Bayesian Hierarchical Models and Moment Bounds for High-Dimensional Time Series

Yicheng Li

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Adrian E. Raftery, Chair

Fang Han, Chair

Yanqin Fan

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bayesian Hierarchical Models and Moment Bounds for High-Dimensional Time Series

Yicheng Li

Co-Chairs of the Supervisory Committee:
Adrian E. Raftery

Fang Han

In this dissertation, I explore two statistical tasks involving high-dimensional time series. The first task is to forecast high-dimensional time series using Bayesian hierarchical models (BHM). The data under modeling is related to smoking epidemic and human mortality measures obtained from multiple populations around the world. I propose a BHM for estimating and forecasting the all-age smoking attributable fraction (ASAF), which serves as a summarizing statistical measure of the effect of smoking on mortality. The projected ASAF is used to forecast the dynamics of the between-gender gap of life expectancy at birth. In addition, I propose a general framework to incorporate smoking-related information into life expectancy at birth forecast. The framework includes forecasting an age-specific smoking attributable fraction (ASSAF), a non-smoking life expectancy at birth, and a male-female life expectancy gap. Assessed by out-of-sample validation, the new framework improves forecast accuracy and calibration compared with other commonly considered methods for mortality forecasts.

The second task is to obtain expectation bounds for the deviation of large sample autocovariance matrices from their means under weak data dependence. While the accuracy of covariance matrix estimation corresponding to independent data has been well understood, much less is known in the case of dependent data. We make a step towards filling this gap, and establish deviation bounds that depend only on the parameters controlling the "intrinsic dimension" of the data up to some logarithmic terms. Our results have immediate impacts

on high dimensional time series analysis, and we apply them to the high dimensional linear VAR(d) model, the vector-valued ARCH model, and a model used in Banna et al. (2016).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

APC: Age-Period-Cohort analysis.

ARCH: autoregressive conditional heteroskedasticity model.

ASAF: all-age smoking attributable fraction.

ASSAF: age-specific smoking attributable fraction.

BHM: Bayesian hierarchical model.

COPD: chronic obstructive pulmonary disease.

CPS: American Cancer Society Cancer Prevention Study

CRPS: continuous ranked probability score.

MAE: mean absolute error.

MCMC: Markov chain Monte Carlo.

OECD: Organisation for Economic Co-operation and Development

SAF: smoking attributable fraction.

UN: United Nations.

US: United States of America.

VAR: vector autoregression.

WPP: World Population Prospects

ACKNOWLEDGMENTS

First of all, I would like to thank Adrian E. Raftery and Fang Han for their mentorship. I came to this department with very little knowledge about how to conduct research until I started to work with them. I would like to thank them for their patience in mentoring me, especially when I encountered hardship. Their enthusiasm for research inspired me to finish this dissertation. I would also like to thank Hana Sevcikova for her helpful discussion on questions of my research and life. Moreover, I would like to thank Yanqin Fan and Haidong Wang for serving on my supervisory committee and offering their advice.

I would also like to thank the Department of Statistics at the University of Washington for their generous financial support and the embracing environment. Specifically, I would like to thank Ellen Reynolds, Eileen Heimer, Jasmine Wang, Tracy Pham, and Kristine Chan for their warmhearted help during my studying here.

Lastly, I am grateful to my parents, Jun Li and Yanjie Bi, for supporting me over these years. I would also like to thank Peter for standing by me and encouraging me to overcome many hardships in this journey.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Motivation

Estimation and inference on high-dimensional time series data are common tasks in modern statistics, and they exhibit large differences in various aspects from their independent data counterparts. This dissertation is motivated by two different prospects of high-dimensional time series data. The first aspect, motivated by a task in statistical demography, is to forecast multi-population smoking-related mortality measures through a Bayesian hierarchical model. The other one, motivated by commonly used time series models under high-dimensional setting, is to derive expectation bounds for the deviation of large sample autocovariance matrices from their means under weak dependence.

Estimating mortality-related measures is one of the main components of human population projection. With the improvements in food supply, medical care, and the general living environment, the human mortality rate has been steadily decreasing since the last century. However, unhealthy lifestyles associated with modernization such as smoking, alcohol consumption, and obesity have become major risk factors of cardiovascular diseases, malignant cancers, diabetes, and many others. Among these risk factors, smoking is the leading preventable cause of death. Globally, tobacco use causes approximately 6 million deaths per year. In the United States, tobacco use kills more than 480,000 on average per year.

