributions after rejection of a potential burn-in period for several different initialisations $\theta^{(0)}$ of the algorithm to check for identical distributions.

Convergence plots give much insight into the behaviour of the Markov chain and point out possible flaws in the algorithm. Analyses, however, have to be done carefully, because a stationary behaviour in a realised path that is hypothesised to be the limit, may only be due to what is usually referred to as *metastability*, i.e. a temporal equilibrium caused by pure chance or the existence of several stationary distributions. Moreover, inference for the convergence of a multidimensional quantity can be biased since only marginal information is used.

More formal techniques have been proposed over the last decades to overcome the tendency of subjectivity in assessing graphical output. For a function $t\colon S \to \mathbb{R}$, Raftery and Lewis (1992) suggest to compute the burn-in length and the subsequent number of required iterations to estimate confidence boundaries for $t(\theta)$ with pre-specified confidence level and error tolerance. Robert and Casella (2004) point out that, with this

method, convergence is only analysed for the confidence boundaries rather than the entire Markov chain. Another approach is derived from common time series analysis. If the sample path is believed to have reached an equilibrium, ergodic sample averages can be computed for different windows at the beginning and end of the path, respectively. Under the hypothesis of convergence of the Markov chain, both empirical means should be similar. In fact, denoting by $\bar{t}_a$ and $\bar{t}_b$ the versions of the average $\bar{t} = \sum_i t(\theta^{(i)})/n$ at both windows, respectively, for simultaneously increasing sample sizes the distribution of the pivotal quantity

$$\frac{\bar{t}_a - \bar{t}_b}{\sqrt{\widehat{\mathrm{Var}}(\bar{t}_a - \bar{t}_b)}}$$

converges to the standard normal distribution. Values of large absolute magnitude indicate a lack of convergence, though small values do not necessarily guarantee convergence. Note that the variance estimator must be chosen to account for the fact that the realisations $\theta^{(i)}$ are identically distributed but not independent. Geweke (1991) employs spectral time series analysis to estimate the sample's variance, in which case the test is also referred to as *Geweke's diagnostics*. Alternatively, the MCMC output can be sufficiently thinned to eliminate serial autocorrelation, and the test is applied to the resulting pseudo-independent sample. Then the variance is estimated via the sum of plain variance estimators for the pseudo-independent sample means $\bar{t}_a$ and $\bar{t}_b$. If several paths of Markov chains with different, preferably overdispersed, initialisations are available, convergence to a common equilibrium can be assessed via the estimated variances between and within paths. If convergence has not been reached, the individual paths are still influenced by their initial values. The ergodic means for the different paths show a wide spread between each other, and the sample variance of such ergodic means overestimates the true variance. Conversely, the average of ergodic variances within each path underestimates the true variance, because the Markov chains have not yet traversed their burn-in range to the support of the equilibrium. A numerical output is given by the proportional comparison of the two different estimates for the variance. It is far from one if there is disagreement between the various paths, but formal hypothesis tests are not available. For more information on these and further methods, see the review by Robert and Casella (2004).

C Vector Error Correction Models

In this appendix, the *VECM*, a well-known approach in multivariate time series analysis, is presented. The VECM is an extension of the *VAR model*, which itself is the vector-valued equivalent to standard AR approaches in time series analysis. While multiple time series can be modelled simultaneously in VAR models with due regard being given to serial cross-correlation of current and past multi-dimensional values, the VECM additionally accounts for long-term equilibriums *between* the single time series, i.e. there exists a linear combination of the different time series which is stationary, whereas the marginal time series may be non-stationary. This so-called *cointegration* can often be observed between different macroeconomic measures such as stock market indices or prices for commodities and financial products. Indeed, the VECM was developed in the context of financial econometrics, mainly by a series of pioneering papers, starting with the work by Granger (1981) and further exploration in the famous seminal paper by Engle and Granger (1987) with much contribution from several authors in the years thereafter. Most notably, Johansen (1988, 1991) and Johansen and Juselius (1990, 1992) developed an ML estimation framework, sometimes loosely referred to as the *Johansen procedure*, which is widely used nowadays. A thorough overview on the VECM can be found in Lütkepohl (2007) and Johansen (1995), for instance, which the following discussion is mainly based on. The reader is expected to bring general knowledge of common time series analysis as it can be found in many standard textbooks, for example in Box et al. (2013) or Brockwell and Davis (2009). Detailed coverage of general multivariate time series analysis is provided in the books by Lütkepohl (2007) and Reinsel (2003). The remainder of this appendix is organised as follows. VAR processes and the concept of cointegration are reviewed in Sections C.1 and C.2, respectively, before introducing the VECM in Section C.3. Sections C.4 and C.5 describe both frequentist and Bayesian estimation techniques, and Section C.6 summarises model diagnostics for the VECM.

C.1 The Vector Autoregressive Model

In the following, let $\{x_t, t = 1, \ldots, T\}$ be a multivariate time series of dimension $m \in \mathbb{N}$, i.e. $x_t \in \mathbb{R}^m$ for each $t = 1, \ldots, T$. The time series $\{x_t\}$ is an *m-dimensional VAR process of order $k \in \mathbb{N}$*, denoted as $x_t \sim \text{VAR}(k)$, if it is defined by the equations

$$x_t = \phi D_t + \sum_{i=1}^{k} A_i x_{t-i} + \varepsilon_t, \quad t = k+1, \ldots, T \tag{C.1}$$

with initial values $x_1, \ldots, x_k$ and m-dimensional *white noise* ε_t, i.e. $\text{E}(\varepsilon_t) = 0$, $\text{E}(\varepsilon_t \varepsilon_t') = \Omega$ for all $t = k+1, \ldots, T$, and $\text{E}(\varepsilon_t \varepsilon_s') = 0$ for all $s \neq t$, where $\Omega \in \mathbb{R}^{m \times m}$ is a positive definite covariance matrix. For this work, it will suffice to think of ε_t as *Gaussian white noise*, i.e. the special case of iid errors $\varepsilon_t \sim N_m(0, \Omega)$. A vector of possibly time-varying constants D_t with some fixed dimension $d \in \mathbb{N}$ enables the user to include deterministic, e.g. linear or seasonal, trends. This influence is measured by the according parameter matrix $\phi \in \mathbb{R}^{m \times d}$. Further parameters are the k AR coefficient matrices $A_i \in \mathbb{R}^{m \times m}$, which describe the impact of recent values in x_t on the current outcome. Adjustment of the so-called *order k* leads to different time horizons for the serial autocorrelation, i.e. x_t is conditionally independent of x_{t-k-1} and previous terms given the intermediate outcomes $x_{t-k}, \ldots, x_{t-1}$. Note that in the definition of the VAR model, it is implicitly assumed that k is the maximum value for which $A_k \neq 0$.

For a given set of parameters, the solution of the system in (C.1) can be stated in terms of the initial values of $\{x_t\}$ and the independent error terms $\varepsilon_t, t = k+1, \ldots, T$, see Theorem 2.1 in Johansen (1995). The *reverse characteristic polynomial* of this process is given by

$$A(z) = I_m - \sum_{i=1}^{k} A_i z^i, \quad z \in \mathbb{C},$$

where I_m is the $m \times m$ identity matrix. Denoting by B the *backshift operator*, defined through the operation $Bx_t = x_{t-1}$, the reverse characteristic polynomial obviously satisfies $A(B)x_t = \phi D_t + \varepsilon_t$ for the VAR(k) process defined in (C.1). Let $|A(z)|$ denote the determinant of $A(z)$. If D_t is bounded by a polynomial in t, and if $|A(z)| \neq 0$ for all z with $|z| \leq 1$, i.e. the VAR process does not have explosive or seasonal roots, then the

process $y_t := x_t - \mathrm{E}(x_t)$ is said to be *stable*, and, in this case, $\{x_t\}$ can be expressed as an MA process via

$$x_t = \sum_{i=0}^{\infty} C_i(\varepsilon_{t-i} + \phi D_{t-i}) = C_0(B)(\varepsilon_t + \phi D_t),$$

where the coefficient matrices C_i are given by $C_0(z) = \sum_{i=0}^{\infty} C_i z^i = A(z)^{-1}$ for $|z| < 1+\delta$ for some $\delta > 0$, and can be solved for recursively. Stability is the required property of any time series to enable statistical analyses, since if $\{y_t\}$ is stable, then it is also *stationary*[1] with zero mean, i.e. it holds that $\mathrm{E}(y_t) = 0$ and that $\mathrm{E}(y_t y_{t-h}') = \gamma(h)$ depends on $h \in \mathbb{N}_0$ only for all t. Often, stability and stationarity are used interchangeably, although a stationary process need not be stable, and, in the literature, the condition $|A(z)| \neq 0$ for $|z| \leq 1$ is not only referred to a stability but also as *stationarity condition*.

Note that the VAR(k) model is a special case of the more general family of VARMA models, which are also discussed by Lütkepohl (2007) as well as Box et al. (2013) and Brockwell and Davis (2009). In analogy to the univariate case, the representation of stationary multivariate time series through VARMA models is motivated by *Wold's Decomposition Theorem* due to Wold (1938). It states that any stationary process can be represented as the sum of two uncorrelated processes, one of which is a purely deterministic AR process and the other one has an MA representation, both having a possibly infinite order. In particular, if the deterministic part only consists of a mean term, Wold's Theorem states that the process has a pure MA representation. As a consequence, if the MA coefficients C_i are absolutely convergent, and the limit $C_0(z) = \sum_{i=0}^{\infty} C_i z^i$ is invertible for all z with $|z| \leq 1$ such that $C_0(z)^{-1}$ can be expressed as a convergent series $A(z) = I_m - \sum_{i=0}^{\infty} A_i z^i$, then $\{x_t\}$ is a VAR process of possibly infinite order with coefficient matrices A_i. The absolute convergence of $A(z)$ implies convergence of A_i to the zero matrix and therefore a reasonable approximation by a finite order VAR(k) model for sufficiently large $k \in \mathbb{N}$. This result is the core motivation behind VAR models, because it guarantees the usefulness of this approach for many eventually stationary time series after possible transformations to account for deterministic components other than the

[1] This work adopts the convention in time series analysis that stationarity refers to *weak* or *wide-sense stationarity* of time-invariant means and covariances. Other definitions, such as *strict stationarity* of time-invariant joint distributions of vectors consisting of consecutive variables of the time series, are not discussed here.

mean. Note that in the definition of the VAR model through (C.1), deterministic trends or seasonal patterns can be naturally included through ϕD_t. Lütkepohl (2007) reviews computation of the so-called *Yule-Walker equations* for the derivation of autocovariance functions and least-squares estimators for the model parameters, ML estimation under Gaussian white noise, determination of the lag length k, forecasting and checks of model adequacy, among many more detailed topics.

C.2 Cointegration

The previous section showed that the VAR model is appropriate for stationary time series with deterministic components, which can be modelled via ϕD_t. Obviously, many observed time series data will not meet such restrictive assumptions, for example time series with stochastic trends, heteroskedastic covariances, or periodic patterns with time-varying coefficients. Of particular interest in this work are time series models for stochastic trends, a feature observed in many real-life applications that can be accounted for by applying the already established methodology to transformed quantities of the underlying data. This section summarises approaches in the analysis of multivariate time series with stochastic trends. For the many other possible complications with observed time series, the interested reader is referred to the previously cited literature.

As with univariate time series, a time series is non-stationary when $|A(z)| = 0$ for some z with $|z| \leq 1$, i.e. the reverse characteristic polynomial has roots inside or on the edge of the unit disk. If for a univariate time series at least one root is strictly inside the unit disk, the time series is *explosive* in that its variance diverges with exponential rate. Although fruitful for analyses of, say, bacterial growth, in statistical time series applications, such models are usually considered unreasonable, and attention is devoted to the borderline case, when $|A(z)| = 0$ for $|z| \leq 1$ implies $|z| = 1$. If, in particular, the only root on the unit circle is the so-called *unit root* $z = 1$ and all other roots still fulfil $|z| > 1$, the behaviour is similar to that of a *random walk*, which is exactly the AR(1) model with its only root being the unit root. In contrast to explosive time series, such processes are characterised by linearly increasing variances and asymptotic correlation of one for large lags. Moreover, trends are linear if there is no superimposed deterministic trend. Although of interest in applications with periodic data, the case of roots on the

unit circle other than the unit root $z = 1$ is not discussed here.

For the moment, let $\{x_t, t = 1, \ldots, T\}$ be a univariate AR process without deterministic components, which has $d \in \mathbb{N}$ unit roots with all other roots being outside the unit disk. Note that then $A(z) = \alpha(z)(1 - z)^d$, where $\alpha(z)$ is the characteristic polynomial of a stable process, because all its roots are now outside the unit disk. Hence, the process $\{x_t\}$, given through $A(B)x_t = \varepsilon_t$, can be written as

$$\alpha(B)(1 - B)^d x_t = \varepsilon_t,$$

and it can be seen that the process $y_t := \Delta^d x_t := (1 - B)^d x_t$, which is the original time series differenced d times, is stationary. Noting that $x_t = ((1 - B)^{-1})^d y_t$ with $(1 - B)^{-1} = \sum_{i=0}^{\infty} B^i$, the original time series x_t is obtained by summing – or, in other words, integrating – d times the stationary time series y_t. Hence, the process is called *integrated*[2] *of order d*, commonly denoted as $x_t \sim I(d)$. For consistency, a stationary time series is then sometimes denoted as $x_t \sim I(0)$.

Now consider the general case of an m-dimensional VAR(k) process $\{x_t\}$ with $A(z) = I_m - \sum_{i=1}^{k} A_i z^i$ and without deterministic components, i.e.

$$A(B)x_t = \varepsilon_t.$$

The *adjugate matrix* of $A(B)$ is the matrix $\operatorname{adj}(A(B))$ for which $A(B)\operatorname{adj}(A(B)) = |A(B)|I_m$. Multiplying by $\operatorname{adj}(A(B))$ from the left gives the alternative representation

$$|A(B)|x_t = \operatorname{adj}(A(B))\varepsilon_t. \tag{C.2}$$

Since it can be shown that $|A(z)|$ is a characteristic polynomial, $|A(B)|$ is a univariate polynomial in the backshift operator, i.e. the left-hand side of (C.2) is an m-dimensional vector of univariate AR processes with identical AR operators. By definition of the adjugate matrix, the right-hand side of (C.2) is an m-dimensional MA process of finite order.As in the univariate case, the AR operator $|A(B)|$ is tested for unit roots,

[2]Note that in the general setting of Autoregressive Moving Average (ARMA) models, *any* process is called *integrated of order d* if it can be characterised as a stationary and invertible ARMA model after differencing d times.

where w.l.o.g. the case of roots with $|z| \leq 1$ but $z \neq 1$ is neglected. Again, if the AR operator consists of d unit roots with all other roots being outside the unit disk, then $|A(B)| = \alpha(B)(1-B)^d$ for some polynomial α with $\alpha(z) \neq 0$ for $|z| \leq 1$, and the vector $\Delta^d x_t$ of d times marginally differenced time series is an m-dimensional stationary process.

Lütkepohl (2007) points out that d is an upper bound for the integration order of each marginal time series. In particular, all components may be integrated with orders strictly less than d, and such orders may also vary. Moreover, if one allows $\{x_t\}$ to be VAR(k) *with* deterministic trend vector ϕD_t, the component-wise differencing technique can lead to a stable representation $\alpha(B)\Delta^d x_t = \mathrm{adj}(A(B))\varepsilon_t$ without deterministic trends in the marginals, i.e. the underlying deterministic relations between the marginal time series in x_t are cancelled out. See the cited reference for simple examples. As a consequence of the latter finding, considering marginal time series as $I(d)$ processes and differencing them individually can distort the structure of the multivariate time series, and therefore leads to loss of information on the relationship between the components. As Box et al. (2013) point out, dealing with non-stationarity becomes substantially different and more complicated in the multivariate case when the marginal time series share *common stochastic trends*. In this situation, the individual time series $x_{1t}, \ldots, x_{mt}$ in $x_t = (x_{1t}, \ldots, x_{mt})'$ are integrated of possibly different orders, but wander jointly by satisfying a linear combination $\beta' x_t = \sum_{i=1}^m \beta_i x_{it}$, which itself is stationary with zero mean. If the maximum order of integration among the marginals is d, and the process $z_t := \beta' x_t$ fulfils $\beta \neq 0$ and $z_t \sim I(d-b)$, then the multivariate process $\{x_t\}$ of integrated univariate time series is called *cointegrated of order* (d, b), written as $x_t \sim C(d, b)$, with *cointegrating vector* β. Cointegration can be interpreted as a deterministic *long-run equilibrium* $\beta' x_t = 0$, superimposed by stochastic but stable deviations through a univariate stationary process. Such noisy equilibriums are often observed between macroeconomic quantities. With the foregoing definition, one can think of x_t being of integration order d itself, since $\Delta^d x_t$ is stable but $\Delta^{d-1} x_t$ is not, although the order of integration for certain marginals may be less than d. This clearly simplifies terminology, but must be understood carefully when analysing and interpreting the individual components. Moreover, note that the cointegration vector β is not unique. Not only is $c\beta$ a cointegrating vector for any constant $c \neq 0$, but also several linearly independent cointegrating vectors $\beta_1, \ldots, \beta_r$ may exist for a given cointegrated process.