The onset of the smoking epidemic could be traced back to the mid-19th century when the cigarette industry in industrialized regions started to sprout. Over the past century, the dynamics of the smoking epidemic have shares similar trends with variations in duration, magnitude, and velocity of development from population to populations. As described in Pampel (2005), the smoking epidemic is a diffusion process from male to female population,

and from developed to developing regions. The common trend of smoking prevalence, as observed more completely in male populations of the developed world, shows an increasing-peaking-decreasing pattern, where the turning point happened around the mid-20th century when adverse effects of smoking were realized by the public and anti-smoking movements began to thrive. Female populations from most developed regions have also experienced a decline of the smoking epidemic but usually one or two decades later than that of the male population. In the developing world, the smoking epidemic started later and some countries such as China and India are still experiencing the stage of increasing or leveling. Female smoking epidemics in developing regions remain at a low level in general since female smoking is not encouraged in most of these areas due to gender disparity. Therefore, estimating and forecasting mortality attributable to smoking is important for monitoring and controlling the current and future level of smoking effect on public health, which motivates the work presented in Chapter 2.

The smoking effect is mainly responsible for several noticeable characteristics in the dynamics of mortality change over the past century. First of all, smoking is responsible for the non-linear decline in mortality rate and non-linear increase in life expectancy. Janssen et al. (2013) argued that as the smoking effect is removed, mortality decline becomes more linear. Bongaarts (2006) estimated the life expectancy at birth after removing the smoking effect, which appears to be more linear than observed life expectancy at birth. Secondly, smoking accounts largely for geological differences in mortality measures. For example, a regional disadvantage in mortality of the population from the southern United States compared with those from other regions has shown to be largely due to smoking according to Fenelon and Preston (2012). Last but not least, smoking accounts largely for the life expectancy gap between males and females. During the last century, female populations from almost all countries have a higher life expectancy at birth than those of the male populations, but a shrinking between-gender gap of life expectancy has been observed among developed countries starting from the last few decades, which has been shown to be associated with the shrinking between-gender smoking effects (cf. Preston and Wang (2006), Pampel (2005)).

Therefore, incorporating smoking-related information in mortality modeling will not only help to understand the role of the smoking epidemic on mortality measures, but also gain more accuracy and confidence in future mortality forecast. Chapter 3 focuses mainly on building up such a model, inspired by Bongaarts (2006) and Janssen et al. (2013), to make better life expectancy projections with the assistance of smoking-related data.

On the other hand, motivated from a theoretical aspect, the second main topic of this dissertation studies the autocovariance matrix estimation of a class of high-dimensional time series models. Concentration inequalities and moment bounds for high-dimensional covariance matrix estimation based on independent samples are well-established in the literature. However, such inequalities are not available until recently for many commonly used high-dimensional time series models such as vector autoregressive model of lag d (VAR(d)) and vector-valued autoregressive conditionally heteroscedastic (ARCH) model, and existing literature often requires special structures on autocovariance matrices. Deriving optimal tail probability bound and moment bound for autocovariance matrix estimation under a more general class of high-dimensional time series models motivates Chapter 4 in this dissertation. The considered class is closely related to a weakly dependent data structure called the τ-mixing process proposed by Dedecker and Prieur (2005), which includes the previously mentioned models and many others such as Bernoulli shifts, contracting Markov Chains, and so on. The result heavily depends on a brand new Bernstein-type concentration inequality for the sum of a sequence of τ-mixing random matrices, which is mainly inspired by Merlevède et al. (2009) and Banna et al. (2016).

1.2 Background

We now provide some brief background knowledge on mortality forecasts and non-asymptotic theory of covariance estimation. This section is not meant to be a comprehensive literature review but tries to provide information mostly related to this dissertation and references on general readings for interested readers. Additional background for each subtopics can be found in the Introduction section at the beginning of each chapter (Sections 2.1, 3.1, and

4.1).

1.2.1 Mortality Estimation and Projection

Quantitative mortality forecasts methods have thrived in recent decades, largely required by government and insurance companies due to the continuously population aging. In contrast to qualitative methods which largely depend on experts' opinions, modern mortality forecasts methods, which are extrapolative in nature, are believed to be more objective, reliable, and applicable. The most well-known class of mortality forecasts is the Lee-Carter-type method, which is originated from Lee and Carter (1992). The original Lee-Carter method decomposes the logarithmic transformed age-specific mortality rate of $d_{x,t}$ by three components

$$\log(d_{x,t}) = a_x + b_x \times k_t + \varepsilon_{x,t},$$

where a_x is average log-transformed mortality at age x, b_x evaluates the responding change of overall level of mortality over time at age x, k_t measures the overall level of mortality at time t, and $\varepsilon_{x,t}$ is standard Gaussian error. For forecasts, Lee-Carter method extrapolates k_t linearly based on the historical data. Although the Lee-Carter model suffers from a severe underestimation of prediction variance and heavily depends on the linearity assumption of the time trend, it is often considered as a benchmark of mortality forecasts for comparison.

To overcome these shortcomings, many variants of the Lee-Carter method are developed. Lee and Miller (2001) proposed a procedure to estimate the time effect so that the generated life expectancy estimates match the observed values. Booth et al. (2002) suggested searching for a fitting period, in which the linear assumption holds. Other variations of the Lee-Carter method include adding a cohort effect (Renshaw and Haberman, 2006; Currie et al., 2004; Plat, 2009), applying functional data approach (De Jong and Tickle, 2006; Hyndman and Ullah, 2007; Shang, 2016), and incorporating biomedical information (Janssen et al., 2013; Stoeldraijer et al., 2015; Vidra et al., 2017; Trias Llimós and Janssen, 2019). For a more comprehensive summary and comparison among variants of Lee-Carter-type methods, see Booth et al. (2006), Booth and Tickle (2008), and Janssen (2018).

Probabilistic forecast under the Bayesian framework has been a recent advance. Bayesian methods naturally incorporate variability in the observed data into the forecast. On the one hand, several Bayesian Lee-Carter methods are developed. Wiśniowski et al. (2015) introduced a unified Bayesian framework of Lee-Carter-type modeling for mortality, fertility, and migration. King and Soneji (2011) suggested a Bayesian linear model incorporating two health risk factors–smoking and obesity. Pedroza (2006) and Fung et al. (2017) approached the mortality dynamics by rewriting the original model into a Bayesian state-space framework. On the other hand, Raftery et al. (2013) proposed a Bayesian hierarchical model (BHM) on life expectancy directly, by modeling its non-linear growth. This method is currently adopted by the United Nations Population Division's *World Population Prospects* (WPP). Godwin and Raftery (2017) modified the BHM method by incorporating HIV epidemics related data, which improved the forecast performance for HIV-epidemic countries. As Janssen (2018) commented, mortality forecasts could be more accurate and interpretable with the help of extra epidemiology information compared to pure extrapolative methods. Chapter 3 builds a new method for mortality forecast based on Raftery et al. (2013) by considering smoking epidemics.

1.2.2 Dependent Data Framework

This section will introduce the framework of dependent data considered in Chapter 4 of this dissertation. Mixing measure, which quantifies the dependence among random variables, is defined on their corresponding σ-algebras. Consider an absolute probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for any sub-$\sigma$-algebras $\mathcal{U},\ \mathcal{V} \subset \mathcal{F}$, two well-studied strong mixing measures, α-mixing measure and β-mixing measure, which were introduced by Rosenblatt (1956) and Volkonskii and Rozanov (1959) respectively, are defined as

$$\alpha(\mathcal{U}, \mathcal{V}) := \sup_{U \in \mathcal{U}, V \in \mathcal{V}} |\mathbb{P}(U \cap V) - \mathbb{P}(U)\mathbb{P}(V)|,$$

$$\beta(\mathcal{U}, \mathcal{V}) := \frac{1}{2} \sup_{I, J \geq 1, \{U_i\}_{i=1}^{I}, \{V_j\}_{j=1}^{J}} \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(U_i \cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j)|,$$

where the supremum in the definition of $\beta(\mathcal{U}, \mathcal{V})$ is taken over all measurable partitions $\{U_i\}_{i=1}^I, \{V_j\}_{j=1}^J$ of Ω. It can be shown that $0 \leq \alpha(\mathcal{U}, \mathcal{V}) \leq \beta(\mathcal{U}, \mathcal{V}) \leq 1$ for all $\mathcal{U}, \mathcal{V} \subset \mathcal{F}$.