C.3 Vector Error Correction Models

In the following, the special case of $x_t \sim C(1,1)$ is considered, i.e. the marginals are $x_{it} \sim I(0)$ or $x_{it} \sim I(1)$ for all $i = 1, \ldots, m$, with the latter holding for at least one i, and there exists at least one non-trivial linear combination $z_t \sim I(0)$ of $x_{it}, i = 1, \ldots, m$. If $x_t \sim \text{VAR}(k)$ without deterministic components, then

$$x_t = \sum_{i=1}^{k} A_i x_{t-i} + \varepsilon_t, \quad t = k+1, \ldots, T,$$

and $|A(z)|$ has, say, d unit roots, where all other roots lie outside the unit disk. In particular, it follows that $|A(1)| = |I_m - \sum_{i=1}^{k} A_i| = 0$, and so the m-dimensional square matrix

$$\Pi := -\left(I_m - \sum_{i=1}^{k} A_i\right)$$

must be singular. Let $r := \text{rk}(\Pi)$ be the rank of Π. Since Π is singular, it holds that $r < m$. Also, suppose w.l.o.g. that $r > 0$, because for $r = 0$ the terms including Π would simply vanish in the following discussion. Then there exists a decomposition $\Pi = \alpha\beta'$ with α and β both being non-zero $m \times r$ matrices of rank r. Differencing x_t once yields

$$\begin{aligned}
\Delta x_t &= x_t - x_{t-1} \\
&= \sum_{i=1}^{k} A_i x_{t-i} + \varepsilon_t - x_{t-1} \\
&= -I_m x_{t-1} + A_1 x_{t-1} + \sum_{i=2}^{k} A_i x_{t-i} + \varepsilon_t \\
&= -I_m x_{t-1} + A_1 x_{t-1} + A_2 x_{t-1} + \cdots + A_k x_{t-1} + \sum_{i=2}^{k} A_i (x_{t-i} - x_{t-1}) + \varepsilon_t \\
&= \Pi x_{t-1} - A_2 \Delta x_{t-1} + \sum_{i=3}^{k} A_i (x_{t-i} - x_{t-2} - \Delta x_{t-1}) + \varepsilon_t \\
&= \Pi x_{t-1} - \sum_{i=2}^{k} A_i \Delta x_{t-1} + \sum_{i=3}^{k} A_i (x_{t-i} - x_{t-2}) + \varepsilon_t
\end{aligned}$$

and hence, by induction,

$$\begin{aligned}\Delta x_t &= \Pi x_{t-1} - \sum_{i=2}^{k} A_i \Delta x_{t-1} - \sum_{i=3}^{k} A_i \Delta x_{t-2} - \cdots - A_k \Delta x_{t-k+1} + \varepsilon_t \\ &= \alpha\beta' x_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \varepsilon_t \end{aligned} \tag{C.3}$$

with $\Gamma_i := \sum_{j=i+1}^{k} A_j$ for $i = 1, \ldots, k-1$. Since, by assumption, the differenced time series $\{\Delta x_{t-k+1}, \ldots, \Delta x_{t-1}, \Delta x_t\}$ and ε_t are stationary processes, rearranging of (C.3) reveals that the term $\alpha\beta' x_{t-1}$ as a linear combination of stationary processes is stationary itself. Multiplication of $\alpha\beta' x_{t-1}$ by a matrix from the left does not distort stationarity, and the special case of multiplication by $(\alpha'\alpha)^{-1}\alpha'$ reveals that even $\beta' x_{t-1}$ is a vector of r different stationary processes, which hence must be the cointegrating long-run relationships. Therefore, β is called *cointegration matrix* and consists of the r cointegrating vectors, given by its columns. The *loading matrix* α then consists of m rows, each measuring the effect of the long-run equilibriums on the corresponding time series through a weighted sum. The rank r of Π can be interpreted as the number of linearly independent equilibrium relationships between the univariate marginals, and is often called the *cointegration rank*. Since $\Pi = \alpha\beta' = \alpha Q Q^{-1}\beta' = \tilde{\alpha}\tilde{\beta}'$ with $\tilde{\alpha} := \alpha Q$ and $\tilde{\beta} := \beta(Q^{-1})'$ for every regular matrix $Q \in \mathbb{R}^{r\times r}$, the decomposition of Π is not unique, which corresponds to the non-uniqueness of representations for the individual cointegrating equations. Identifiability can be obtained by imposing restrictions on α or β, and most commonly one sets $\beta = (I_r, \beta_l')'$ for some lower block matrix $\beta_l \in \mathbb{R}^{(m-r)\times r}$. This constraint is often referred to as *linear normalisation*.

The VAR(k) model for $\{x_t\}$ written as in (C.3) is referred to as the *VECM representation*. More precisely, this model equation is called the *transitory version* of the VECM, compared to the equivalent *long-run specification*

$$\Delta x_t = \sum_{i=1}^{k-1} \tilde{\Gamma}_i \Delta x_{t-i} + \alpha\beta' x_{t-k} + \varepsilon_t, \quad t = k+1, \ldots, T$$

with coefficients $\tilde{\Gamma}_i := -(I_m - \sum_{j=1}^{i} A_j)$ for $i = 1, \ldots, k-1$. This work adopts the transitory specification of the VECM, because Π can be conveniently interpreted as the

error correction effect on the previous observation rather than the outcome at a possibly large lag k. Again, it is worth emphasising that although the maximum integrating order of each univariate time series is at most 1, a VAR$(k-1)$ representation for the first differences in $\{x_t\}$ would eliminate the cointegration term $\alpha\beta' x_{t-1}$ and, hence, would not contain the full information of the VAR(k) process $\{x_t\}$ as derived in (C.3). In particular, the derivation above reveals that a cointegrated process does not yield a VAR representation for the first differences of the original time series. Starting from a VECM representation, however, is a valuable approach, as one obtains a stationary VAR$(k-1)$ process for the first differences if $r = 0$ or, equivalently, $\Pi = 0$, and a stationary VAR(k) process for the integrated time series if $r = m$, because a full rank m implies that Π is regular such that $|A(1)| \neq 0$, which means that $\{x_t\}$ has no unit roots.

As a final step, the VECM presentation in (C.3) can be extended by an additional term ϕD_t for deterministic trends. Since D_t is time-dependent, it can be easily used to incorporate constant means or any linear, quadratic, or higher order trend. The resulting VECM

$$\Delta x_t = \phi D_t + \alpha\beta' x_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \varepsilon_t, \quad t = k+1, \ldots, T \tag{C.4}$$

is *unrestrictive* in the deterministic component. Lütkepohl (2007) shows that if each ϕD_t can be decomposed into a sum with one addend being of the form $-\alpha\beta'\mu_0$, then this latter term can be absorbed as intercepts into the cointegration relations such that the process $\{\beta' x_t\}$ has constant mean $\mu_0 \in \mathbb{R}^m$. Similarly, parts of a linear trend in $\{x_t\}$ can be absorbed into an expanded error correction term, representing linear trends in the cointegrating relationships, whereas remaining time-invariant addends of ϕD_t would generate linear trends in the marginal time series. In this work, the focus is limited to the unrestrictive specification without exploring the nature of deterministic components in more detail. The reader is referred to Johansen (1995) for a thorough discussion.

By analogy to the MA representation of any VAR(k) process, under minimum conditions on the parameters and initial values in the VECM, the process $\{x_t\}$ can be represented as a function in the white noise variables $\varepsilon_{k+1}, \ldots, \varepsilon_t$ and the initial values $x_1, \ldots, x_k$. More specifically, x_t is decomposed into $m - r$ stochastic trends, represented by a weighted

sum of m random walks that determine the long-run behaviour of x_t, an $I(0)$ process denoting the disequilibrium error, as well as terms for the initial values and deterministic components. This result is known as the *Granger Representation Theorem*[3], see Theorem 4.2 in Johansen (1995), which is of particular importance in deriving asymptotic properties of parameter estimators.

Based on the findings with regard to Wold's Theorem, this work focuses on the VECM as the direct extension of the generally applicable VAR model through the error correction term. Note, however, that the broader class of VARMA models contains stochastic processes, which cannot be represented through pure VAR equations. If such processes are not stationary, they may be modelled as integrated time series and, again, possible cointegration for long-run relationships between the marginal time series must be taken into account. The generalisation of the VECM to cointegrated VARMA models is discussed by Lütkepohl (2007) and the references therein, but not further considered here. Note also that the concept of cointegration, which was introduced in multivariate time series through the work by Granger (1981) and Engle and Granger (1987), is closely connected to the idea of even earlier developed *error correction models*, which were designed to overcome spurious correlation in hitherto used linear regression approaches. Besides the above derivation of the VECM, Lütkepohl (2007) motivates this model by extension of error correction models by time series concepts. A general survey on error correction models can be found in, e.g., Salmon (1982).

C.4 Frequentist Estimation and Forecasting

Estimation of the VECM becomes more complicated than for a VAR specification, because in addition to the order k, the cointegration rank r is generally unknown and must be estimated. As a result, asymptotic behaviour for the estimators is different from what is known for the stationary VAR(k) model. The starting point for the general estimation

[3] As Hansen (2005) points out, the famous *Granger Representation Theorem* in this form is due to Johansen (1991), but should not be confused with the same-titled theorem proven by Engle and Granger (1987), which makes statements about the existence of an error correction representation for a process with stationary and invertible VARMA specifications for its first difference and cointegrating relationships.

of a VECM specification is the simplified version

$$\Delta x_t = \alpha\beta' x_{t-1} + \varepsilon_t, \quad t = 2, \dots, T$$

with initial value x_1, standard white noise for the error terms, no deterministic components, and a lag order of $k = 1$ for the VAR model, i.e. the innovation process does not have any AR components. Unknown parameters that have to be estimated, are the cointegration rank r and the matrices α, β, as well as the variance and covariance terms for the error terms. The estimation procedure will be generalised later for an arbitrary lag k.

For the following estimators, it is assumed that a sample of time series data $x_2, \dots, x_T$ and the initial value x_1 are observed. Given that the cointegration rank r and the covariance matrix Ω for the white noise are known with $r \neq 0$, the *unrestricted least-squares estimator* is given by

$$\widehat{\Pi} := \left(\sum_{t=2}^{T} \Delta x_t x_{t-1}'\right)\left(\sum_{t=2}^{T} x_{t-1}x_{t-1}'\right)^{-1}, \tag{C.5}$$

and Lütkepohl (2007) shows that this is an unbiased and asymptotically normal estimator. In particular, denoting by vec and $\otimes$ the vectorisation operator and Matrix Kronecker product from Definitions D.2 and D.3, respectively, it holds that

$$\sqrt{T-1}\,\mathrm{vec}\left(\widehat{\Pi} - \Pi\right) \xrightarrow{d} N_{m^2}(0, \beta\,\mathrm{Cov}(z_t, z_{t-1})^{-1}\beta' \otimes \Omega),$$

where $z_t := (\beta, \alpha_\perp)' x_t$ for some $t \in \{2, \dots, T\}$ with an orthogonal complement $\alpha_\perp$ of α, i.e. an $m \times m - r$ matrix of full column rank with $\alpha'\alpha_\perp = 0$. The matrix $\beta\,\mathrm{Cov}(z_t, z_{t-1})^{-1}\beta'$ can be consistently estimated via

$$\left((T-1)^{-1}\sum_{t=2}^{T} x_{t-1}x_{t-1}'\right)^{-1}.$$

When Ω is unknown, the usual residual covariance matrix yields a consistent estimator for Ω and allows for t-tests on individual entries in $\widehat{\Pi}$. However, more general Wald tests with more flexible restrictions suffer from singularities in the limiting distribution, and must be treated differently, see Lütkepohl (2007).

With the linear parametrisation $\beta = (I_r, \beta_l')'$ for some $\beta_l \in \mathbb{R}^{(m-r)\times r}$, the estimators for α and β are derived in a two-step procedure. Noting that under this normalisation the first r columns of $\Pi = \alpha\beta'$ are equal to α, a consistent estimator $\widehat{\alpha}$ for this matrix is given by the corresponding columns in $\widehat{\Pi}$ as in (C.5). Then the *estimated generalised least-squares estimator* for β_l is defined through

$$\widehat{\beta}_l' := \left(\widehat{\alpha}'\widehat{\Omega}^{-1}\widehat{\alpha}\right)^{-1}\widehat{\alpha}'\widehat{\Omega}^{-1}\left(\sum_{t=2}^{T}\left(\Delta x_t - \widehat{\alpha}x_{t-1}^{(1)}\right)x_{t-1}^{(2)\prime}\right)\left(\sum_{t=2}^{T}x_{t-1}^{(2)}x_{t-1}^{(2)\prime}\right)^{-1},$$

where $x_t = \left(x_t^{(1)\prime}, x_t^{(2)\prime}\right)'$ with r-dimensional $x_t^{(1)}$ and $(m-r)$-dimensional $x_t^{(2)}$, and $\widehat{\Omega}$ is the consistent residual covariance matrix estimator for Ω. Asymptotic normality can be established for this estimator as well, and it brings the advantage of asymptotic χ^2 distributions for Wald tests.

Under the additional assumption of Gaussian white noise, the normal distribution on ε_t can be used to establish ML estimation. The maximisation problem of the (log) likelihood is equivalent to a determinant minimisation problem, which can be solved by an eigen-decomposition of the matrix

$$\left(\sum_{t=2}^{T}x_{t-1}x_{t-1}'\right)^{-1/2}\left(\sum_{t=2}^{T}x_{t-1}\Delta x_t'\right)\left(\sum_{t=2}^{T}\Delta x_t\Delta x_t'\right)\left(\sum_{t=2}^{T}\Delta x_t x_{t-1}'\right)\left(\sum_{t=2}^{T}x_{t-1}x_{t-1}'\right)^{-1/2}.$$

Denoting the orthonormal eigenvectors of this matrix by $v_1, \ldots, v_m$, where the indices correspond to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$, the *ML estimator* is $\widetilde{\Pi} := \widetilde{\alpha}\widetilde{\beta}'$ with

$$\widetilde{\beta} := (v_1, \ldots, v_r)'\left(\sum_{t=2}^{T}x_{t-1}x_{t-1}'\right)^{-1/2},$$
$$\widetilde{\alpha} := \left(\sum_{t=2}^{T}\Delta x_t x_{t-1}'\widetilde{\beta}\right)\left(\sum_{t=2}^{T}\widetilde{\beta}'x_{t-1}x_{t-1}'\widetilde{\beta}\right)^{-1}.$$

The ML estimator has the same asymptotic properties as the unrestricted least-squares estimator $\widehat{\Pi}$. Estimators for the normal linearisation are obtained by multiplication of $\widetilde{\beta}$ with the inverse of its own upper $r \times r$ block matrix, which yields $\left(I_r, \widetilde{\beta}_l'\right)'$, and by substitution of $\widetilde{\beta}$ in the second formula to obtain the corresponding estimator for α.

Again, asymptotic properties correspond to those found for the estimated generalised least-squares estimator.

The previously mentioned residual covariance matrix is estimated as

$$\widehat{\Omega} := (T-1)^{-1} \sum_{t=2}^{T} \left(\Delta x_t - \widehat{\Pi} x_{t-1}\right) \left(\Delta x_t - \widehat{\Pi} x_{t-1}\right)',$$

where $\widehat{\Pi}$ can be any of the above estimators for Π. Simple versions of the above estimators for the special case of $r = 0$ are specifically derived by Lütkepohl (2007).

For the general case of k AR lags, i.e. for the VECM of the form

$$\Delta x_t = \alpha\beta' x_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \varepsilon_t, \quad t = k+1, \ldots, T$$

with initial values $x_1, \ldots, x_k$, each of the above estimators can be generalised. With the compact matrix formulation already known from Section 3.2.2, the model can be written as

$$\Delta X = \alpha\beta' X_1 + \Gamma X_2 + \varepsilon$$

with $m \times T-k$ matrices $\Delta X = (\Delta x_{k+1}, \ldots, \Delta x_T)$, $X_1 = (x_k, \ldots, x_{T-1})$, $\varepsilon = (\varepsilon_{k+1}, \ldots, \varepsilon_T)$, an $m \times (k-1)m$ matrix $\Gamma = (\Gamma_1, \ldots, \Gamma_{k-1})$, and a $(k-1)m \times T-k$ matrix

$$X_2 = \left((\Delta x_k', \ldots, \Delta x_2')', (\Delta x_{k+1}', \ldots, \Delta x_3')', \ldots, (\Delta x_{T-1}', \ldots, \Delta x_{T-k+1}')'\right).$$

The consistent and asymptotically normal *unrestricted least-squares estimator* can be shown to be

$$\left(\widehat{\Pi}, \widehat{\Gamma}\right) := (\Delta X X_1', \Delta X X_2') \begin{pmatrix} X_1 X_1' & X_1 X_2' \\ X_2 X_1' & X_2 X_2' \end{pmatrix}^{-1}$$

with consistent estimator for the covariance matrix

$$\widehat{\Omega} := (T - (m+1)k)^{-1} \left(\Delta X - \widehat{\Pi} X_1 - \widehat{\Gamma} X_2\right) \left(\Delta X - \widehat{\Pi} X_1 - \widehat{\Gamma} X_2\right)'.$$

Remarks w.r.t. t and Wald tests made for the unrestricted least-squares estimator in the case without AR components apply here, too. Now, let

$$\begin{aligned} M &= I_{T-k} - X_2'(X_2X_2')^{-1}X_2, \\ R_0 &= \Delta X M, \\ R_1 &= X_1 M, \end{aligned}$$

and split $R_1 = (R_1^{(1)\prime}, R_1^{(2)\prime})'$ into block matrices with r and $m-r$ rows, respectively. By analogy to the previous case, under linear normalisation the *estimated generalised least-squares estimator* for β_l is

$$\widehat{\beta}_l' := \left(\widehat{\alpha}'\widehat{\Omega}^{-1}\widehat{\alpha}\right)^{-1}\widehat{\alpha}'\widehat{\Omega}^{-1}\left(R_0 - \widehat{\alpha}R_1^{(1)}\right)R_1^{(2)\prime}\left(R_1^{(2)}R_1^{(2)\prime}\right)^{-1},$$

when $\widehat{\Pi}, \widehat{\Gamma}, \widehat{\Omega}$ are given through the unrestricted least-squares estimation above and $\widehat{\alpha}$ equals the first r rows of $\widehat{\Pi}$. Its asymptotic behaviour corresponds to the simple case where $k = 1$.

Under iid $\varepsilon_t \sim N_m(0, \Omega)$, the previous ML estimator can be generalised for an arbitrary lag order k. With above definitions of M, R_0, R_1 and the matrices $S_{ij} = R_i R_j'/(T-k)$ for $i, j = 0, 1$, (log) likelihood maximisation can again be achieved through an eigen-decomposition. Let $\lambda_1 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of the matrix $S_{11}^{-1/2}S_{10}S_{00}^{-1}S_{01}S_{11}^{-1/2}$ with corresponding orthonormal eigenvectors $v_1, \ldots, v_m$. The *ML estimator* is then given by $\widetilde{\Pi} := \widetilde{\alpha}\widetilde{\beta}'$ with

$$\begin{aligned} \widetilde{\beta}' &:= (v_1, \ldots, v_r)'S_{11}^{-1/2}, \\ \widetilde{\alpha} &:= S_{01}\widetilde{\beta}\left(\widetilde{\beta}'S_{11}\widetilde{\beta}\right)^{-1}. \end{aligned}$$

The corresponding estimators for Γ and Ω are

$$\begin{aligned} \widetilde{\Gamma} &:= \left(\Delta X - \widetilde{\Pi}X_1\right)X_2'\left(X_2X_2'\right)^{-1}, \\ \widetilde{\Omega} &:= (T-k)^{-1}\left(\Delta X - \widetilde{\Pi}X_1 - \widetilde{\Gamma}X_2\right)\left(\Delta X - \widetilde{\Pi}X_1 - \widetilde{\Gamma}X_2\right)'. \end{aligned}$$

All estimators are consistent and asymptotically normal with the same limiting distri-

bution as the least-squares estimator, and $\widetilde{\Omega}$ is asymptotically independent of all other parameter estimators. For an identifiable decomposition of $\widetilde{\Pi}$, the same technique as for the case of $k = 1$ leads to the ML estimator of β_l and corresponding α under the linear normalisation. Again, the asymptotic distribution coincides with that of the estimated generalised least-squares estimator. As before, special formulae for the case of $r = 0$ are again derived in Lütkepohl (2007).

Generalisation of the already introduced estimators through inclusion of deterministic components is straightforward. Consider the unrestrictive specification of the VECM in (C.4) with deterministic component D_t and corresponding parameters ϕ. The compact representation of the VECM introduced in this section remains valid if Γ and X_2 are replaced by

$$\begin{aligned}\Gamma &= (\Gamma_1, \ldots, \Gamma_{k-1}, \phi)\,, \\ X_2 &= \left(\left(\Delta x'_k, \ldots, \Delta x'_2, D'_{k+1}\right)', \ldots, \left(\Delta x'_{T-1}, \ldots, \Delta x'_{T-k+1}, D'_T\right)' \right).\end{aligned}$$

Lütkepohl (2007) shows that with above notation, both least-squares and ML estimators are given by their previous formulae. Asymptotic behaviour remains principally the same, too.

Once all parameters, including the covariance matrix in the VECM representation, are estimated, it is possible to return to the VAR formulation in (C.1) for prediction purposes. For the AR parameters $A_1, \ldots, A_k$ in levels, it holds that $A_1 = I_m + \Pi$ if $k = 1$, and

$$\begin{aligned}A_1 &= I_m + \Pi + \Gamma_1, \\ A_i &= \Gamma_i - \Gamma_{i-1}, \quad i = 2, \ldots, k-1, \\ A_k &= -\Gamma_{k-1}\end{aligned}$$

otherwise. Being linear transformations of Γ and Π, consistent and asymptotic normal estimators $\widehat{A}_1, \ldots, \widehat{A}_k$ are obtained by replacing Γ and Π by their least-squares or ML estimators. The estimators $\widehat{\Omega}$ and, if applicable, $\widehat{\phi}$ can be directly carried forward. Consequently, for $h \in \mathbb{N}$, the *h-step forecasts* for the time series in levels are then

recursively given through

$$\widehat{x}_{T+h} = \widehat{\phi} D_{T+h} + \sum_{i=1}^{k} \widehat{A}_i \widehat{x}_{T+h-i},$$

where $\widehat{x}_t = x_t$ for all $t \leq T$. Under the assumption of future iid Gaussian white noise with covariance matrix Ω, the *forecast errors* $x_{T+h} - \widehat{x}_{T+h}$ are multivariate normal with zero mean and a covariance matrix that is estimated by

$$\widehat{\Omega}_h := \sum_{i=0}^{h-1} \widehat{C}_i \widehat{\Omega} \widehat{C}_i',$$

where $\widehat{C}_i := \sum_{j=1}^{i} \widehat{C}_{i-j} \widehat{A}_j$ for $i = 1, \ldots, h-1$ can be solved for recursively with starting value $\widehat{C}_0 = I_m$ and $\widehat{A}_j = 0$ for $j > k$. With this result, confidence intervals for individual components or confidence regions for the multivariate h-step forecast become computable. For simultaneous confidence intervals for several components, inflation of type I errors must be dealt with via the, say, Bonferroni method.

All above approaches rely on the assumption that both the lag order and cointegration rank are known, which will be unreasonable in most applications. As it is the case for the lag order k in any AR model, the rank, too, has to be determined in a prior step, and estimation is conducted as described before based on this finding. Since a VECM with $k-1$ lags can be written as an equivalent VAR(k) model, lag order selection is adopted from the general VAR framework and applied with the latter representation. For the VECM representation, the number of lags is reduced by one such that the VAR(k) process need not be analysed for $k = 0$. As known from univariate AR processes, the value of k may be estimated through different approaches. A common strategy involves consecutive Wald or likelihood-ratio tests. Starting with a predetermined upper bound K for k, by decreasing the possible lag order k by 1, one tests $A_k = 0$ versus $A_k \neq 0 \mid A_K = \cdots = A_{k+1} = 0$ until no rejection occurs. Due to the significant increase of type I errors in multiple testing schemes with the complicated determination of a correct nominal significance level, a popular alternative is so-called information criteria. The *final prediction error criterion* by Akaike (1969) is recommended if forecasting is the main purpose, as it chooses the lag order such that the mean prediction error is minimised. The *Akaike information criterion*, also due to Akaike (1974), the *Bayesian* or *Schwarz*

information criterion due to Schwarz (1978), as well as the *Hannan-Quinn information criterion* by Hannan and Quinn (1979) are prominent alternatives. Lütkepohl (2007) defines each of these criteria and discusses pro and cons in terms of different sample sizes.

Since the lag order selection is done with aid of the VAR representation, it can be conveniently conducted without knowledge of the number of cointegration relationships, whence it is typical to derive the lag order first and the cointegration rank subsequently. Although information criteria for rank determination can be found in the literature, here it is common to apply consecutive likelihood-ratio tests under the assumption of Gaussian white noise. As mentioned earlier, the maximised (log) likelihoods are obtained by minimisation of a determinant, which for a VECM without deterministic components is the product of the previously defined eigenvalues $\lambda_1, \ldots, \lambda_m$. For a test of $H_0\colon\ \mathrm{rk}(\Pi) = r_0$ against $H_1\colon r_0 < \mathrm{rk}(\Pi) \leq r_1$ with pre-specified integers $r_0 < r_1$, the *likelihood-ratio test statistic* becomes

$$\lambda_{LR}(r_0, r_1) := -(T-k) \sum_{i=r_0+1}^{r_1} \log(1-\lambda_i).$$

In situations with more complex VECM representations, e.g. with linear trends, the test statistic remains the same but with eigenvalues of a correspondingly generalised matrix given in Lütkepohl (2007). Two particular choices for r_1 w.r.t. r_0 are of main interest: with $r_1 = m$, the null value r_0 is tested against any other possible rank exceeding r_0, called the *trace test*, and with $r_1 = r_0 + 1$, one obtains the *maximum eigenvalue test* in which the null value is only tested against the next higher value. Tabulated percentage points for the asymptotic non-standard distributions are given by Johansen (1995), for instance. The strategy for rank determination, often explicitly referred to as the *Johansen procedure*, is to start with $r_0 = 0$ and to test it against higher alternatives. The cointegration rank of r_0 is accepted if H_0 cannot be rejected, otherwise the test is carried forward to $r_0 = 1$. The proposal is increased until the test cannot be rejected for the first time, and the according value for r_0 determines r. If the resulting cointegration rank turns out to be either 0 or m, one concludes that a VAR model in first differences or a stationary VAR model for the original variables is appropriate, respectively. The strategy allows for both the trace or maximum eigenvalue tests, and power analyses show that none of these tests is generally preferable over the other. It is noteworthy that,

despite the fact that the Johansen procedure is still the most common practice in frequentist analysis, some authors have criticised inconsistency problems in the estimation of r and proposed alternatives, see, e.g., Chao and Phillips (1999).

C.5 Bayesian Estimation and Forecasting

Following the arguments in Appendix A, Bayesian estimation of VAR(k) models in general and, in particular, in VECM representation has evolved an interesting alternative to least-squares and ML approaches as described above. As before, for parameter estimation, the lag order k is generally assumed known or determined in advance. In a fully Bayesian framework, this may be done as outlined by Villani (2001). However, whereas pioneering work in the Bayesian estimation of VECM also requires a fixed value for the cointegration rank, as it is the case in all frequentist techniques, modern approaches incorporate the determination of r in the general parameter estimation. Their outcomes can be regarded highly advantageous since incoherence in this step is diminished. A short overview on different Bayesian estimation procedures is given in the remainder of this section.

Starting with a general VAR(k) model as in (C.1) but without deterministic components, Bayesian estimation of $A_1, \ldots, A_k$ is often done via a multivariate normal prior for the stacked vector $\mathrm{vec}(A_1, \ldots, A_k)$ of all columns of all individual AR parameter matrices with pre-determined mean $\mu_A \in \mathbb{R}^{km^2}$ and covariance matrix $V_A \in \mathbb{R}^{km^2 \times km^2}$. The likelihood is chosen based on Gaussian error terms with zero mean and covariance matrix Ω that is assumed known for the moment. Lütkepohl (2007) derives the posterior distribution for $\mathrm{vec}(A_1, \ldots, A_k)$, which is again multivariate normal. The exact form of the posterior depends on the choice of the hyperparameters μ_A and V_A. Most commonly, the mean value μ_A is set to zero to reflect weak belief that AR correlation is significant. The covariance matrix V_A is typically chosen to be a diagonal matrix in order to reflect prior independence between all entries within and between the matrices $A_1, \ldots, A_k$. Using a suitable version for stable processes of the so-called *Minnesota* or *Litterman prior* by Doan et al. (1984) and Litterman (1986), the diagonal elements of

V_A are given by

$$v_{ijl} = \begin{cases} \left(\frac{\lambda}{l}\right)^2, & i = j \\ \left(\frac{\lambda\theta\sigma_i}{l\sigma_j}\right)^2, & i \neq j \end{cases}$$

for $i, j = 1, \ldots, m$ and $l = 1, \ldots, k$, where v_{ijl} is the diagonal element in V_A referring to the entry (i, j) in A_l. The hyperparameters become the prior standard deviation $\lambda > 0$ for all diagonal entries in A_1 and a tuning parameter $\theta \in (0, 1)$. The entries $\sigma_i, i = 1, \ldots, m$, represent the square roots of the diagonal elements of Ω, still assumed to be known for the moment. The hyperparameter λ determines the general uncertainty in the parameters in $\text{vec}(A_1, \ldots, A_k)$ and the tuning parameter θ represents the reduction of uncertainty for the off-diagonal elements, as they are believed to be closer to 0. Also, the variance and covariance terms for the entries in $A_1, \ldots, A_k$ decrease with increasing lags, because AR terms for higher orders are believed to be less significant than for smaller orders. The ratio σ_i^2/σ_j^2 is a normalisation factor to take into account the differences in residual variability. As a final note, for the Bayesian estimation in a proper VAR(k) model, the error variance Ω is generally not known and, in strict Bayesian philosophy, priors for its elements must be included in the analysis. However, since this would highly complicate the computation of the posterior distribution for all variables, usually an *empirical Bayes approach* is entertained in which Ω is replaced by its least-squared or ML estimator. Also, priors may be adjusted to allow for deterministic trends in (C.1).

The Bayesian methodology becomes far more complicated with regards to models in VECM form. Not only is the cointegration rank to be additionally estimated, but also the determination of $\Pi = \alpha\beta'$ reveals a non-linear estimation task with possible identification problems. Due to these restrictions, Lütkepohl (2007) points out that many modellers circumnavigate these challenges by setting up the model in VAR(k) form and estimating its parameters via the, e.g., original Minnesota prior as described by Doan et al. (1984) and Litterman (1986), thereby ignoring the fact that the individual time series may be cointegrated. In contrast to the previously discussed modification of the Minnesota prior, the original version for possibly integrated multivariate AR processes is still multivariate normal with zero mean for almost all parameters, except for the means of the diagonal elements for the first lag coefficient matrix. These are set equal

to one to express the prior belief of having m individually integrated time series, i.e. m independent random walks. The variance for the normal prior is still chosen as outlined above and allows for uncertainty about the simple assumption of independent random walks. Alternatives to the Minnesota prior are found in the literature of the early 1990s, for example see DeJong (1992), who applies a non-informative prior for $A_1, \ldots, A_k$ in the VAR(k) representation. Since these models, however, do not include the concept of cointegration and bear the risk of undesired changes in the prior when transforming the VAR(k) model into the VECM representation for cointegration analysis, the specification of priors for VAR(k) models has become less important in favour of more elaborate techniques, which are reviewed in the excellent survey paper by Koop et al. (2006).

With the pioneering work on Bayesian cointegration during the mid-1990s, mainly by Bauwens and Lubrano (1993), Geweke (1996), and Kleibergen and van Dijk (1994), focus in estimation of model parameters has been shifted to the VECM representation directly. These approaches have in common that a cointegration rank r is determined and fixed a priori such that the dimension of α and β is known. For suitable informative priors, the full conditionals for these and all the other parameters can be derived, and hence the posterior can be efficiently approximated through Gibbs sampling. The most prominent example is the work by Geweke (1996), who considers the linear normalisation $\beta = (I_r, \beta_l')'$ and chooses normal priors for $\alpha, \beta_l, \phi, \Gamma_1, \ldots, \Gamma_{k-1}$ and an Inverse Wishart prior for Ω. The uncertainty about the cointegration rank can be incorporated into the estimation procedure by carrying out the analysis for every possible value of r and applying Bayesian model selection techniques to the results. However, problems with these approaches exist. Koop et al. (2006) point out that, while the behaviour of standard priors are well-understood in linear estimation, the reduced rank restriction introduces a non-linear estimation problem of $\Pi = \alpha\beta'$, for which the posterior properties of standard priors are not known. Furthermore, the exact form of α depends on the normalisation of β, which makes a suitable prior choice rather difficult. Strachan and Dijk (2004) give an example of inconsistency for the linear normalisation with a diffuse prior for β_l, whose posterior in turn states that such a parametrisation, although specified as such due to prior beliefs, is unlikely. Kleibergen and van Dijk (1994) also find local non-identification problems for certain values of α, leading to improper posteriors for β_l when its prior is improper. Therefore, posterior moments need not exist and the Gibbs sampler may not converge.

Due to these shortcomings with early approaches in direct Bayesian cointegration analysis, these models – although intuitive and convenient – have not become particularly popular. The late 1990s and the years thereafter have seen much further work on Bayesian estimation within the VECM framework, most notably the concept of estimating the so-called *cointegration space* spanned by the columns of β. Other modern approaches include the use of *Jeffreys' prior*, named after Jeffreys (1998) and first discussed by Kleibergen and van Dijk (1994), and the *embedding model*, also introduced by Kleibergen and van Dijk (1994), which nests the VECM as a special case. The general idea and many of the extensions in later years by several authors, as well as problems with these approaches, are reviewed by Koop et al. (2006) and will not be further discussed here. The line of research with the cointegration space, referred to as the *Grassman approach*, is applied in this work and presented in detail. The following summary is adopted from Koop et al. (2006) and Villani (2005).

The Grassman approach goes back to the seminal work by Villani (2000) and was further developed by Villani (2005), Strachan (2003), Strachan and Inder (2004), and Strachan and Dijk (2004). The principal idea is to avoid problems arising from the lack of global identification in the product of $\Pi = \alpha\beta'$ through direct estimation of the cointegration space for $\beta \in \mathbb{R}^{m\times r}$ with full column rank, i.e. $\text{sp}(\beta) := \{\alpha\beta' \colon \alpha \in \mathbb{R}^{m\times r}, \text{rk}(\alpha) = r\}$, the only uniquely estimable quantity given the data. In Bayesian methodology, it follows that the analyst has to specify a prior for all possible outcomes of this quantity of interest rather than for the individual parameters α and β. A diffuse prior then distributes all probability mass uniformly over the support of the cointegration space. The appeal of this approach becomes apparent from a result shown by Strachan and Inder (2004) that such a desired non-informative prior belief would not yield a non-informative but an undesired informative prior for β_l in the linear normalisation.

For illustration purposes of the Grassman approach, consider the simple example of $m = 2$ and $r = 1$, i.e. there are two individually non-stationary processes which share a stationary cointegration relationship. Then β is a 2×1 vector, which can be depicted as an arrow starting at the origin in a two-dimensional coordinate system. The cointegration space for one true vector β is the infinite line through the origin containing the arrow corresponding to β. A prior must distribute the probability mass over the support of all lines through the origin in this coordinate system. In this simple example, this is easily

doable by introducing polar coordinates through $\beta = (\cos\theta, \sin\theta)'$ with $\theta \in [-\pi/2, \pi/2)$. Implicitly, the length of β is w.l.o.g. constrained to unity. A non-informative prior for the support of the cointegration space is hence equivalent to a uniform prior for θ over its support. A more complex example is the case for $m = 3$ and $r = 2$, in which the cointegration space is a two-dimensional plane in a three-dimensional coordinate system, spanned by two linearly independent cointegration vectors given by the columns of the 3×2 matrix β, and the support are all such planes through the origin. Generally, the cointegration space is an r-dimensional hyperplane in the m-dimensional space. Due to the restriction of linear independence between different cointegration vectors, priors in such higher dimensions cannot be simply expressed by distributions over marginal angles for the individual vectors. However, there exist unique distributions over the so-called *Grassman manifolds*, which are equivalent to any chosen prior over the support of the cointegration space. The Grassman manifold $\mathbb{G}_{m,r}$ is the set of all r-dimensional hyperplanes in the m-dimensional space. Obviously, the cointegration space $\mathrm{sp}(\beta)$ must be an element of $\mathbb{G}_{m,r}$. A non-informative distribution over the support for all cointegration spaces is naturally given by a uniform distribution over $\mathbb{G}_{m,r}$. For practical purposes, these abstract distributions can be again transformed to more convenient representations, e.g. for β_l, if the normal linearisation is chosen. Villani (2000, 2005) shows that a diffuse prior on the cointegration space equals a Matrix-t distribution on β_l in this parametrisation. Strachan and Dijk (2004) and Strachan and Inder (2004) discuss possible disadvantages of the linear normalisation and propose alternative and more general approaches, which work directly with the Grassman manifolds without identification restrictions on β, but the generally convenient discussion of the linear normalisation as in Villani (2005) is considered sufficient for this work. For more information on the general estimation procedure w.r.t. Grassman manifolds and, for example, informative priors on the cointegration space, the interested reader is referred to the above cited literature and to, e.g., James (1954) for a general review on this well-understood field of mathematics.

Note that, so far, the Grassman approach has only been concerned with the estimation of β through more abstract quantities. Villani (2005) uses a joint prior for the cointegration space and all remaining parameters. In particular, the priors for $\phi, \Gamma_1, \ldots, \Gamma_{k-1}$ turn out to be non-informative in his work. Motivated by the Bayesian analysis of VAR(k) processes, Warne (2006) extends this approach to apply a Minnesota-like prior with $\Gamma_1, \ldots, \Gamma_{k-1}$. The priors for the cointegration space are further specified in terms of β_l

under the linear normalisation. As a result, in both studies, an efficient Gibbs sampling algorithm becomes available. Note that when the order of the individual time series is chosen such that the last $m - r$ series are not cointegrated solely among themselves, then the analysis based on this prior is invariant to the normalisation. The exact form of the priors and the resulting posterior distributions are presented in Section 3.3. Another main advantage of the approach by Villani (2005) is the possibility of computing a consistent posterior distribution for the cointegration rank r. Consequently, only the lag order k must be specified a priori, whereas the inconsistency of a pre-specified value for r, as seen in many other Bayesian and frequentist approaches, can be eliminated. Alternatively, if the inconsistency between lag order determination and parameter estimation becomes problematic, joint posterior distributions for both k and r can be derived through the extension by Warne (2006).