Consider a random process, i.e., $\{X_i\}_{i\in\mathbb{Z}}$, where $\mathbb{Z}$ is the set of integers. Then the corresponding mixing measures defined on the random process depend on how apart two subsequences are, i.e., for any integer $k > 0$,

$$\alpha(k; \{X_i\}_{i\in\mathbb{Z}}) := \sup_{j\in\mathbb{Z}} \alpha(\sigma(\{X_\ell\}_{\ell\leq j}), \sigma(\{X_\ell\}_{\ell\geq j+k})),$$

$$\beta(k; \{X_i\}_{i\in\mathbb{Z}}) := \sup_{j\in\mathbb{Z}} \beta(\sigma(\{X_\ell\}_{\ell\leq j}), \sigma(\{X_\ell\}_{\ell\geq j+k})).$$

A random process is called α-mixing if $\alpha(k; \{X_i\}_{i\in\mathbb{Z}}) \to 0$ as $k \to \infty$. If $\alpha(k; \{X_i\}_{i\in\mathbb{Z}}) \leq c\alpha^k$ for some arbitrary constant c and $0 \leq \alpha < 1$ for all k, then the process is called geometric α-mixing. Such definitions could be applied to β-mixing measure similarly. Common examples of α-mixing and β-mixing include m-dependent random process, strictly stationary countable-state Markov chain, strictly stationary Markov chain with geometric ergodicity, and many others. For the interested readers, see Bradley (2005a) for a more complete survey of the theory of strong mixing conditions.

Unfortunately, strong mixing conditions are usually hard to verify or violated in many commonly used time series models, especially under high-dimensional settings (cf. Andrews (1984) and Section 1.5 of Dedecker et al. (2007)). Dedecker and Prieur (2005) introduced a class of weak dependence measure, which turns out to be more easily calculated and contains a large range of pertinent examples. τ-mixing measure is one such weak dependence measure and the one considered in Chapter 4 of this dissertation. Consider probability space $(\Omega, \mathcal{F}, \mathbb{P})$, X an L_1-integrable random variable taking value in a Polish space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, and a sub-$\sigma$-algebra $\mathcal{A} \subset \mathcal{F}$. The τ-measure of dependence between X and $\mathcal{A}$ is defined to be

$$\tau(\mathcal{A}, X; \|\cdot\|_{\mathcal{X}}) := \Big\| \sup_{g\in\Lambda(\|\cdot\|_{\mathcal{X}})} \Big\{ \int g(x)\mathbb{P}_{X|\mathcal{A}}(\mathsf{d}x) - \int g(x)\mathbb{P}_X(\mathsf{d}x) \Big\} \Big\|_{L(1)},$$

where $\mathbb{P}_X$ is the distribution of X, $\mathbb{P}_{X|\mathcal{A}}$ is the conditional distribution of X given $\mathcal{A}$, and $\Lambda(\|\cdot\|_{\mathcal{X}})$ stands for the set of 1-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ with respect to the norm $\|\cdot\|_{\mathcal{X}}$.

A nice coupling property makes τ-mixing measure incredibly useful. First, a similar coupling lemma on β-mixing measure was proved by Berbee (1979):

Lemma 1. *Let X and Y be two random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in Borel spaces B_1 and B_2 respectively. Assume that there exists a random variable U uniformly distributed over $[0,1]$, independent of $\sigma(X)$ and $\sigma(Y)$. Then there exists a random variable $\tilde{Y}$, measurable with respect to $\sigma(X) \vee \sigma(Y) \vee \sigma(U)$, independent of $\sigma(X)$ and distributed as Y, such that*

$$\beta(\sigma(X), \sigma(Y)) = \mathbb{E}(\mathbf{1}(Y \neq \tilde{Y})). \tag{1.1}$$

Lemma 1 allows one to replace one half of a dependent random sequence with an identically-distributed copy that is independent of the other half of the original sequence, and the difference introduced by the copy can be quantified by the β-mixing measure of the sequence.