As a final note, Bayesian predictions of the time series are obtained through simulation of the VECM parameters from their posterior distribution, which is obtained via Gibbs sampling. For each set of realised parameters, the h-step forecast is simulated using the normality of the white noise in the VECM formulation. The necessary initial values $x_{T-k+1}, \ldots, x_T$ can be their observed values for simplicity or, to be more accurate, draws from their posterior predictive distribution.

C.6 Goodness-of-Fit Diagnostics

This section provides a short overview on tools for diagnostics to assess how the model fits the data. Quantitative and qualitative goodness-of-fit checks constitute an important step in any statistical analysis. Model diagnostics in the VECM adopt many tools for VAR models, which are multivariate extensions of well-known diagnostics in the univariate case. Most importantly, residuals should be checked for remaining autocorrelation, unexplained by the model, and non-normality when assuming Gaussian white noise. Residual checks for whiteness are of less concern when the model's main objective is forecasting, and when predictions perform reasonably well. The following review on main techniques is again based on Lütkepohl (2007), who assumes parameter estimation to be done via the frequentist approaches outlined in Appendix C.4. In Bayesian frameworks, the tests are applied with techniques from Appendix A.4.

The estimated *residuals* for the VECM in (C.4) are defined as

$$\widehat{\varepsilon}_t := \Delta x_t - \widehat{\phi} D_t - \widehat{\alpha}\widehat{\beta}' x_{t-1} - \sum_{i=1}^{k-1} \widehat{\Gamma}_i \Delta x_{t-i}$$

for all $t = k+1, \ldots, T$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are the unrestricted least-squares or ML estimators for the general case of k lags, along with the corresponding frequentist estimators for all other parameters. The *residual autocovariances* for any lag $h \in \{0, 1, \ldots, T-k-1\}$ are computed as

$$\widehat{C}_h := \frac{1}{T-k} \sum_{t=k+1+h}^{T} \widehat{\varepsilon}_t \widehat{\varepsilon}'_{t-h}.$$

The *residual autocorrelations* are then given by $\widehat{R}_h = \widehat{D}^{-1}\widehat{C}_h\widehat{D}^{-1}$ for $h = 0, 1, \ldots, T-k-1$, where $\widehat{D}$ is a diagonal matrix with its entries being the square roots of the diagonal elements of $\widehat{C}_0$. Lütkepohl (2007) derives asymptotic normality for both the residual autocovariance and autocorrelation terms under minimum conditions. Plots of the estimated autocorrelation and cross-correlation versus the lag indices for each marginal time series, along with empirical confidence bounds, give a rough check of significance in the residual autocorrelation. Estimates, which exceed approximate bounds for, say, 95% confidence intervals, indicate a lack of fit and give insight in how to adjust the model. More quantitative checks are available through formal hypothesis tests known from standard time series analysis. Brüggemann (2004) shows that the well-known *Portmanteau test* in the VECM framework has the statistic

$$Q_h := (T-k) \sum_{i=1}^{h} \operatorname{tr}\left(\widehat{C}_i' \widehat{C}_0^{-1} \widehat{C}_i \widehat{C}_0^{-1}\right)$$

with some pre-specified $h \in \{1, \ldots, T-k-1\}$ to test $H_0\colon R_1 = \cdots = R_h = 0$ against $H_1\colon R_i \neq 0$ for some i. Here, $\operatorname{tr}(M)$ denotes the trace of a quadratic matrix M, i.e. the sum over all diagonal elements. Under H_0, for sufficiently large h, the statistic has an asymptotic χ^2 distribution with $hm^2 - m^2(k-1) - mr$ degrees of freedom. The limiting distribution for the Portmanteau test is obtained when both the sample size and h go to infinity. This test is hence not suitable to test the significance of residual autocorrelation of low order. Modified versions of the test statistic exist to account for this shortcoming.

The *Lagrange Multiplier* or *Breusch-Godfrey test* due to Breusch (1978) and Godfrey (1978) is an alternative procedure to test for residual autocorrelation of a pre-specified, preferably small lag order h. Here it is assumed that the error terms follow a VAR(h) model, i.e.

$$\varepsilon_t = \Lambda_1 \varepsilon_{t-1} + \cdots + \Lambda_h \varepsilon_{t-h} + \delta_t$$

with white noise δ_t. The null hypothesis of no residual autocorrelation corresponds to $H_0 \colon \Lambda_1 = \cdots = \Lambda_h = 0$ versus $H_1 \colon \Lambda_i \neq 0$ for some i. The Lagrange Multiplier test statistic is then derived as the statistic for a score test on the auxiliary regression model

$$\widehat{\varepsilon}_t = \widehat{\phi} D_t + \widehat{\alpha}\widehat{\beta}' x_{t-1} + \sum_{i=1}^{k-1} \widehat{\Gamma}_i \Delta x_{t-i} + \Lambda_1 \widehat{\varepsilon}_{t-1} + \cdots + \Lambda_h \widehat{\varepsilon}_{t-h} + \delta_t$$

for $t = k+1, \ldots, T$, where $\widehat{\varepsilon}_t = 0$ for $t \leq k$, which Lütkepohl (2007) shows to be

$$\lambda_{LM} := (T-k) \widehat{c}_h' \widehat{\Omega}_c(h)^{-1} \widehat{c}_h,$$

where $\widehat{c}_h := \operatorname{vec}(\widehat{C}_1, \ldots, \widehat{C}_h)$ and $\widehat{\Omega}_c(h) := 1/(T-k) \left(\widehat{E}\widehat{E}' - \widehat{E}X' \left(XX'\right)^{-1} X\widehat{E}' \right) \otimes \widehat{\Omega}$, with $\widehat{\Omega}$ as before, and the following quantities by analogy to the compact matrix formulation from Appendix C.4 for the VECM representation,

$$\begin{aligned}
X_t &= \left(1, x_t', \ldots, x_{t-k+1}'\right)', \quad t = k, \ldots, T-1, \\
X &= \left(X_k, \ldots, X_{T-1}\right), \\
F_i &= \left(0_{T-k \times i}, I_{T-k}\right)' \left(I_{T-k}, 0_{T-k \times i}\right), \quad i = 1, \ldots, h, \\
F &= \left(F_1, \ldots, F_h\right), \\
\widehat{\varepsilon} &= \left(\widehat{\varepsilon}_{k+1}, \ldots, \widehat{\varepsilon}_T\right), \\
\widehat{E} &= \left(I_h \otimes \widehat{\varepsilon}\right) F'.
\end{aligned}$$

Under the same minimum conditions as for the Portmanteau test, the asymptotic distribution for λ_{LM} is $\chi^2(hm^2)$.

Particular interest in goodness-of-fit is devoted to the normality of the residuals if a Gaussian distribution was assumed for the error terms. Rejection of normality indicates

misspecification of the likelihood and, hence, of all statistical inference for the model. A common way to test the residuals for normality is to check whether their skewness and kurtosis, i.e. the third and fourth moments, equal the theoretical values of zero and three, respectively. Such a test, the *Lomnicki-Jarque-Bera test* due to Jarque and Bera (1987) and Lomnicki (1961), is available for the univariate case and can be extended to the multivariate case of VAR models, including the VECM specification. For iid error terms $\varepsilon_t \sim N_m(0, \Omega)$, one obtains $w_t = (w_{1t}, \ldots, w_{mt})' := P^{-1}\varepsilon_t \sim N_m(0, I_m)$, where P is a matrix fulfilling $PP' = \Omega$, and hence $\mathrm{E}(w_{it}^3) = 0$ and $\mathrm{E}(w_{it}^4) = 3$ for all $i = 1, \ldots, m$. Substituting the error terms by the estimated residuals, the test statistics for the null hypotheses of $\mathrm{E}(\widehat{w}_{it}^3) = 0$ and $\mathrm{E}(\widehat{w}_{it}^4) = 3$ are asymptotically $\chi^2(m)$-distributed pivotal quantities of the vectors of empirical third and fourth moments for these standardised residual terms, see Lütkepohl (2007) for exact definitions. Rejection of either of the null hypotheses indicates problems with the normality assumption, where, for small sample sizes, the test should be understood as a rough check only. Further tests in assessment of the fit of a VECM are derived for checks of structural changes within the time series over time, since the statistical inference and forecasts rely on the time invariance of parameters. The interested reader is referred to Lütkepohl (2007) for more information.

D Further Mathematical and Probabilistic Preliminaries

D.1 The Generalised Gamma Function

The following definition introduces a generalisation of the Gamma function as used in the normalising constant of the Grassman prior in (3.10). For simplicity, the function's support is restricted to the case of natural numbers. More information on generalised Gamma functions can be found in, e.g., James (1964).

Definition D.1 (Generalised Gamma Function)**.** *Let $b \in \mathbb{N}_0$. Then the Generalised Gamma function Γ_b is defined for all $a \in \mathbb{N}_0$ with $a \geq b$ through $\Gamma_b(a) := 1$ if $b = 0$, and*

$$\Gamma_b(a) := \prod_{i=1}^{b} \Gamma\left(\frac{a-i+1}{2}\right)$$

if $b > 0$, where $\Gamma(\cdot)$ is the (positive real) Gamma function given by

$$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x)\, dx, \quad t > 0.$$

In particular, since $(a-i+1)/2$ is a multiple of $1/2$ for all $a, b \in \mathbb{N}$ and $i \in \{1, \ldots, b\}$, it follows from the basic properties of the standard Gamma function that

$$\Gamma\left(\frac{a-i+1}{2}\right) = \begin{cases} \left(\frac{a-i-1}{2}\right)!, & \frac{a-i}{2} \notin \mathbb{N} \\ \frac{(a-i)!}{2^{a-i}\left(\frac{a-i}{2}\right)!}\sqrt{\pi}, & \frac{a-i}{2} \in \mathbb{N} \end{cases}.$$

D.2 Definitions in Matrix Algebra

In probabilistic matrix algebra, it is convenient to transform matrices into vectors in order to make use of well-known multivariate distributions instead of rather uncommon matrix-valued distributions.

Definition D.2 (Vectorisation Operator)**.** *Let $M = (m_1 \cdots m_b) \in \mathbb{R}^{a \times b}$ be a matrix with b columns $m_1, \ldots, m_b \in \mathbb{R}^a$. The vectorisation (or vector or column stacking) operator* $\operatorname{vec}\colon \mathbb{R}^{a \times b} \to \mathbb{R}^{ab}$ *transforms the matrix M into the vector* $\operatorname{vec}(M) := (m_1', \ldots, m_b')'$ *of length ab by stacking together all columns from left to right. The vector* $\operatorname{vec}(M)$ *is usually referred to as the vectorised or vectorisation of M.*

When random matrices are transformed into random vectors, corresponding covariance matrices must be changed accordingly. The following notation is useful.

Definition D.3 (Matrix Kronecker Product)**.** *For matrices $M_1 \in \mathbb{R}^{a_1 \times b_1}$, with elements m_{ij} for $i = 1, \ldots, a_1$ and $j = 1, \ldots, b_1$, and $M_2 \in \mathbb{R}^{a_2 \times b_2}$, define their Matrix Kronecker product as*

$$M_1 \otimes M_2 := \begin{pmatrix} m_{11}M_2 & m_{12}M_2 & \cdots & m_{1b_1}M_2 \\ m_{21}M_2 & m_{22}M_2 & \cdots & m_{2b_1}M_2 \\ \vdots & \vdots & \ddots & \vdots \\ m_{a_11}M_2 & m_{a_12}M_2 & \cdots & m_{a_1b_1}M_2 \end{pmatrix} \in \mathbb{R}^{a_1a_2 \times b_1b_2}.$$

Then $\otimes$ is called the Matrix Kronecker product operator.

For more information, see, e.g., Neudecker (1968).

D.3 Matrix-Valued Distributions

When dealing with multivariate problems in Bayesian applications, it becomes necessary to simulate from matrix-valued distributions for the covariance matrices. The following distributions are generalisations of the Normal, t and χ^2 distributions.

A generalisation of the normal distribution for matrix-valued random variables is the Matrix-Normal distribution.

Definition D.4 (Matrix-Normal Distribution)**.** *A random matrix $X \in \mathbb{R}^{a\times b}$ is Matrix-Normal distributed with parameter matrices $M \in \mathbb{R}^{a\times b}, U \in \mathbb{R}^{a\times a}, V \in \mathbb{R}^{b\times b}$, the latter two being positive definite, denoted by $X \sim MN_{a\times b}(M, U, V)$, if X has the density*

$$f(X) = \frac{1}{(2\pi)^{ab/2}\,|V|^{a/2}\,|U|^{b/2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left[V^{-1}(X-M)'U^{-1}(X-M)\right]\right).$$

Note that with the vector operator and Matrix Kronecker product from Definitions D.2 and D.3, $X \sim MN_{a\times b}(M, U, V)$ is equivalent to $\operatorname{vec}(X) \sim N_{ab}(\operatorname{vec}(M), V \otimes U)$, i.e. the vectorised X can be conveniently expresses through a multivariate vector-valued normal distribution. For more information on the Matrix-Normal distribution, see Dawid (1981), for instance.

In analogy to univariate and multivariate normal distributions, a shift to t-distributions is required when variance components are not known but sampled. In matrix-valued probability theory, this analogue is the Matrix-t distribution.

Definition D.5 (Matrix-t Distribution)**.** *A random matrix $X \in \mathbb{R}^{a\times b}$ is Matrix-t distributed with parameters $\mu \in \mathbb{R}^{a\times b}, P \in \mathbb{R}^{a\times a}, Q \in \mathbb{R}^{b\times b}$, the latter two being positive definite, and $n \geq 0$, denoted by $X \sim Mt_{a\times b}(\mu, P, Q, n)$, if X has the density*

$$f(X) = \frac{\Gamma_b(n+a+b-1)\,|P|^{b/2}}{\Gamma_b(n+b-1)\,\pi^{ab/2}\,|Q|^{a/2}} \left|I_b + Q^{-1}(X-\mu)'P(X-\mu)\right|^{-(n+a+b-1)/2}.$$

More information on the Matrix-t distribution can be found in Box and Tiao (2011), for instance.

The Wishart distribution, named after Wishart (1928), generalises the χ^2 distribution to matrix-valued random variables. It has become popular, since in Bayesian methodology, it is a conjugate prior for the precision, i.e. the inverse of the covariance matrix, given a multivariate normal distributed random vector. In this case, the covariance matrix itself is said to be Inverse Wishart distributed.

Definition D.6 (Wishart Distribution)**.** *A positive definite random matrix $X \in \mathbb{R}^{a\times a}$ is Wishart distributed with positive definite scale matrix $V \in \mathbb{R}^{a\times a}$ and $n \geq a$ degrees of*

freedom, denoted by $X \sim W_a(V, n)$*, if* X *has the density*

$$f(X) = \frac{|X|^{(n-a-1)/2}}{2^{na/2}\pi^{a(a-1)/4}\Gamma_a(n)\,|V|^{n/2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(V^{-1}X\right)\right).$$

The Wishart distribution is motivated as distribution for the precision, because for iid $y_i \sim N_a(0, V), i = 1, \ldots, n$, it follows that the scatter matrix $X = (y_1 \cdots y_n)(y_1 \cdots y_n)' \in \mathbb{R}^{a\times a}$ is Wishart distributed with parameters V and n.

Definition D.7 (Inverse Wishart Distribution)**.** *A positive definite random matrix* $X \in \mathbb{R}^{a\times a}$ *is Inverse Wishart distributed with positive definite scale matrix* $S \in \mathbb{R}^{a\times a}$ *and degrees of freedom* $n \geq a$*, denoted by* $IW_a(S, n)$*, if* X *has the density*

$$f(X) = \frac{|S|^{n/2}}{2^{na/2}\pi^{a(a-1)/4}\Gamma_a(n)}\,|X|^{-(n+a+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left(A^{-1}X\right)\right).$$

The mean of X is $\mathrm{E}(X) = S/(n-a-1)$ if $n > a+1$. If $X \sim W_a(V, n)$, then $X^{-1} \sim IW_a(V^{-1}, n)$. See Zellner (1971) for more properties such as variance formulae.

E Numerical Details on the Markov Chain Monte Carlo Algorithm

E.1 Computation of the Acceptance Probability

In the Metropolis-Hastings step for updates of κ_t, the acceptance probability

$$a\left(\kappa_t^{(i-1)}, \kappa_t^*\right) = \min\left\{1, \frac{f\left(\kappa_t^* \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}{f\left(\kappa_t^{(i-1)} \mid \mathcal{D}, \mathcal{E}, \mathcal{H}^{(i)}, \mathcal{K}_{-t}^{(i-1/i)}, k, r\right)}\right\}.$$

must be computed in each iteration step i, where κ_t^* denotes the proposal value. Here, $\mathcal{H}^{(i)} = \left\{\phi^{(i)}, \alpha^{(i)}, \beta^{(i)}, \Gamma^{(i)}, \Omega^{(i)}\right\}$ denotes the set of hyperparameters, which have been already updated at step i via the Gibbs sampler. The expression

$$\mathcal{K}_{-t}^{(i-1/i)} = \left\{\kappa_1^{(i)}, \ldots, \kappa_{t-1}^{(i)}, \kappa_{t+1}^{(i-1)}, \ldots, \kappa_T^{(i-1)}\right\}$$

comprises the latest updated realisations of all other parameters. In Section 3.3.3, it was derived that the posterior density for some κ_t is of the form

$$f\left(\kappa_t \mid \mathcal{D}, \mathcal{E}, \mathcal{H}, \mathcal{K}_{-t}, k, r\right) \propto \exp\left(L_t + P_t\right)$$

with the likelihood-driven component

$$L_t = \sum_x \sum_p \left[D_{xpt} \log\left(\log\left(1 + \exp(\eta_{xpt})\right)\right) - E_{xpt} \log\left(1 + \exp(\eta_{xpt})\right)\right]$$

and prior-driven component

$$P_t = \begin{cases} -\frac{1}{2}\left(\sum_{s=k+1}^{k+t} \varepsilon_s'\Omega^{-1}\varepsilon_s + (\kappa_t - \mu_t)'\,\Sigma_t^{-1}\,(\kappa_t - \mu_t)\right), & t \leq k \\ -\frac{1}{2}\sum_{s=t}^{k+t} \varepsilon_s'\Omega^{-1}\varepsilon_s, & k < t \leq T-k \\ -\frac{1}{2}\sum_{s=t}^{T} \varepsilon_s'\Omega^{-1}\varepsilon_s, & t > T-k \end{cases}.$$

The proportionality factors can be neglected in the evaluation of the acceptance probability as they cancel out. The following sections provide formulae for the numerical evaluation of the acceptance probability both under standard as well as single-component Metropolis-Hastings algorithms.