The coupling lemma for τ-mixing measure is proved in Dedecker et al. (2007):

Lemma 2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{A}$ be a sigma algebra of $\mathcal{F}$, and X be a random variable with values in a Polish space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$. Assume that $\int \|x - x_0\|_{\mathcal{X}} \mathbb{P}_X(\mathsf{d}x)$ is finite for any $x_0 \in \mathcal{X}$. Assume that there exists a random variable U uniformly distributed over $[0,1]$, independent of the σ-algebra generated by X and $\mathcal{A}$. Then there exists a random variable $\tilde{X}$, measurable with respect to $\mathcal{A} \vee \sigma(X) \vee \sigma(U)$, independent of $\mathcal{A}$ and distributed as X, such that*

$$\tau(\mathcal{A}, X; \|\cdot\|_{\mathcal{X}}) = \mathbb{E}\|X - \tilde{X}\|_{\mathcal{X}}. \tag{1.2}$$

Notice that the coupling property of τ-mixing measures differs from that of β-mixing mainly by changing the distance function $d(x, \tilde{x}) = \mathbf{1}(x \neq \tilde{x})$ to $d(x, \tilde{x}) = |x - \tilde{x}|$. Dedecker et al. (2007) provides a more comprehensive survey of the theory and applications of these weak dependent measures.

1.2.3 Non-asymptotic Theory of Covariance Estimation

In high-dimensional settings, the sample autocovariance matrix $\hat{\mathbf{\Sigma}}_m := (n-m)^{-1}\sum_{i=1}^{n-m}\boldsymbol{Y}_i\boldsymbol{Y}_{i+m}^{\mathsf{T}}$ for $0 \leq m \leq n-1$ based on a random sample size of n may not be a consistent estimator for the population autocovariance matrix $\mathbf{\Sigma}_m$ of a random vector $\boldsymbol{Y} \in \mathbb{R}^p$ when $p > n$. Vershynin (2012) first introduced the "effective rank" of a matrix $\mathbf{\Sigma}$

$$r(\mathbf{\Sigma}) := \frac{\mathrm{tr}(\mathbf{\Sigma})}{\|\mathbf{\Sigma}\|},$$

as the "intrinsic dimension" of the data to quantify the minimal sample size required for sample autocovariance matrix to be consistent and to achieve the optimal rate of convergence of $\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|$ under the i.i.d and bounded support assumption. Lounici (2014) and Bunea and Xiao (2015) established a similar rate optimal bound for sample covariance matrix estimation based on a sample of i.i.d, subgaussian random vectors:

$$\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\| \leq C\|\mathbf{\Sigma}_0\|\Big\{\sqrt{\frac{r(\mathbf{\Sigma}_0)\log(ep)}{n}} + \frac{r(\mathbf{\Sigma}_0)\log(ep)}{n}\Big\} \tag{1.3}$$

for some arbitrary constant $C > 0$. The key component in proving such results is to derive a corresponding Bernstein-type large deviation inequality for a sum of random matrices of interest. Vershynin (2012) and Lounici (2014) proved different versions of Bernstein's inequality under some boundedness assumptions while Bunea and Xiao (2015) proved one for unbounded matrices.

Deriving rate optimal expectation bounds for the deviations of high-dimensional sample autocovariance matrices from their means under dependence is more challenging. The major issue is to derive a similar Bernstein-type inequality for a sequence of dependent random matrices. Merlevède et al. (2009) and Merlevède et al. (2011) first derived the Bernstein-type inequalities for one-dimensional α-mixing and τ-mixing random processes by introducing a new decoupling technique using a Cantor-like partition of the sequence. Banna et al. (2016) extended the result to matrix settings under β-mixing by carefully applying Berbee's coupling Lemma 1. Motivated by the fact that the τ-mixing measure possesses similar nice coupling properties and includes a larger verifiable class of applications, Theorem 10 in Chapter 4

further extends the Bernstein-type inequality to a sequence of τ-mixing random matrices with bounded spectral norm. This Bernstein-type inequality provides a base for deriving rate optimal moment bounds for high-dimensional sample autocovariance matrix estimation under dependence in Chapter 4 of this dissertation.

Recently, in a remarkable series of papers (Koltchinskii and Lounici, 2017a,b,c), Koltchinskii and Lounici showed that, for subgaussian independent data, the extra multiplicative p term in Inequality 1.3 could be further removed. The proof rests on Talagrand's majorizing measures (Talagrand, 2014) and a corresponding maximal inequality due to Mendelson (Mendelson, 2010). In the most general cases, it is still unknown if Talagrand's approach could be extended to weakly dependent data, but it motivates Theorem 4 in Chapter 4 to remove all logarithm factors in Inequality (1.3) under a Gaussian random process.