E.2 General Formulae

Let $\eta_{xpt}^{(i-1)}$ and η_{xpt}^{*} denote the linear predictor from the right-hand side of equation (3.2) when plugging in $\kappa_t^{(i-1)}$ and κ_t^{*}, respectively. Furthermore, for $s = k+1, \ldots, k+t$ if $t \leq k$, and for $s = t, \ldots, \min\{k+t, T\}$ if $t > k$, define

$$\varepsilon_s^{(i-1/i)} = \kappa_s^{(i-1)} - \phi^{(i)} - \left(I_m + \alpha^{(i)}\beta^{(i)\prime}\right)\kappa_{s-1}^{(i-1/i)} - \sum_{j=1}^{k-1}\Gamma_j^{(i)}\Delta\kappa_{s-j}^{(i-1/i)}$$

where

$$\kappa_{s-1}^{(i-1/i)} = \begin{cases} \kappa_{t-1}^{(i)}, & s = t \\ \kappa_{s-1}^{(i-1)}, & s > t \end{cases},$$

$$\Delta\kappa_{s-j}^{(i-1/i)} = \begin{cases} \kappa_{s-j}^{(i)} - \kappa_{s-j-1}^{(i)}, & s-j < t \\ \kappa_{t}^{(i-1)} - \kappa_{t-1}^{(i)}, & s-j = t \\ \kappa_{s-j}^{(i-1)} - \kappa_{s-j-1}^{(i-1)}, & s-j > t \end{cases}.$$

Replacing $\kappa_t^{(i-1)}$ by κ_t^* leads to the residuals

$$\varepsilon_{k+1}^* = \varepsilon_{k+1}^{(i-1/i)} + \Gamma_{k-1}^{(i)} \left(\kappa_1^* - \kappa_1^{(i-1)} \right)$$

if $t = 1$,

$$\begin{aligned} \varepsilon_s^* &= \varepsilon_s^{(i-1/i)} - \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \left(\kappa_t^* - \kappa_t^{(i-1)} \right), \quad s = k+1, \ldots, k+t-1, \\ \varepsilon_{k+t}^* &= \varepsilon_{k+t}^{(i-1/i)} + \Gamma_{k-1}^{(i)} \left(\kappa_t^* - \kappa_t^{(i-1)} \right) \end{aligned}$$

if $t = 2, \ldots, k-1$,

$$\begin{aligned} \varepsilon_{k+1}^* &= \varepsilon_{k+1}^{(i-1/i)} - \left(I_m + \alpha^{(i)} \beta^{(i)\prime} + \Gamma_1^{(i)} \right) \left(\kappa_k^* - \kappa_k^{(i-1)} \right), \\ \varepsilon_s^* &= \varepsilon_s^{(i-1/i)} - \left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)} \right) \left(\kappa_k^* - \kappa_k^{(i-1)} \right), \quad s = k+2, \ldots, 2k-1, \\ \varepsilon_{2k}^* &= \varepsilon_{2k}^{(i-1/i)} + \Gamma_{k-1}^{(i)} \left(\kappa_k^* - \kappa_k^{(i-1)} \right) \end{aligned}$$

if $t = k$,

$$\begin{aligned} \varepsilon_t^* &= \varepsilon_t^{(i-1/i)} + \kappa_t^* - \kappa_t^{(i-1)}, \\ \varepsilon_{t+1}^* &= \varepsilon_{t+1}^{(i-1/i)} - \left(I_m + \alpha^{(i)} \beta^{(i)\prime} + \Gamma_1^{(i)} \right) \left(\kappa_t^* - \kappa_t^{(i-1)} \right), \\ \varepsilon_s^* &= \varepsilon_s^{(i-1/i)} - \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \left(\kappa_t^* - \kappa_t^{(i-1)} \right), \quad s = t+2, \ldots, k+t-1, \\ \varepsilon_{k+t}^* &= \varepsilon_{k+t}^{(i-1/i)} + \Gamma_{k-1}^{(i)} \left(\kappa_t^* - \kappa_t^{(i-1)} \right) \end{aligned}$$

if $t = k+1, \ldots, T-k$,

$$\begin{aligned} \varepsilon_t^* &= \varepsilon_t^{(i-1/i)} + \kappa_t^* - \kappa_t^{(i-1)}, \\ \varepsilon_{t+1}^* &= \varepsilon_{t+1}^{(i-1/i)} - \left(I_m + \alpha^{(i)} \beta^{(i)\prime} + \Gamma_1^{(i)} \right) \left(\kappa_t^* - \kappa_t^{(i-1)} \right), \\ \varepsilon_s^* &= \varepsilon_s^{(i-1/i)} - \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \left(\kappa_t^* - \kappa_t^{(i-1)} \right), \quad s = t+2, \ldots, T, \end{aligned}$$

if $t = T-k+1, \ldots, T-2$,

$$\begin{aligned} \varepsilon_{T-1}^* &= \varepsilon_{T-1}^{(i-1/i)} + \kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}, \\ \varepsilon_T^* &= \varepsilon_T^{(i-1/i)} - \left(I_m + \alpha^{(i)} \beta^{(i)\prime} + \Gamma_1^{(i)} \right) \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)} \right) \end{aligned}$$

if $t = T - 1$, and

$$\varepsilon_T^* = \varepsilon_T^{(i-1/i)} + \kappa_T^* - \kappa_T^{(i-1)}$$

if $t = T$. Note that for all cases to be well-defined, it is assumed that $k < T/2 + 1$.

Evaluation of the acceptance probability w.r.t. $\kappa_t^{(i-1)}$ and κ_t^* gives

$$a\left(\kappa_t^{(i-1)}, \kappa_t^*\right) = \min\left\{1, \exp\left(\Delta L_t^{(i-1/*)} + \Delta P_t^{(i-1/*)}\right)\right\} \tag{E.1}$$

with

$$\begin{aligned}\Delta L_t^{(i-1/*)} =& \sum_x \sum_p D_{xpt}\left[\log\left(\log\left(1+\exp\left(\eta_{xpt}^*\right)\right)\right) - \log\left(\log\left(1+\exp\left(\eta_{xpt}^{(i-1)}\right)\right)\right)\right] \\ &+ \sum_x \sum_p E_{xpt}\left[\log\left(1+\exp\left(\eta_{xpt}^{(i-1)}\right)\right) - \log\left(1+\exp\left(\eta_{xpt}^*\right)\right)\right]\end{aligned}$$

and

$$\begin{aligned}&\Delta P_t^{(i-1/*)} \\ &= \begin{cases} \frac{1}{2}\left[\sum_{s=k+1}^{k+t}\left(\varepsilon_s^{(i-1)\prime}\left(\Omega^{(i)}\right)^{-1}\varepsilon_s^{(i-1)} - \varepsilon_s^{*\prime}\left(\Omega^{(i)}\right)^{-1}\varepsilon_s^*\right) + \zeta_t^{(i-1/*)}\right], & t \le k \\ \frac{1}{2}\sum_{s=t}^{k+t}\left(\varepsilon_s^{(i-1)\prime}\left(\Omega^{(i)}\right)^{-1}\varepsilon_s^{(i-1)} - \varepsilon_s^{*\prime}\left(\Omega^{(i)}\right)^{-1}\varepsilon_s^*\right), & k < t \le T-k\,, \\ \frac{1}{2}\sum_{s=t}^{T}\left(\varepsilon_s^{(i-1)\prime}\left(\Omega^{(i)}\right)^{-1}\varepsilon_s^{(i-1)} - \varepsilon_s^{*\prime}\left(\Omega^{(i)}\right)^{-1}\varepsilon_s^*\right), & t > T-k \end{cases}\end{aligned}$$

where

$$\zeta_t^{(i-1/*)} = \left(\kappa_t^{(i-1)} - \mu_t\right)' \Sigma_t^{-1}\left(\kappa_t^{(i-1)} - \mu_t\right) - \left(\kappa_t^* - \mu_t\right)' \Sigma_t^{-1}\left(\kappa_t^* - \mu_t\right)$$

for $t = 1, \ldots, k$, and the residuals ε_s^* are chosen for t as outlined above. Using that

$$a'Ma - (a+b)'M(a+b) = -\left(a'Mb + b'Ma + b'Mb\right) = -\left(2a'Mb + b'Mb\right)$$

for a symmetric matrix $M \in \mathbb{R}^{n\times n}$ and vectors $a, b \in \mathbb{R}^n$, it can be further deduced that

$$\begin{aligned}
&\Delta P_1^{(i-1/*)} \\
&= -\left(\kappa_1^{(i-1)} - \mu_1\right)' \Sigma_1^{-1} \left(\kappa_1^* - \kappa_1^{(i-1)}\right) - \frac{1}{2}\left(\kappa_1^* - \kappa_1^{(i-1)}\right)' \Sigma_1^{-1} \left(\kappa_1^* - \kappa_1^{(i-1)}\right) \\
&\quad - \varepsilon_{k+1}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_1^* - \kappa_1^{(i-1)}\right) \\
&\quad - \frac{1}{2}\left(\kappa_1^* - \kappa_1^{(i-1)}\right)' \Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_1^* - \kappa_1^{(i-1)}\right)
\end{aligned}$$

if $t = 1$,

$$\begin{aligned}
&\Delta P_t^{(i-1/*)} \\
&= -\left(\kappa_t^{(i-1)} - \mu_t\right)' \Sigma_t^{-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) - \frac{1}{2}\left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \Sigma_t^{-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad + \sum_{s=k+1}^{k+t-1} \varepsilon_s^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2} \sum_{s=k+1}^{k+t-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \varepsilon_{k+t}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2}\left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)
\end{aligned}$$

if $t = 2, \ldots, k-1$,

$$\begin{aligned}
&\Delta P_k^{(i-1/*)} \\
&= -\left(\kappa_k^{(i-1)} - \mu_k\right)' \Sigma_k^{-1} \left(\kappa_k^* - \kappa_k^{(i-1)}\right) - \frac{1}{2}\left(\kappa_k^* - \kappa_k^{(i-1)}\right)' \Sigma_k^{-1} \left(\kappa_k^* - \kappa_k^{(i-1)}\right) \\
&\quad + \varepsilon_{k+1}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_k^* - \kappa_k^{(i-1)}\right) \\
&\quad - \frac{1}{2}\left(\kappa_k^* - \kappa_k^{(i-1)}\right)' \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_k^* - \kappa_k^{(i-1)}\right) \\
&\quad + \sum_{s=k+2}^{2k-1} \varepsilon_s^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)}\right) \left(\kappa_k^* - \kappa_k^{(i-1)}\right) \\
&\quad - \frac{1}{2} \sum_{s=k+2}^{2k-1} \left(\kappa_k^* - \kappa_k^{(i-1)}\right)' \left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)}\right) \left(\kappa_k^* - \kappa_k^{(i-1)}\right) \\
&\quad - \varepsilon_{2k}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_k^* - \kappa_k^{(i-1)}\right) \\
&\quad - \frac{1}{2}\left(\kappa_k^* - \kappa_k^{(i-1)}\right)' \Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_k^* - \kappa_k^{(i-1)}\right)
\end{aligned}$$

if $t = k$,

$$
\begin{aligned}
&\Delta P_t^{(i-1/*)} \\
&= -\varepsilon_t^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) - \frac{1}{2} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad + \varepsilon_{t+1}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad + \sum_{s=t+2}^{k+t-1} \varepsilon_s^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2} \sum_{s=t+2}^{k+t-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \varepsilon_{k+t}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)
\end{aligned}
$$

if $t = k+1, \ldots, T-k$,

$$
\begin{aligned}
&\Delta P_t^{(i-1/*)} \\
&= -\varepsilon_t^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) - \frac{1}{2} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad + \varepsilon_{t+1}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad + \sum_{s=t+2}^{T} \varepsilon_s^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right) \\
&\quad - \frac{1}{2} \sum_{s=t+2}^{T} \left(\kappa_t^* - \kappa_t^{(i-1)}\right)' \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right) \left(\kappa_t^* - \kappa_t^{(i-1)}\right)
\end{aligned}
$$

if $t = T-k+1, \ldots, T-2$,

$$
\begin{aligned}
&\Delta P_{T-1}^{(i-1/*)} \\
&= -\varepsilon_{T-1}^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}\right) - \frac{1}{2} \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}\right) \\
&\quad + \varepsilon_T^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}\right) \\
&\quad - \frac{1}{2} \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}\right)' \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right) \left(\kappa_{T-1}^* - \kappa_{T-1}^{(i-1)}\right)
\end{aligned}
$$

if $t = T - 1$, and

$$\Delta P_T^{(i-1/*)} = -\varepsilon_T^{(i-1)\prime} \left(\Omega^{(i)}\right)^{-1} \left(\kappa_T^* - \kappa_T^{(i-1)}\right) - \frac{1}{2} \left(\kappa_T^* - \kappa_T^{(i-1)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\kappa_T^* - \kappa_T^{(i-1)}\right)$$

for $t = T$. For numerical efficiency, the following terms should be computed once at the beginning as they occur often, but do not change in the course of the Metropolis-Hastings algorithm:

- $\left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}$
- $\Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}$
- $\left(\Omega^{(i)}\right)^{-1} \left(\Gamma_j^{(i)} - \Gamma_{j-1}^{(i)}\right)$ for $j = 2, \ldots, k-1$
- $\left(\Gamma_j^{(i)} - \Gamma_{j-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_j^{(i)} - \Gamma_{j-1}^{(i)}\right)$ for $j = 2, \ldots, k-1$
- $\left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)$
- $\left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)$

E.3 Expressions under Single-Component Metropolis-Hastings

As a special case of the procedure in the previous section, assume that for a given calendar year $t \in \{1, \ldots, T\}$, each component of κ_t is visited and updated individually. In particular, at iteration step i in the Metropolis-Hastings algorithm, the random vector κ_t^* is not simulated once from an m-dimensional normal distribution. Instead, for each component $j = 1, \ldots, m$, a univariate proposal value for the j-th entry of κ_t is simulated from a one-dimensional normal distribution with the corresponding marginal mean and variance parameters. Updates of the components are done individually, i.e. for each j, the acceptance probability is computed and a realisation of a uniform random variable is simulated. As a consequence, during each iteration i, the Metropolis-Hastings step is now conducted mT instead of T times. However, simulation of proposal values has become a univariate process, and acceptance rates generally increase, because more plausible components will not be rejected as often as before, when other components caused the proposed *vector* to be less likely. Note that with this procedure, the proposals for

different components in κ_t become independent. For more information on the tuning effects of this algorithm, the reader is referred to Gilks (2005).

For the following discussion, assume that $j \in \{1, \ldots, m\}$ is arbitrary but fixed. A new proposal for the j-th component of κ_t is simulated from a normal distribution with mean being the j-th component of $\kappa_t^{(i-1)}$ and variance being the entry (j, j) of Σ_{MH}. Let δ be the difference of the realised value and the mean, i.e. the current value for the j-th component. Then in the vector notation as before, where $\kappa_t^{(i-1)}$ is now implicitly assumed to consist of the i-th values for components $1, \ldots, j-1$ and $(i-1)$-th components for $j, \ldots, m$, one can write

$$\kappa_t^* = \kappa_t^{(i-1)} + \delta e_j,$$

where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)'$ is the canonical vector having entry 1 at j-th position. Hence, in the results from Section E.2, the terms $\kappa_t^* - \kappa_t^{(i-1)}$ can be replaced by δe_j. Noting that δ can be factored out as a scalar, the remaining multiplication of matrices with e_j from the right simply leaves their j-th columns, and one similarly gets the entry (j, j) when multiplying with e_j from both sides. Denoting by subscripts j and jj the j-th column and the entry (j, j) of a matrix, the formulae for the prior-driven term $\Delta P_t^{(i-1/*)}$ in (E.1) then simplify to

$$\begin{aligned}\Delta P_1^{(i-1/*)} = -\delta \Bigg[&\left(\kappa_1^{(i-1)} - \mu_1\right)' \left(\Sigma_1^{-1}\right)_j + \frac{1}{2}\delta \left(\Sigma_1^{-1}\right)_{jj} \\ &+ \varepsilon_{k+1}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}\right)_j + \frac{1}{2}\delta \left(\Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}\right)_{jj} \Bigg]\end{aligned}$$

if $t = 1$,

$$\begin{aligned}\Delta P_t^{(i-1/*)} = -\delta \Bigg[& \left(\kappa_t^{(i-1)} - \mu_t\right)' \left(\Sigma_t^{-1}\right)_j + \frac{1}{2}\delta \left(\Sigma_t^{-1}\right)_{jj} \\ & - \sum_{s=k+1}^{k+t-1} \varepsilon_s^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right)\right)_j \\ & + \frac{1}{2}\delta \sum_{s=k+1}^{k+t-1} \left(\left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)}\right)\right)_{jj} \\ & + \varepsilon_{k+t}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}\right)_j \\ & + \frac{1}{2}\delta \left(\Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}\right)_{jj} \Bigg]\end{aligned}$$

if $t = 2, \ldots, k-1$,

$$\begin{aligned}\Delta P_k^{(i-1/*)} = -\delta \Bigg[& \left(\kappa_k^{(i-1)} - \mu_k\right)' \left(\Sigma_k^{-1}\right)_j + \frac{1}{2}\delta \left(\Sigma_k^{-1}\right)_{jj} \\ & - \varepsilon_{k+1}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)\right)_j \\ & + \frac{1}{2}\delta \left(\left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)}\right)\right)_{jj} \\ & - \sum_{s=k+2}^{2k-1} \varepsilon_s^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)}\right)\right)_j \\ & + \frac{1}{2}\delta \sum_{s=k+2}^{2k-1} \left(\left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)}\right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-k}^{(i)} - \Gamma_{s-k-1}^{(i)}\right)\right)_{jj} \\ & + \varepsilon_{2k}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}\right)_j \\ & + \frac{1}{2}\delta \left(\Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)}\right)_{jj} \Bigg]\end{aligned}$$

if $t = k$,

$$\begin{aligned}\Delta P_t^{(i-1/*)} = -\delta \Bigg[& \varepsilon_t^{(i-1)\prime} \left(\Omega^{(i)}\right)_j^{-1} + \frac{1}{2}\delta \left(\Omega^{(i)}\right)_{jj}^{-1} \\ & - \varepsilon_{t+1}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right) \right)_j \\ & + \frac{1}{2}\delta \left(\left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right) \right)_{jj} \\ & - \sum_{s=t+2}^{k+t-1} \varepsilon_s^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \right)_j \\ & + \frac{1}{2}\delta \sum_{s=t+2}^{k+t-1} \left(\left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \right)_{jj} \\ & + \varepsilon_{k+t}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \right)_j \\ & + \frac{1}{2}\delta \left(\Gamma_{k-1}^{(i)\prime} \left(\Omega^{(i)}\right)^{-1} \Gamma_{k-1}^{(i)} \right)_{jj} \Bigg]\end{aligned}$$

if $t = k+1, \ldots, T-k$,

$$\begin{aligned}\Delta P_t^{(i-1/*)} = -\delta \Bigg[& \varepsilon_t^{(i-1)\prime} \left(\Omega^{(i)}\right)_j^{-1} + \frac{1}{2}\delta \left(\Omega^{(i)}\right)_{jj}^{-1} \\ & - \varepsilon_{t+1}^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right) \right)_j \\ & + \frac{1}{2}\delta \left(\left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right) \right)_{jj} \\ & - \sum_{s=t+2}^{T} \varepsilon_s^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \right)_j \\ & + \frac{1}{2}\delta \sum_{s=t+2}^{T} \left(\left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right)' \left(\Omega^{(i)}\right)^{-1} \left(\Gamma_{s-t}^{(i)} - \Gamma_{s-t-1}^{(i)} \right) \right)_{jj} \Bigg]\end{aligned}$$

if $t = T - k + 1, \ldots, T - 2$,

$$\begin{aligned}\Delta P_{T-1}^{(i-1/*)} = -\delta \Bigg[& \varepsilon_{T-1}^{(i-1)\prime} \left(\Omega^{(i)}\right)_j^{-1} + \frac{1}{2}\delta \left(\Omega^{(i)}\right)_{jj}^{-1} \\ & - \varepsilon_T^{(i-1)\prime} \left(\left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right) \right)_j \\ & + \frac{1}{2}\delta \left(\left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right)' \left(\Omega^{(i)}\right)^{-1} \left(I_m + \alpha^{(i)}\beta^{(i)\prime} + \Gamma_1^{(i)} \right) \right)_{jj} \Bigg]\end{aligned}$$

if $t = T - 1$, and

$$\Delta P_T^{(i-1/*)} = -\delta \left[\varepsilon_T^{(i-1)\prime} \left(\Omega^{(i)}\right)_j^{-1} + \frac{1}{2}\delta \left(\Omega^{(i)}\right)_{jj}^{-1} \right]$$

if $t = T$.

Further note that for such $j \in \{1, \ldots, m\}$, which correspond to global parameters that are not specific for a certain population p, the likelihood-driven term is conveniently computed for the average population p^* only. This minimises the computational effort in the computation of the acceptance probability without any loss of information.

F The R Package `bmpmp`

This appendix introduces the R package `bmpmp` for flexible and efficient estimation and creation of output of the BMPMP model. The package is specifically designed for joint forecasts of different countries with distinction of both genders. It consists of the following five main functions, which will be described in more detail in the respective sections F.1 to F.5.

- `create.data` reads country-specific observed numbers of deaths D_{xgt} and exposure-to-risk E_{xgt} for both genders as provided by the Human Mortality Database (2014) and creates one input data file for the estimation procedure of the BMPMP model.
- `bmpmp.estimation` runs the MCMC algorithm for the BMPMP model using the input data file created by `create.data` and the modeller's choices for the prior distributions.
- `bmpmp.estimation.continue` continues the MCMC algorithm for the BMPMP model with the same input data and prior distributions, when output has already been created by `bmpmp.estimation` (or `bmpmp.estimation.continue`).
- `bmpmp.plots` constructs posterior distribution, convergence, validation, and forecast plots for the output from `bmpmp.estimation`
- `ml.estimation.plots` estimates the CBD model and VECM equations in the BMPMP model via ML (if not over-parametrised), using input data files created by `create.data` and `bmpmp.estimation`, and constructs forecast plots.

It is noteworthy that, although the term BMPMP model is used throughout this appendix, the R package allows all functions to be applied for the case of univariate single-country mortality forecasts.

F.1 The Function `create.data`

Description

`create.data` reads country-specific observed numbers of deaths D_{xgt} and exposure-to-risk E_{xgt} for both genders as provided by the Human Mortality Database (2014) and creates one input data file for the estimation procedure of the BMPMP model.

Usage

```
create.data(x_min = 40, x_max = 100, t_min = 1956, t_max = 2009,
            countries, name_overall = "Total",
            death_files, exposure_files,
            output_directory = paste(getwd(),"/output/", sep = ""))
```

Arguments

`x_min`	The minimum age x_0 for the model. Must be a scalar in $\mathbb{N}_0$. Default is 40.
`x_max`	The maximum age for the model. Must be a scalar in $\mathbb{N}$ exceeding `x_min` but not exceeding the maximum possible age in the input data. Default is 100.
`t_min`	The minimum calendar year for the calibration period. Must be a scalar in $\mathbb{N}$ chosen such that calibration period is covered by all input data. Default is 1956.
`t_max`	The maximum calendar year for the calibration period. Must be a scalar in $\mathbb{N}$ exceeding `t_min` and chosen such that calibration period is covered by all input data. Default is 2009.
`countries`	The names (or any other identification) of the countries which shall be analysed. Must be a character object in case of one single country and a vector with character entries in case of several countries.
`name_overall`	Character object with the name (or any other identification) for the reference population, i.e. the overall population consisting of all countries under consideration. Only required if several countries are to be analysed. Default is `"Total"`.

`death_files`	Character object/vector containing the location paths of the input data file(s) in `txt` format for the observed numbers of deaths. Must be of the same length as `countries` and match in order. See 'Details' for the required input format.
`exposure_files`	Character object/vector containing the full paths of the input data file(s) in `txt` format for the exposure-to-risk. Must be of the same length as `countries` and match in order. See 'Details' for the required input format.
`output_directory`	Character object specifying the directory for the output data file `input_data.RData`. Must end with `/"`. Default is `paste(getwd(),"/output/", sep = ""))`, i.e. a folder called `output` in the current workspace. For the current workspace simply set `output_directory = "/"`.

Details

The routine extracts the required observations from the individual input data files for the calibration window defined through the parameters `x_min`, `x_max`, `t_min`, `t_max`. Data for both genders of each country are stored, and in case of several countries, the reference population of all males in all countries is derived. Furthermore, other important quantities, such as the dimension of the corresponding time series for the VECM, are computed. While the routine is running, status messages are provided.

For `death_files` and `exposure_files`, each of the vector entries (or the object itself in case of one country) must be a character specifying the full path of the corresponding input data file (either full path or starting from the current workspace). The input data files must be the *1x1 Deaths* and *1x1 Exposure-to-Risk* `txt` files provided by the Human Mortality Database (2014), available at `http://www.mortality.org` under *Complete Data Series (Period data)* for each country[1]. As an example for the observed number of deaths in Ireland (effective 24 September 2014), the input data file's header should look like the following.

[1]Note that when certain files have entries `.` to denote missing data, as found for Belgium for the years 1914–1918, they must be replaced by `0.00` for the routine to work.

deaths_ireland.txt

```
Ireland, Deaths 1x1     Last modified: 12-Nov-2010, MPv5 May07

Year     Age        Female             Male            Total
1950      0           1246.00          1676.00          2922.00
1950      1            101.00           142.00           243.00
1950      2             60.00            73.00           133.00
1950      3             52.00            59.00           111.00
1950      4             48.00            45.00            93.00
1950      5             39.00            30.00            69.00
1950      6             22.00            18.00            40.00
1950      7             23.00            34.00            57.00
1950      8             18.00            30.00            48.00
1950      9             11.00            17.00            28.00
1950     10             21.00            18.00            39.00
```

Value

The routine does not return but saves the output as a list in `input_data.RData` in the output directory as required by the function `bmpmp.estimation` for estimation of the BMPMP model for the specified input data.

The function `load` can be used to read the list into the current workspace. It consists of the following components.

`age_levels`	A vector containing all ages under consideration from `x_min` to `x_max`.
`calendar_years`	A vector containing all calendar years under consideration from `t_min` to `t_max`.
`countries`	The input object `countries` in case of one country or the vector `countries` extended by `name_overall` otherwise.
`x_min`	The supplied value for `x_min`.
`x_max`	The supplied value for `x_max`.
`C`	The number of countries including the reference population (will equal one in case of one country).
`T`	The number of calendar years of the calibration window.

M	The number of ages of the calibration window.
m	The dimension of the multivariate time series for the VECM.
D_m	In case of one country: A matrix containing all observed number of deaths for males for all ages (given by rows) and calendar years (given by columns). In case of several countries: A list containing all observed numbers of deaths for males for all countries (given by list entries), ages (given by respective rows), and calendar years (given by respective columns).
D_f	In case of one country: A matrix containing all observed number of deaths for females for all ages (given by rows) and calendar years (given by columns). In case of several countries: A list containing all observed numbers of deaths for females for all countries (given by list entries), ages (given by respective rows), and calendar years (given by respective columns).
E_m	In case of one country: A matrix containing all exposure-to-risk for males for all ages (given by rows) and calendar years (given by columns). In case of several countries: A list containing all exposure-to-risk for males for all countries (given by list entries), ages (given by respective rows), and calendar years (given by respective columns).
E_f	In case of one country: A matrix containing all exposure-to-risk for females for all ages (given by rows) and calendar years (given by columns). In case of several countries: A list containing all exposure-to-risk for females for all countries (given by list entries), ages (given by respective rows), and calendar years (given by respective columns).
logit_Q_m	In case of one country: A matrix containing all observed logits for mortality rates under assumption (2.1) for males for all ages (given by rows) and calendar years (given by columns). In case of several countries: A list containing all observed logits of mortality rates under assumption (2.1) for males for all countries (given by list entries), ages (given by respective rows), and calendar years (given by respective columns).

logit_Q_f	In case of one country: A matrix containing all observed logits for mortality rates under assumption (2.1) for females for all ages (given by rows) and calendar years (given by columns). In case of several countries: A list containing all observed logits of mortality rates under assumption (2.1) for females for all countries (given by list entries), ages (given by respective rows), and calendar years (given by respective columns).
own_col	An object (if one country) or vector (if several countries) of characters defining colours for graphical output.
death_files	The supplied object/vector for death_files.
exposure_files	The supplied object/vector for exposure_files.

Examples

```
### Example for Big Five as in Section 4.1

# Set workspace first
# Save input data from http://www.mortality.org
# under the following paths
death_files <- c("input_data\\deaths_west_germany.txt",
                 "input_data\\deaths_spain.txt",
                 "input_data\\deaths_france.txt",
                 "input_data\\deaths_italy.txt",
                 "input_data\\deaths_united_kingdom.txt")
exposure_files <- c("input_data\\exposure-to-risk_west_germany.txt",
                    "input_data\\exposure-to-risk_spain.txt",
                    "input_data\\exposure-to-risk_france.txt",
                    "input_data\\exposure-to-risk_italy.txt",
                    "input_data\\exposure-to-risk_united_kingdom.txt")

# Give corresponding names for countries and reference population
countries <- c("DE","ES","FR","IT","UK")
name_overall <- "EU"
```

```
# Set output directory for combined dataset (folder must exist)
output_directory <- paste(getwd(),"/output/", sep = "")

# Define calibration window
x_min <- 40
x_max <- 100
t_min <- 1956
t_max <- 1995

# Run function
create.data(x_min = x_min, x_max = x_max, t_min = t_min, t_max = t_max,
            countries, name_overall, death_files, exposure_files,
            output_directory = paste(getwd(),"/output/", sep = ""))

### End example
```

F.2 The Function `bmpmp.estimation`

Description

`bmpmp.estimation` runs the MCMC algorithm for the BMPMP model using the input data file created by `create.data` and the modeller's choices for the prior distributions.

Usage

```
bmpmp.estimation(N = 1000000, thinning = 100, save = 2000,
                 output_directory = paste(getwd(),"/output/", sep = ""),
                 cbd_plots = TRUE, cbd_plots_directory = "cbd_plots/",
                 titles = TRUE, tikz_format = FALSE,
                 k = 2, r = 1, w = 1, lambda_A = 1, lambda_alpha = 1,
                 lambda_b = 5, lambda_l = 1, mhc = 1)
```

Arguments

`N` Number of iterations N for the MCMC algorithm. Must be a scalar in $\mathbb{N}_0$. Default is 1,000,000. If `N` is chosen to be zero, the routine will only initialise all quantities for the MCMC algorithm.

`thinning` Thinning factor for the MCMC algorithm if $N > 0$, i.e. every `thinning`-th iteration is stored. Must be a positive divisor of `N`. Default is 100.

`save` Number of iterations for which intermediate results subject to thinning should be stored if $N > 0$. Must be a positive divisor of `N` and a multiple of `thinning`. Default is 2,000.

`output_directory` Character object specifying the directory for both the input data file `input_data.RData` created through the routine `create.data` and the output data files `initial.RData`, `constants.RData` and, if $N > 0$, `bmpmp.RData`. Must end with `/"`. Default is `paste(getwd(),"/output/", sep = ""))`, i.e. a folder called `output` in the current workspace. For the current workspace simply set `output_directory = "/"`.

`cbd_plots` Logical value indicating whether starting values for the CBD model, i.e. its ML estimates, should be plotted. Default is `TRUE`.

`cbd_plots_directory` Character object specifying the directory within the output directory given through `output_directory` for plots of the starting values for the CBD model if `cbd_plots = TRUE`. Must end with `/"`. Default is `"cbd_plots/"`, i.e. a folder called `cbd_plots` in the output directory. For the `output` directory itself simply set `cbd_plots_directory = "/"`.

`titles` Logical value if `cbd_plots = TRUE`, indicating whether main and axis titles should be included in the graphical output for the starting values. Default is `TRUE`.

`tikz_format` Logical value if `cbd_plots = TRUE`, indicating whether the output for the starting values should not be graphs stored in `pdf` format as by default, but files in `tikz` format to allow easy inclusion of graphs in LaTeX documents. Default is `FALSE`.

`k` Lag order k in the VAR model representation, i.e. the lag order for the VECM is $k - 1$. Must be a scalar in $\mathbb{N}$. Default is 2.

`r`	Cointegration rank r in the VECM. Must be a scalar in $\mathbb{N}_0$. Default is 1.
`w`	Posterior weight w for the likelihood and prior in the Metropolis-Hastings algorithm. Must be a scalar in $(0, 1]$. Default is 1, i.e. the original Metropolis-Hastings algorithm.
`lambda_A`	Tuning factor λ_A in the tuning function f_A for the determination of A. Must be a scalar exceeding 0. Default is 1. See 'Details' for information on the initialisation of A.
`lambda_alpha`	Constant λ_α for the prior of α. Must be a scalar exceeding 0. Default is 1.
`lambda_b`	Baseline constant λ_b for the overall magnitude of variance in the prior of Γ if $k > 1$. Must be a scalar exceeding 0. Default is 5.
`lambda_l`	Lag constant λ_b for the shrinkage in variance with increasing order in the prior of Γ if $k > 2$. Must be a scalar exceeding 0. Default is 1.
`mhc`	Tuning constant c_{MH} for the Metropolis-Hastings proposal variance Σ_{MH}. Must be a scalar exceeding 0. Default is 1. See 'Details' for information on the initialisation of Σ_{MH}.

Details

The routine initialises and, if $N > 0$, runs the Bayesian estimation algorithm for the BMPMP model with the data supplied through `input_data.RData` in the output directory. When there is more than one country to be analysed, the exact equation for the CBD model applied in this function is given by

$$\begin{aligned}\log\left(\frac{q_{xpgt}}{1-q_{xpgt}}\right) &= \kappa_t^0 + \kappa_t^p + \kappa_t^g + \kappa_t^{pg} + \left(\kappa_t^x + \kappa_t^{xp} + \kappa_t^{xg}\right)(x - x_0) \\ &\quad + \left(\kappa_t^{x^2} + \kappa_t^{x^2 p} + \kappa_t^{x^2 g}\right)(x - x_0)^2, \quad x \geq x_0,\end{aligned}$$

with t, x, p running through the calendar years, ages, and populations (including the reference population) specified in `input_data.RData`, and g denoting the effects for females compared to males. The minimum age x_0 coincides with `x_min`. In case of only one country, the above model is applied with all population-specific parameters set equal to zero.

For the Metropolis-Hastings algorithm within the estimation procedure, the single-component method is applied. The starting values are the ML estimates for the CBD parameters, A for Ω, and zero matrices or vectors for all other hyperparameters. The matrix A is initialised through $A = f_A(\widehat{\Omega})$, where $f_A(M) = (\lambda_A^2 m_{ij}^2)_{ij}$ for a square matrix $M = (m_{ij})_{ij}$ with λ_A as specified in `lambda_A`, and $\widehat{\Omega}$ is the empirical covariance matrix of the time series in differences. If this choice of A becomes numerically singular, its off-diagonal elements are set to zero. The constant q is automatically given by $m + 2$, where m is the dimension of the time series, stored in `m` in `input_data.RData`. All other prior constants $\lambda_\alpha, \lambda_b, \lambda_l$ can be directly selected by the modeller through `lambda_alpha`, `lambda_b`, `lambda_l`. The constants μ_t and Σ_t for the priors regarding the first k CBD parameters are automatically determined through an empirical Bayes approach, i.e. $\mu_1, \ldots, \mu_k$ are chosen to be the corresponding ML estimates and $\Sigma_1 = \cdots = \Sigma_k = A$. The Metropolis-Hastings proposal variances are chosen to be the diagonal elements of $\Sigma_{\text{MH}} = c_{\text{MH}} A$, where c_{MH} can be determined by the user through `mhc`. The function further allows for the choice of a posterior weight in `w`.

Since the required number of iterations N must be chosen to be large in general, sufficiently high thinning and a low value for `save` must be selected to guarantee enough data allocation space within the R workspace and to reduce the necessary storage space for the output.

If `cbd_plots = TRUE`, the routine will construct plots of the starting values in `pdf` or `tikz` format, stored in the folder specified through `cbd_plots_directory`.

The function `bmpmp.estimation` requires the `MCMCpack` package if $N > 0$ and the `tikzDevice` package if `tikzformat = TRUE`. While the routine is running, status messages are provided. The routine automatically sets seeds for reproducible output. Error messages will be given for improper choices of k and r w.r.t. the latent dimension and horizon of the time series. Note that the choice of $r = 0$ corresponds to a stationary VAR(k) model for the time series in levels, where $r = m$ corresponds to a stationary VAR($k - 1$) model for the time series in differences.