1.3 Outline of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 and 3 focus on the Bayesian hierarchical modeling of high-dimensional time series data on human mortality data. Chapter 2 presents a method for projecting all-age smoking attributable fraction (ASAF) for over 60 countries using a Bayesian hierarchical model. The result of this chapter is used for modeling male-female life expectancy gap in Chapter 3. Chapter 3 focuses on forecasting life expectancy at birth with the smoking effect incorporated for both genders and over 60 countries simultaneously. Chapter 4 switches gears to provide the moment bounds of high-dimensional autocovariance matrices estimation under weak dependence with applications to some commonly used models in econometrics.

Chapter 2

ESTIMATING AND FORECASTING THE SMOKING-ATTRIBUTABLE MORTALITY FRACTION FOR BOTH GENDERS JOINTLY IN OVER 60 COUNTRIES

2.1 Introduction

Smoking is known to have adverse impacts on health and is one of the leading preventable causes of death (Peto et al., 1992; Bongaarts, 2014; Mons and Brenner, 2017). It is a major risk factor for lung cancer, chronic obstructive pulmonary disease (COPD), respiratory diseases, and vascular diseases, and tobacco use causes approximately 6 million deaths per year (Britton, 2017). For instance, tobacco use causes more than 480,000 deaths per year in the United States, accounting for about 20% of the total deaths of US adults, even though smoking prevalence in United States has declined from 42% in the 1960s to 14% in 2018 (Mons and Brenner, 2017).

The smoking attributable fraction (SAF) is the proportion by which mortality would be reduced if the population were not exposed to smoking. It is defined as

$$\text{SAF} = \frac{n_S}{n_D},$$

where n_S is the number of smokers who died because of their smoking habit and n_D is the total number of people who died. It can be shown that this is equivalent to

$$\text{SAF} = \frac{p(r-1)}{p(r-1)+1}, \tag{2.1}$$

where p is the underlying prevalence of smoking in the population and r is the risk of dying of smokers divided by the risk of dying of nonsmokers in the population (Rosen, 2013).

Estimating and forecasting the SAF of mortality is essential for assessing how the smoking epidemic influences mortality measures from the past to the future. First of all, nonlinear

patterns of increase in life expectancy over time are partially due to the smoking epidemic. Bongaarts (2006) used the SAF to calculate the non-smoking life expectancy, which turned out to evolve in a more linear fashion than overall life expectancy (including smoking effects). Janssen et al. (2013) used a similar technique to calculate the non-smoking attributable mortality, and showed that its decline is more linear than that of overall mortality.

Second, smoking partly accounts for regional variations in mortality. In most developed regions in the world including Western Europe, North America and some East Asian countries, the smoking epidemic among males started earlier than elsewhere, in the first half of the 20th century. The adverse effect of the smoking epidemic accumulated for several decades, leading to SAF peaking in these countries around the 1980s. With the continuous decline of male smoking prevalence in these countries due to anti-smoking movements and tobacco control, years of life lost due to smoking began to decrease in recent decades. In contrast, many developing countries are currently in the early stage of the smoking epidemic, with high and increasing smoking prevalence among males, even though tobacco control policies are in place.

Smoking also accounts for some subnational differences in mortality. For example, Fenelon and Preston (2012) found that smoking accounts for the southern mortality disadvantage relative to other regions of the United States. They showed that smoking explained 65% of the subnational variation in male mortality in 2004.

Third, changes in smoking mortality largely account for changes in the between-gender differences in mortality. The gap in mortality between males and females has tended to first widen and then narrow in most developed countries, and reduced between-gender differences in smoking largely explain the current closing of the between-gender mortality gap (Pampel, 2006; Preston and Wang, 2006). Indeed, in these countries the female smoking epidemic usually started one or two decades later than the male epidemic, and thereafter followed a similar pattern. In mid- to low-income countries, female smoking-related mortality remains low but still follows a similar rising-peaking-falling trend to the male one. The SAF for males and females clearly follows the same general increasing-peaking-decreasing trend but with

different times of onset, times-to-peak and maximum values (see Figure 2.1).