Value

The routine does not return but saves the output as lists in the files `initial.RData`,

constants.RData, and bmpmp.RData in the output directory as required by the function bmpmp.plots for graphical outputs for the analysis of the BMPMP model.

The function load can be used to read any of the lists into the current workspace. The list initial.RData consists of the following components.

data	An $m \times T - k$ matrix containing the multivariate time series in levels (without the k initial values). The marginal time series (given by the rows) are ordered as follows: $\kappa^0, \kappa^g, \kappa^x, \kappa^{x^2}, \kappa^{xg}, \kappa^{x^2 g}$, all κ^p in the order of countries, then all κ^{pg}, all κ^{xp}, and all $\kappa^{x^2 p}$ in the same order, respectively. In case of the analysis of one single country, data consists of the first six time series only.
delta_x	An $m \times T - k - 1$ matrix containing the multivariate time series given in data in differences.
Z_0	The $m \times T - k$ matrix $\Delta K = (\Delta\kappa_{k+1}, \ldots, \Delta\kappa_T)$ in the compact matrix form for the VECM.
Z_1	The $m \times T - k$ matrix $K_1 = (\kappa_k, \ldots, \kappa_{T-1})$ in the compact matrix form for the VECM.
Z_2	The $(k-1)m \times T - k$ matrix K_2 with entries $\left(\Delta\kappa'_{T-1}, \ldots, \Delta\kappa'_{T-k+1}\right)'$ in the compact matrix form for the VECM.
Omega	The initial value for Ω.
Omega_inverse	The inverse of the initial value of Ω.
Phi	The initial value for ϕ.
Gamma	The initial value for Γ.
alpha	The initial value for α.
beta	The initial value for β.
Psi	The initial value for β_l.

The list constants.RData consists of the following components.

C	The number of countries including the reference population (will equal one in case of one country).
T	The number of calendar years of the calibration window.
M	The number of ages of the calibration window.

`k`	The lag order k in the VAR model representation, i.e. the lag order for the VECM is $k-1$.
`r`	The cointegration rank r in the VECM.
`m`	The dimension of the multivariate time series for the VECM.
`d`	The dimension of the deterministic parameters in the VECM, which is one by default.
`D_t`	The matrix of time-varying coefficients for the deterministic parameters in the VECM, i.e. a $1 \times T$ matrix of ones.
`D_0`	The last $T-k$ columns of `D_t`.
`lambda_alpha`	The constant λ_α.
`A`	The constant matrix A.
`Sigma_inverse`	The inverse of Ω_Γ.
`mu`	A matrix (or vector if $k=1$) consisting of $\mu_1, \ldots, \mu_k$.
`B_inverse`	The inverse of $\Sigma = \Sigma_1 = \cdots = \Sigma_k = A$.
`mhc`	The tuning constant c_{MH} for the Metropolis-Hastings proposal variance Σ_{MH}.
`w`	The posterior weight w for the likelihood and prior in the Metropolis-Hastings algorithm.
`n_Omega`	The value $n_\Omega = T - k + q + r + m(k-1)$.
`c`	The matrix $c = (I_r, 0_{r \times m-r})'$.
`c_orthogonal`	The matrix $c_\perp = (0_{m-r \times r}, I_{m-r})'$.
`H`	The matrix $H = I_r \otimes c_\perp$.
`h`	The vector $h = \text{vec}(H)$.
`age_levels`	A vector containing all ages under consideration from `x_min` to `x_max`.

In case of $N > 0$, the list `bmpmp.RData` consists of the following components.

`mcmc`	A list, whose i-th entry is a list containing all values of the quantities in `initial.RData` for the i-th iteration (after thinning).

`eta` A list, whose i-th entry is a list with two entries. In case of several countries, the first entry is a list with $T-k$ entries, whose t-th entry is a matrix of the linear predictors for all male populations at t-th calendar year realised at the i-th iteration (after thinning), where the rows indicate the countries as given in `countries` (with the last row being the reference population) and the columns indicate the ages as given in `age_levels`. In case of a single country, the first entry is a list with $T-k$ entries, whose t-th entry is a vector of the linear predictors for the male population at t-th calendar year realised at the i-th iteration (after thinning), where the entries indicate the ages as given in `age_levels`. The second entry of i-th entry of `eta` is the corresponding list of all female linear predictors, arranged in the analogous way.

`N_thinned` The number of iterations after thinning.

In the alternative case of $N = 0$, the list `bmpmp.RData` consists of the following components.

`mcmc` A list of one entry, which is a list containing all values of the quantities in `initial.RData`.

`eta` A list of one entry, which is a list with two entries. In case of several countries, the first entry is a list with $T - k$ entries, whose t-th entry is a matrix of the linear predictors for all male populations at t-th calendar year realised for the initial values, where the rows indicate the countries as given in `countries` (with the last row being the reference population) and the columns indicate the ages as given in `age_levels`. In case of a single country, the first entry is a list with $T-k$ entries, whose t-th entry is a vector of the linear predictors for the male population at t-th calendar year realised for the initial values, where the entries indicate the ages as given in `age_levels`. The second entry of the single entry of `eta` is the corresponding list of all female linear predictors, arranged in the analogous way.

`N_thinned` The number zero.

In case of $N > 0$, additional output files are the intermediate results for each increment of `save` iterations, which are automatically contained in `bmpmp.RData`. Their lists are set up as in the case of `bmpmp.RData`, where the entries are called `mcmc_thinned`, `eta_thinned`, `number_thinned`.

Finally, if `cbd_plots = TRUE`, the routine saves graphical output in `pdf` or `tikz` format as outlined under 'Details'.

Examples

```
### Example for Big Five as in Section 4.1
### (Example for create.data continued)

# Required package (must be installed)
library(MCMCpack) # (will also be loaded within the function)

# Set workspace first and create input_data.RData through create.data()
# Set output directory as before
output_directory <- paste(getwd(),"/output/", sep = "")

# Set directory for CBD plots within the output directory
# (folder must exist)
cbd_plots_directory <- "cbd_plots/"
cbd_plots <- TRUE
titles <- TRUE
tikz_format <- FALSE

# Setup of model
k <- 2
r <- 5
w <- 1
lambda_A <- sqrt(10)
lambda_alpha <- 1
lambda_b <- 5
```

```
lambda_l <- 1  # (not required)
mhc <- 10

# Setup algorithm
N <- 1000000
thinning <- 100
save <- 1000

# Run function
bmpmp.estimation(N = N, thinning = thinning, save = save,
                 output_directory = output_directory,
                 cbd_plots = cbd_plots,
                 cbd_plots_directory = cbd_plots_directory,
                 titles = titles, tikz_format = tikz_format,
                 k = k, r = r, w = w, lambda_A = lambda_A,
                 lambda_alpha = lambda_alpha,
                 lambda_b = lambda_b, lambda_l = lambda_l, mhc = mhc)

### End example
```

F.3 The Function `bmpmp.estimation.continue`

Description

`bmpmp.estimation.continue` continues the MCMC algorithm for the BMPMP model with the same input data and prior distributions, when output of `bmpmp.estimation` (or `bmpmp.estimation.continue`) has already been created.

Usage

```
bmpmp.estimation.continue(N = 2000000, thinning = 100, save = 2000,
                          output_directory
                          = paste(getwd(),"/output/", sep = ""))
```

Arguments

`N`	Number of iterations N for the entire MCMC algorithm, i.e. the final number of iterations including previously obtained results. Must be a scalar in $\mathbb{N}$ exceeding N_{prev}, i.e. the number of iterations before thinning used in `bmpmp.estimation` (or the last call of `bmpmp.estimation.continue`). Default is 2,000,000.
`thinning`	Thinning factor for the new section of the MCMC algorithm, i.e. every `thinning`-th iteration in the current part of the MCMC algorithm is stored. Must be a positive divisor of the additional number of iterations $N - N_{\text{prev}}$. Default is 100.
`save`	Number of iterations for which intermediate results subject to thinning should be stored in the new section of the MCMC algorithm. Must be a positive divisor of the additional number of iterations $N - N_{\text{prev}}$ and a multiple of `thinning`. Default is 2,000.
`output_directory`	Character object specifying the directory for the input data files `input_data.RData`, `constants.RData`, as well as the previous results in `bmpmp.RData` created through the routine `bmpmp.estimation` (or `bmpmp.estimation.continue`), and the new output data file, which will override the existing `bmpmp.RData`. Must end with `/"`. Default is `paste(getwd(),"/output/", sep = ""))`, i.e. a folder called `output` in the current workspace. For the current workspace simply set `output_directory = "/"`.

Details

The routine continues the Bayesian estimation algorithm for the same BMPMP model as in the previously used `bmpmp.estimation` (or `bmpmp.estimation.continue`). In particular, the same input data and choices for the prior distributions given in the respective files in the output directory are used. The routine carries forward the previous results as stored in `bmpmp.RData` and resumes the MCMC algorithm with the realisation for the last iteration therein. New realisations are combined with the already given iterations such that the output comprises the entire output of the algorithm. As a special case, `bmpmp.estimation.continue` starts the MCMC algorithm in case `bmpmp.estimation` was used for initialisation only, i.e. with `N` begin zero.

The function `bmpmp.estimation.continue` requires the `MCMCpack` package. While the

routine is running, status messages are provided. The routine automatically sets seeds for reproducible output (even if next time `bmpmp.estimation` was used for all iterations at once). Notes made in the 'Details' section for the function `bmpmp.estimation` regarding the parameters `N`, `thinning`, and `save` apply here, too.

Value

The routine does not return but saves the output as a list in `bmpmp.RData`, thereby overriding the already existing file. The function `load` can be used to read any of the lists into the current workspace. As for the output of `bmpmp.estimation`, it consists of the following entries.

`mcmc`	A list, whose i-th entry is a list containing all values of the quantities in `initial.RData` for the i-th iteration (after thinning) of the entire MCMC algorithm.
`eta`	A list, whose i-th entry is a list with two entries. In case of several countries, the first entry is a list with $T - k$ entries, whose t-th entry is a matrix of the linear predictors for all male populations at t-th calendar year realised at the i-th iteration (after thinning) of the entire MCMC algorithm, where the rows indicate the countries as given in `countries` (with the last row being the reference population) and the columns indicate the ages as given in `age_levels`. In case of a single country, the first entry is a list with $T - k$ entries, whose t-th entry is a vector of the linear predictors for the male population at t-th calendar year realised at the i-th iteration (after thinning) of the entire MCMC algorithm, where the entries indicate the ages as given in `age_levels`. The second entry of i-th entry of `eta` is the corresponding list of all female linear predictors, arranged in the analogous way.
`N_thinned`	The number of iterations after thinning for the entire MCMC algorithm.

Additional output files are the intermediate results for each increment of `save` iterations for the new realisations, which are automatically contained in `bmpmp.RData`. Their lists are set up as in the case of `bmpmp.RData`, where the entries are called `mcmc_thinned`,

eta_thinned, number_thinned.

Examples

```
### Example for bmpmp.estimation continued

# Required package (must be installed)
library(MCMCpack) # (will also be loaded within the function)

# Set workspace first and create input_data.RData through create.data()
# and constants.RData and bmpmp.RData through bmpmp.estimation()
# Set output directory as before
output_directory <- paste(getwd(),"/output/", sep = "")

# Setup 100,000 additional realisations for the algorithm
N <- 500000 # (previous value was N = 400,000)
thinning <- 100
save <- 2000

# Run function
bmpmp.estimation.continue(N = N, thinning = thinning, save = save,
                          output_directory = output_directory)

### End example
```

F.4 The Function bmpmp.plots

Description

bmpmp.plots constructs posterior distribution, convergence, validation, and forecast plots for the output from bmpmp.estimation.

Usage

```
bmpmp.plots(burn_in = 0, year = 1980, age = 60,
            forecast_begin = 2000, forecast_end = 2050,
```

```
output_directory = paste(getwd(),"/output/", sep = ""),
distribution_plots_directory = "distribution_plots/",
convergence_plots_directory = "convergence_plots/",
validation_plots_directory = "validation_plots/",
forecast_plots_directory = "forecast_plots/",
titles = TRUE, tikz_format = FALSE)
```

Arguments

`burn_in`	Length of burn-in period, i.e. the number of initial realisations (after thinning) in the MCMC algorithm that should be discarded. Must be a scalar in $\mathbb{N}_0$ smaller than `N_thinned` in `bmpmp.RData`. Default is 0.
`year`	The fixed reference calendar year for posterior distribution and validation plots of the logit of mortality rates (linear predictors) versus age levels. Must be an entry in `calendar_years` in `input_data.RData`. Default is 1980.
`age`	The fixed age for posterior distribution, validation, and forecast plots of the logit of mortality rates (linear predictors) versus calendar years. Must be an entry in `age_levels` in `input_data.RData`. Default is 60.
`forecast_begin`	Calendar year, for which the forecast period should begin. Must be a scalar in $\mathbb{N}$ exceeding $t_{\min} + k$ but not larger than $t_{\max} + 1$, where $t_{\min}$ and $t_{\max}$ are the input parameters `t_min` and `t_max` from `create.data`, and k is the chosen lag order `k` for `bmpmp.estimation`. Default is 2000.
`forecast_end`	Calendar year, for which the forecast period should end. Must be a scalar in $\mathbb{N}$ exceeding `forecast_begin`. Default is 2050.
`output_directory`	Character object specifying the directory for both input data files `input_data.RData` created through the routine `create.data` and `bmpmp.RData` created through `bmpmp.estimation`. Must end with `/"`. Default is `paste(getwd(),"/output/", sep = "")`, i.e. a folder called `output` in the current workspace. For the current workspace simply set `output_directory = "/"`.

distribution_plots_directory	Character object specifying the directory within the output directory given through output_directory for posterior distribution plots of hyperparameters and parameters. Must end with /". Default is "distribution_plots/", i.e. a folder called distribution_plots in the output directory. For the output directory itself simply set distribution_plots_directory = "/".
convergence_plots_directory	Character object specifying the directory within the output directory given through output_directory for MCMC convergence plots of hyperparameters and parameters. Must end with /". Default is "convergence_plots/", i.e. a folder called convergence_plots in the output directory. For the output directory itself simply set convergence_plots_directory = "/".
validation_plots_directory	Character object specifying the directory within the output directory given through output_directory for internal validation plots of hyperparameters and parameters. Must end with /". Default is "validation_plots/", i.e. a folder called validation_plots in the output directory. For the output directory itself simply set validation_plots_directory = "/".
forecast_plots_directory	Character object specifying the directory within the output directory given through output_directory for forecast (external validation) plots of hyperparameters and parameters. Must end with /". Default is "forecast_plots/", i.e. a folder called forecast_plots in the output directory. For the output directory itself simply set forecast_plots_directory = "/".
titles	Logical value indicating whether main and axis titles should be included in the graphical output. Default is TRUE.
tikz_format	Logical value indicating whether the output should not be graphs stored in pdf format as by default, but files in tikz format to allow easy inclusion of graphs in LaTeX documents. Default is FALSE.

Details

The routine will construct the following plots in pdf or tikz format, stored in the respectively specified folders, for the output in bmpmp.RData and input_data.RData, constructed through the functions bmpmp.estimation and create.data, respectively.

- Posterior distribution plots: (a) Marginal time series plots for all CBD parameters, showing fancharts for the posterior realisations of the time series for all iterations after the burn-in period (after thinning), their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), and their starting values given by the corresponding ML estimates (solid red). The files are named after the corresponding parameters. (b) Plots of the logit of mortality rates (linear predictors) versus calendar years of the calibration period for all countries and genders for the fixed reference age given through `age`. Shown are fancharts for the posterior realisations of the linear predictors for all iterations after the burn-in period (after thinning), their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), their starting values given by the corresponding ML estimates (solid red), and the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality (solid green). The files' names start with `linear_predictor` and indicate the respective gender and country. (c) Plots of the logit of mortality rates (linear predictors) versus age levels of the calibration window for all countries and genders for the fixed reference calendar year given through `year`. Shown are fancharts for the posterior realisations of the linear predictors for all iterations after the burn-in period (after thinning), their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), their starting values given by the corresponding ML estimates (solid red), and the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality (solid green). The files' names start with `mortality_rate` and indicate the respective gender and country.

- MCMC convergence plots: Marginal scatterplots for all hyperparameters and parameters versus iterations after thinning (starting with initial values, i.e. the burn-in period is shown). For the covariance matrices $\Omega^{(i)}$ and the cointegration hyperparameters $\alpha^{(i)}$ and $\beta^{(i)}$, the precision matrices $\left(\Omega^{(i)}\right)^{-1}$ and the cointegration matrices $\Pi^{(i)} = \alpha^{(i)}\beta^{(i)\prime}$ are also plotted. The files are named after the name of the corresponding parameter or hyperparameter and their respective subscripts denoting the vector's or matrix's entry.

- Internal validation plots: (a) Marginal time series plots for all CBD parameters over the calibration period, showing fancharts for the marginal realisations of the VECM for each realisation of the posterior distribution for the hyperparameters

after the burn-in period (after thinning) using the prior distribution for the first k values, their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), and their starting values given by the corresponding ML estimates (solid red). The files are named after the corresponding parameters. (b) Plots of the logit of mortality rates (linear predictors) versus calendar years of the calibration period for all countries and genders for the fixed reference age given through `age`. Shown are fancharts for the simulated realisations of the linear predictors, their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), their starting values given by the corresponding ML estimates (solid red), and the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality (solid green). The files' names start with `linear_predictor` and indicate the respective gender and country. (c) Plots of the logit of mortality rates (linear predictors) versus age levels of the calibration window for all countries and genders for the fixed reference calendar year given through `year`. Shown are fancharts for the simulated realisations of the linear predictors, their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), their starting values given by the corresponding ML estimates (solid red), and the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality (solid green). The files' names start with `mortality_rate` and indicate the respective gender and country.

- Forecast (external validation) plots: (a) Marginal time series plots for all CBD parameters over the calibration and forecast period defined through `forecast_begin` and `forecast_end`, showing the initial values for the calibration period until the begin of the forecast period and, thereafter, fancharts for the marginal realisations of the VECM for each realisation of the posterior distribution for the hyperparameters after the burn-in period (after thinning), using the posterior distribution for the preceding k values, and their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting). The files are named after the corresponding parameters. (b) Plots of the logit of mortality rates (linear predictors) versus calendar years of the calibration and forecast period for all countries and genders for the fixed reference age given through `age`. Shown are fancharts for the simulated realisations of the linear predictors for the forecast period, their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted,

limiting), and the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality for the entire calibration period (solid green). The files' names start with `linear_predictor` and indicate the respective gender and country. (c) Plots of the logit of mortality rates (linear predictors) versus age levels of the calibration window for all countries and genders for the final calendar year of the forecast period given through `forecast_end`. Shown are fancharts for the simulated realisations of the linear predictors in solid black, their pointwise 90%, 95%, 99%, and 100% credibility intervals (solid, dashed, dotted, limiting), and the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality for the last calendar year of the calibration period as comparison (solid green). The files' names start with `mortality_rate` and indicate the respective gender and country.