Therefore, estimating and forecasting the SAF can help to improve mortality forecasts by taking the nonlinearity of mortality decline together with between-country and between-gender differentials into account (Bongaarts, 2006; Janssen et al., 2013; Stoeldraijer et al., 2015). Here we propose a new Bayesian hierarchical model to project SAF that captures the observed increasing-peaking-declining trend so that it could be used for making better mortality forecasts.

Estimating the SAF is not easy for several reasons (Bongaarts, 2014; Tachfouti et al., 2014). First, the smoking habits of individuals can differ in terms of smoking intensity, smoking history, types of tobacco used, as well as first-hand or second-hand smoking, so that estimating the prevalence of smoking (p in Eq. 2.1) based on smoking behavior data is not straightforward. Secondly, to estimate the relative risk of smoking (r in Eq. 2.1) requires accurate cohort data. Such data are challenging to collect because smoking is not a direct killer but rather has a lifelong impact, with deaths occurring mostly at older ages. The American Cancer Society's Cancer Prevention Study II (CPS-II), which began in 1982, is so far the largest study that collects such data (Tachfouti et al., 2014). Thirdly, the quality of registration and survey data varies across countries and between genders, which makes estimation and comparison of SAF across countries difficult.

Three categories of methods have been proposed to estimate SAF. The first is prevalence-based analysis in cohort studies (SAMMEC) (Levin, 1953). This uses estimated smoking prevalence from surveys and relative risk from CPS-II. The second method is prevalence-based analysis in case-control studies. This method is similar to the first one, except that the relative risk is estimated from a case-control study. It has been used for India (Gajalakshmi et al., 2003), Hong Kong (Lam et al., 2001), and China (Niu et al., 1998). The main drawback of prevalence-based methods is the scarcity of reliable historical data on smoking prevalence, especially for developing countries.

The third method, which overcomes this limitation, is an indirect method. It is called the Peto-Lopez method and was first proposed by Peto et al. (1992). This method estimates

the proportion of the population exposed to smoking using lung cancer mortality data, since most lung cancer deaths are due to smoking in developed countries. According to Centers for Disease Control and Prevention (2019), cigarette smoking is associated with more than 80% of lung cancer deaths in the United States. Simonato et al. (2001) also concluded by case-control studies in 6 developed European countries that smoking is associated with over 90% of lung cancer cases. We use this method to estimate the SAF and we describe the procedure in Section 2.2.3.

Another indirect method, the PGW method of Preston et al. (2009), also uses lung cancer mortality rate as an indicator of the cumulative hazard of smoking. Instead of using relative risks from the CPS-II as the Peto-Lopez method does, the PGW method adopts a regression-based procedure. We discuss these two methods in Section 2.5.1. More comparisons among different estimation methods of SAF can be found in Pérez-Ríos and Montes (2008), Tachfouti et al. (2014), Kong et al. (2016), and Peters et al. (2016).

Figure 2.1 plots the estimated all-age SAF (ASAF) of males and females for the United States from 1950 to 2015. It can be seen that the evolution of SAF over time follows a remarkably strong pattern, first rising and then falling. Qualitatively very similar patterns were found in most countries that we studied, although in countries with less good data, higher levels of measurement error can be seen. It seems intuitive to expect that such a regular pattern could be used to obtain good forecasts. Here we describe our method for doing this. It turns out that, indeed, good forecasts can be obtained, thanks to the strong and consistent pattern of SAF over time. Here we propose a new probabilistic projection method for the SAF using a Bayesian hierarchical model. Our method will provide estimates and projections of the SAF for both genders jointly for more than 60 countries.

The paper is organized as follows. The data, the detailed SAF calculation based on the Peto-Lopez method, and the proposed Bayesian hierarchical model are described in Section 2.2. An out-of-sample validation experiment is reported in Section 2.3. We then discuss general estimation and forecasting results for all the countries considered in this work, with detailed case studies for four countries chosen from North America, South America, Asia,

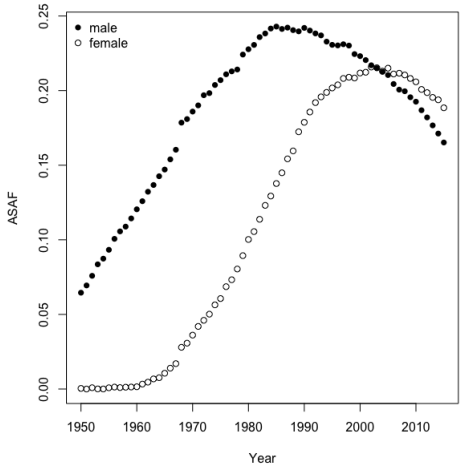


Figure 2.1: United States: All-age smoking attributable fractions of mortality for males and females from 1950 to 2015, estimated using the Peto-Lopez method.

and Europe in Section 2.4. We conclude with a discussion in Section 2.5.