The function `bmpmp.plots` requires the `fanplot` package and, if `tikzformat = TRUE`, the `tikzDevice` package. While the routine is running, status messages are provided. The routine automatically sets seeds for reproducible output. If `bmpmp.RData` has been initialised with the choice of $N = 0$, no plots can be created, and an error message will be provided.

Value

The routine does not return any values, but saves graphical output in `pdf` or `tikz` format as outlined under 'Details'.

Examples

```
### Example for Big Five as in Section 4.1
### (Example for bmpmp.estimation continued)

# Set workspace first and create input_data.RData through create.data()
# and bmpmp.RData through bmpmp.estimation()
# Set output directory as before
output_directory <- paste(getwd(),"/output/", sep = "")

# Set directories for output plots within the output directory
# (folders must exist)
```

```
distribution_plots_directory = "distribution_plots/"
convergence_plots_directory = "convergence_plots/"
validation_plots_directory = "validation_plots/"
forecast_plots_directory = "forecast_plots/"

# Setup for output graphs
burn_in <- 7500 # (after thinning, i.e. 750000 iterations in total)
year <- 1980
age <- 60
forecast_begin <- 1996
forecast_end <- 2014
titles <- TRUE
tikz_format <- FALSE

# Run function
bmpmp.plots(burn_in = burn_in, year = year, age = age,
            forecast_begin = forecast_begin,
            forecast_end = forecast_end,
            output_directory = output_directory,
            distribution_plots_directory = distribution_plots_directory,
            convergence_plots_directory = convergence_plots_directory,
            validation_plots_directory = validation_plots_directory,
            forecast_plots_directory = forecast_plots_directory,
            titles = titles, tikz_format = tikz_format)

### End example
```

F.5 The Function `ml.estimation.plots`

Description

`ml.estimation.plots` estimates the CBD model and VECM equations in the BMPMP model via ML (if not over-parametrised), using input data files created by `create.data` and `bmpmp.estimation`, and constructs forecast plots.

Usage

```
ml.estimation.plots(forecast_begin = 2000, forecast_end = 2050,
                    age = 60, output_directory
                    = paste(getwd(),"/output/", sep = ""),
                    ml_forecast_plots_directory = "ml_forecast_plots/",
                    titles = TRUE, tikz_format = FALSE)
```

Arguments

`forecast_begin`	Calendar year, for which the forecast period should begin. Must be a scalar in $\mathbb{N}$ exceeding $t_{\min} + k$ but not larger than $t_{\max} + 1$, where $t_{\min}$ and $t_{\max}$ are the input parameters `t_min` and `t_max` for `create.data` and k is the chosen lag order `k` for `bmpmp.estimation`. Default is 2000.
`forecast_end`	Calendar year, for which the forecast period should end. Must be a scalar in $\mathbb{N}$ exceeding `forecast_begin`. Default is 2050.
`age`	The fixed age for the forecast plots of the logit of mortality rates (linear predictors) versus calendar years. Must be an entry in `age_levels` in `input_data.RData`. Default is 60.
`output_directory`	Character object specifying the directory for the input data files `input_data.RData`, `constants.RData` and `initial.RData`. Must end with `/"`. Default is `paste(getwd(),"/output/", sep = "")`, i.e. a folder called `output` in the current workspace. For the current workspace simply set `output_directory = "/"`.
`ml_forecast_plots_directory`	Character object specifying the directory within the output directory given through `output_directory` for forecast (external validation) plots of hyperparameters and parameters. Must end with `/"`. Default is `"ml_forecast_plots/"`, i.e. a folder called `ml_forecast_plots` in the output directory. For the `output` directory itself simply set `ml_forecast_plots_directory = "/"`.
`titles`	Logical value indicating whether main and axis titles should be included in the graphical output. Default is `TRUE`.

`tikz_format`	Logical value indicating whether the output should not be graphs stored in `pdf` format as by default, but files in `tikz` format to allow easy inclusion of graphs in LaTeX documents. Default is `FALSE`.

Details

Given the input data through `input_data.RData`, the initial values via `initial.RData`, and the constants through `constants.RData` (where the latter two files can be conveniently created using `bmpmp.estimation` with the choice of `N` = 0), the routine estimates both the CBD model and the VECM as defined in the BMPMP model for `bmpmp.estimation` via ML in a two-step procedure known from classical LC or CBD models. First, the CBD model is estimated via standard ML procedures known from Binomial generalised regression (i.e. the starting values for the Bayesian estimation). In a second step, the VECM is estimated via the ML techniques outlined in Appendix C.4. Note that ML estimation will not be available for many applications, as the number of hyperparameters in the VECM soon exceeds the number of latent parameters in the CBD model. In particular, only a very small number of countries will allow for a lag order k greater than one. In case of singularities in the ML procedure, respective error messages will be provided.

`ml.forecast.plots` further provides the following forecast (external validation) plots in `pdf` or `tikz` format, stored in the folder given by `ml_forecast_plots_directory`: (a) Marginal time series plots for all CBD parameters over the calibration and forecast period defined through `forecast_begin` and `forecast_end`, showing the ML point estimates for the calibration period until the begin of the forecast period and, thereafter, the ML forecasts of the VECM consisting of the pointwise estimates (solid black), and the pointwise 2.5% and 97.5% confidence limits (dashed cyan). The files are named after the corresponding parameters. (b) Plots of the logit of mortality rates (linear predictors) versus calendar years of the calibration and forecast period for all countries and genders for the fixed reference age given through `age`. Shown are the ML point estimates for the calibration period until the begin of the forecast period and, thereafter, the ML forecasts consisting of the pointwise estimates (solid black), and the pointwise 2.5% and 97.5% confidence limits (dashed cyan) along with the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality for the entire calibration period (solid green). The files' names start with `linear_predictor` and indicate the

respective gender and country. (c) Plots of the logit of mortality rates (linear predictors) versus age levels of the calibration window for all countries and genders for the final calendar year of the forecast period given through `forecast_end`. Shown are the ML forecast point estimates of the linear predictors in solid black and their pointwise 2.5% and 97.5% confidence limits (dashed cyan) along with the logits of the observed mortality rates under the assumption of piecewise constant forces of mortality for the last calendar year of the calibration period as comparison (solid green). The files' names start with `mortality_rate` and indicate the respective gender and country.

The function `ml.forecast.plots` requires the `expm` package and, if `tikzformat = TRUE`, the `tikzDevice` package. While the routine is running, status messages are provided.

Value

The routine does not return any values, but saves graphical output in `pdf` or `tikz` format as outlined under 'Details'.

Examples

```
### Example for Big Five as in Section 4.1
### (Example for bmpmp.estimation continued)

# Required package (must be installed)
library(expm) # (will also be loaded within the function)

# Set workspace first and create input_data.RData through create.data()
# and constants.RData
# and bmpmp.RData through bmpmp.estimation() (e.g. with N=0)
# Set output directory as before
output_directory <- paste(getwd(),"/output/", sep = "")

# Set directory for output graphs (folder must exist)
ml_forecast_plots_directory <- "ml_forecast_plots/"
```

```
# Setup for output graphs with forecast period as before
age <- 60
forecast_begin <- 1996
forecast_end <- 2014
titles <- TRUE
tikz_format <- FALSE

# Run function
ml.estimation.plots(forecast_begin = forecast_begin,
                    forecast_end = forecast_end, age = age,
                    output_directory = output_directory,
                    ml_forecast_plots_directory
                    = ml_forecast_plots_directory,
                    titles = titles, tikz_format = tikz_format)

### End example
```

References

Agostinelli, C. and Greco, L. (2012). Weighted likelihood in Bayesian inference. In *Proceedings of 46th Scientific Meeting of the Italian Statistical Society*, Sapienza University of Rome, Rome, Italy.

Ahčan, A., Medved, D., Olivieri, A., and Pitacco, E. (2014). Forecasting mortality for small populations by mixing mortality data. *Insurance: Mathematics and Economics*, 54:12–27.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alho, J. M. (2000). Discussion of Lee (2000). *North American Actuarial Journal*, 4(1):91–93.

Bauwens, L. and Lubrano, M. (1993). Identification restrictions and posterior densities in cointegrated Gaussian VAR systems. In Fomby, T. and Hill, R. C. (Eds.), *Advances in Econometrics*, Vol. 11b. JAI Press, Greenwich, Connecticut.

Biatat, V. D. and Currie, I. D. (2010). Joint models for classification and comparison of mortality in different countries. In *Proceedings of 25rd International Workshop on Statistical Modelling*, 89–94, University of Glasgow, Glasgow, Scotland.

Booth, H., Hyndman, R. J., Tickle, L., and De Jong, P. (2006). Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demographic Research*, 15(9):289–310.

Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, 3(1–2):3–43.

Börger, M. and Aleksic, M.-C. (2011). Coherent projections of age, period, and cohort dependent mortality improvements. *Preprint Series, Faculty of Mathematics and Economics, University of Ulm, Ulm, Germany*.

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (2013). *Time Series Analysis: Forecasting and Control, Fourth Edition*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey.

Box, G. E. and Tiao, G. C. (2011). *Bayesian Inference in Statistical Analysis*, Vol. 40 of *Wiley Classics Library*. John Wiley & Sons, Hoboken, New Jersey.

Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(2):151–164.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17(31):334–355.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, 42:693–734.

Brockwell, P. J. and Davis, R. A. (2009). *Time Series: Theory and Methods, Second Edition*. Springer Series in Statistics. Springer, New York, New York.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373–393.

Brüggemann, R. (2004). *Model Reduction Methods for Vector Autoregressive Processes*, Vol. 536 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin/Heidelberg, Germany.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J., Blake, D., and Dowd, K. (2008). Modelling and management of mortality risk: A review. *Scandinavian Actuarial Journal*, 2008(2–3):79–113.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2011a). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355–367.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011b). Bayesian stochastic mortality modelling for two populations. *Astin Bulletin*, 41(1):29–59.

Chao, J. C. and Phillips, P. C. (1999). Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics*, 91(2):227–271.

Czado, C., Delwarde, A., and Denuit, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, 36(3):260–284.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

de Jong, P. and Tickle, L. (2006). Extending Lee–Carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18.

de Moivre, A. (1725). *Annuities upon Lives: Or, the Valuation of Annuities upon any Number of Lives, as also, of Reversions to which is added, an Appendix Concerning the Expectations of Life, and Probabilities of Survivorship.* Published by Pearson, W., London, England.

DeJong, D. N. (1992). Co-integration and trend-stationarity in macroeconomic time series: Evidence from the likelihood function. *Journal of Econometrics*, 52(3):347–370.

Delwarde, A., Denuit, M., and Partrat, C. (2007). Negative Binomial version of the Lee–Carter model for mortality forecasting. *Applied Stochastic Models in Business and Industry*, 23(5):385–401.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models, Third Edition*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Florida.

Dowd, K., Cairns, A. J., Blake, D., Coughlan, G. D., and Khalaf-Allah, M. (2011). A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2):334–356.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Florida.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Florida.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (Eds.), *Proceedings of the Fourth Valencia International Meeting*, Vol. 4 of *Bayesian Statistics*, 169–193. Oxford University Press, Oxford, England.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75(1):121–146.

Gilks, W. R. (2005). Markov chain Monte Carlo. In Armitage, P. and Colton, T. (Eds.), *Encyclopedia of Biostatistics, Second Edition*. John Wiley & Sons, Chichester, England.

Girosi, F. and King, G. (2008). *Demographic Forecasting*. Princeton University Press, Princeton, New Jersey.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, 46(6):1303–1310.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121–130.

Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195.

Hansen, P. R. (2005). Granger's Representation Theorem: A closed-form expression for $I(1)$ processes. *The Econometrics Journal*, 8(1):23–38.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Human Mortality Database (2014). University of California, Berkeley, California, and Max Planck Institute for Demographic Research, Rostock, Germany. Available at `http://www.mortality.org` or `http://www.humanmortality.de` (data downloaded on 24 September 2014).

James, A. T. (1954). Normal multivariate analysis and the orthogonal group. *The Annals of Mathematical Statistics*, 25(1):40–75.

James, A. T. (1964). Distributions of matrix variates and latent roots derived from normal samples. *The Annals of Mathematical Statistics*, 35(2):475–501.

Jarner, S. F. and Kryger, E. M. (2011). Modelling adult mortality in small populations: The SAINT model. *Astin Bulletin*, 41(2):377–418.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 55(2):163–172.

Jeffreys, H. (1998). *The Theory of Probability, Third Edition*. Oxford University Press, Oxford, England.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Advanced Texts in Econometrics. Oxford University Press, Oxford, England.

Johansen, S. and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration – with applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52(2):169–210.

Johansen, S. and Juselius, K. (1992). Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of Econometrics*, 53(1):211–244.

Kleibergen, F. and van Dijk, H. K. (1994). On the shape of the likelihood/posterior in cointegration models. *Econometric Theory*, 10(3–4):514–551.

Kogure, A., Kitsukawa, K., and Kurachi, Y. (2009). A Bayesian comparison of models for changing mortalities toward evaluating longevity risk in Japan. *Asia-Pacific Journal of Risk and Insurance*, 3(2):1–21.

Koop, G., Strachan, R. W., Van Dijk, H., and Villani, M. (2006). Bayesian approaches to cointegration. In Mills, T. C. and Patterson, K. D. (Eds.), *Econometric Theory*, Vol. 1 of *Palgrave Handbook on Econometrics*, 871–898. Palgrave Macmillan, New York, New York.

Lee, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, 4(1):80–91.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Li, J. S.-H. and Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, 15(2):177–200.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594.

Lin, Y., Liu, S., and Yu, J. (2013). Pricing mortality securities with correlated mortality indexes. *Journal of Risk and Insurance*, 80(4):921–948.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions – five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38.

Lomnicki, Z. (1961). Tests for departure from normality in the case of linear stochastic processes. *Metrika*, 4(1):37–62.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin/Heidelberg, Germany.

Macdonald, A., Cairns, A., Gwilt, P., and Miller, K. (1998). An international comparison of recent trends in population mortality. *British Actuarial Journal*, 4(1):3–141.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability, Second Edition*. Cambridge University Press, Cambridge, England.

Neudecker, H. (1968). The Kronecker matrix product and some of its applications in econometrics. *Statistica Neerlandica*, 22(1):69–82.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48.

Ntamjokouen, A., Haberman, S., and Consigli, G. (2014). A multivariate approach to project the long run relationship between mortality indices for Canadian provinces. In Perna, C. and Sibillo, M. (Eds.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, 153–161. Springer International Publishing, Cham, Switzerland.

Oeppen, J. and Vaupel, J. W. (2002). Broken limits to life expectancy. *Science*, 296(5570):1029–1031.

Pedroza, C. (2006). A Bayesian forecasting model: Predicting US male mortality. *Biostatistics*, 7(4):530–550.

Pitacco, E., Denuit, M., Haberman, S., and Olivieri, A. (2009). *Modelling Longevity Dynamics for Pensions and Annuity Business*. Oxford University Press, Oxford, England.

Plat, R. (2009). Stochastic portfolio specific mortality and the quantification of mortality basis risk. *Insurance: Mathematics and Economics*, 45(1):123–132.

Raftery, A. E., Chunn, J. L., Gerland, P., and Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50(3):777–801.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (Eds.), *Proceedings of the Fourth Valencia International Meeting*, Vol. 4 of *Bayesian Statistics*, 763–773. Oxford University Press, Oxford, England.

Reichmuth, W. H. and Sarferaz, S. (2008). Bayesian demographic modeling and forecasting: An application to US mortality. *Discussion Paper of Sonderforschungsbereich 649 (Economic Risk), Humboldt-Universität zu Berlin, Berlin, Germany*.

Reinsel, G. C. (2003). *Elements of Multivariate Time Series Analysis, Second Edition*. Springer, New York, New York.

Renshaw, A. E. and Haberman, S. (2003a). Lee–Carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):119–137.

Renshaw, A. E. and Haberman, S. (2003b). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.

Renshaw, A. E. and Haberman, S. (2003c). On the forecasting of mortality reduction factors. *Insurance: Mathematics and Economics*, 32(3):379–401.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods, Second Edition*. Springer Texts in Statistics. Springer, New York, New York.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4):1151–1172.

Salmon, M. (1982). Error correction mechanisms. *The Economic Journal*, 92(367):615–629.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, Massachusetts.

Strachan, R. W. (2003). Valid Bayesian estimation of the cointegrating error correction model. *Journal of Business & Economic Statistics*, 21(1):185–195.

Strachan, R. W. and Dijk, H. K. (2004). Valuing structure, model uncertainty and model averaging in vector autoregressive processes. *Econometric Institute Report, Erasmus University Rotterdam, Rotterdam, Netherlands*.

Strachan, R. W. and Inder, B. (2004). Bayesian analysis of the error correction model. *Journal of Econometrics*, 123(2):307–325.

Tuljapurkar, S., Li, N., and Boe, C. (2000). A universal pattern of mortality decline in the G7 countries. *Nature*, 405(6788):789–792.

Villani, M. (2000). *Aspects of Bayesian Cointegration*. PhD thesis, Stockholm University.

Villani, M. (2001). Fractional Bayesian lag length inference in multivariate autoregressive processes. *Journal of Time Series Analysis*, 22(1):67–86.

Villani, M. (2005). Bayesian reference analysis of cointegration. *Econometric Theory*, 21(2):326–357.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.

Warne, A. (2006). Bayesian inference in cointegrated VAR models with applications to the demand for euro area M3. *Working Paper Series of the European Central Bank*, 692:1–41.

Wilmoth, J. R. (1993). Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. Technical report, Department of Demography, University of California, Berkeley, California.

Wilson, C. (2001). On the scale of global demographic convergence 1950–2000. *Population and Development Review*, 27(1):155–171.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 32–52.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksells, Stockholm, Sweden.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, New York, New York.

Zhou, R., Li, J. S.-H., and Tan, K. S. (2013). Pricing standardized mortality securitizations: A two-population model with transitory jump effects. *Journal of Risk and Insurance*, 80(3):733–774.

Zhou, R., Wang, Y., Kaufhold, K., Li, J., and Tan, K. (2012). Modeling mortality of multiple populations with vector error correction models: Applications to Solvency II. *Submitted for publication.*

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ich bin mir bewusst, dass eine unwahre Erklärung rechtliche Folgen haben wird.

Ulm, den 17. Dezember 2014 _____

(Unterschrift)