2.2 Method

2.2.1 Notation

We use the symbol y to denote the estimated (observed) all-age smoking attributable fraction (ASAF), which is defined as the smoking attributable fraction for all age groups combined, and we use the symbol h to denote the true (unobserved) ASAF. All of these quantities are indexed by country c, gender s, and year t. The quantities of interest are the unobserved true past and present ASAF together with their future projections. Here the estimation time period is 1950–2015 and the projection time period is 2015–2050. Section 2.2.3 describes the estimation procedure for ASAF using the Peto-Lopez method for all available countries. A Bayesian hierarchical model will be used to model the estimated ASAF. In the Bayesian hierarchical model, the country-specific parameter vector determining the time evolution

pattern of ASAF for country c and gender s is denoted by $\theta_{c,s}$, and the global parameters by ψ.

2.2.2 Data

We use the annual death counts by country, age group, gender, and cause of death from the WHO Mortality Database (World Health Organization, 2017) which covers data from 1950 to 2015 for more than 130 countries and regions around the world. This dataset comprises death counts registered in national vital registration systems and is coded under the rules of the International Classification of Diseases (ICD). There are 5 raw datasets available by the most recent update on 11 April 2018. The first three datasets are labeled as ICD versions 7, 8, and 9 respectively, and the last two are labeled as ICD version 10.

Each version of ICD codes causes of death differently and a summary of the codes used for estimating ASAF in Section 2.2.3 is given in Table 2.1. For each country, the death counts data can differ by geographical coverage, number of years available and age group breakdown. Some countries such as China only have data from selected regions, and these countries will not be included here.

We use the quinquennial population by five-year age groups from the 2017 Revision of the World Population Prospects (United Nations, 2017) for each country, gender and age group. Since this dataset provides population estimates at five-year intervals, we use linear interpolation to obtain annual population estimates for each five-year age group.

2.2.3 ASAF Estimation

We apply the original Peto-Lopez indirect method to estimate ASAF for male and female separately. This method uses the lung cancer mortality rate as an indicator of the accumulated hazard of smoking to estimate the proportion of population exposed to smoking. As commented in Peto et al. (1992), it is very rare to observe lung cancer cases among non-smokers in developed countries, even in areas with pollution sources such as radon and asbestos. The original papers (Peto et al., 1992, 1994, 2006) applied the method to developed

Table 2.1: ICD codes for different cause of death categories across versions.

Causes	ICD-7 (A-list)	ICD-8 (A-list)	ICD-9 (09A, 09B)
Lung Cancer	A050	A051	B101
Upper Aero-digestive Cancer	A044, A045, A040	A045, A046, A050	B08, B090, B100
Other Cancer	rest of A044-A059	rest of A045-A060	rest of B08-B14
COPD	A092, A093	A093	B323, B324, B325
Other Respiratory	rest of A087-A097	rest of A089-A096	rest of B31-B32
Vascular Disease	A079-A086	A080-A088	B25-B30
Liver Cirrhosis	A105	A102	A347
Other non-med	A138-A150	A138-A150	B47-B56
Other medical	rest	rest	rest
All causes	A000	A000	B00

Causes	ICD-9 (09N)	ICD-10 (101)	ICD-10 (103, 104, 10M)
Lung Cancer	B101	1034	C33-C34
Upper Aero-digestive Cancer	B08, B090, B100	1027, 1028, 1033	C00-C15, C32
Other Cancer	rest of CH02	rest of 1027-1046	rest of C00-C97
COPD	B323, B324, B325	1076	J40-J47
Other Respiratory	rest of CH08	rest of 1072	J00-J99
Vascular Disease	CH07	1064	I00-I99
Liver Cirrhosis	S347	1080	K74, K70
Other non-med	CH17	1095	V00-Y89
Other medical	rest	rest	rest
All causes	B00	1000	AAA

countries only, especially in Western Europe and North America. With the shift of global smoking pattern, and diffusion of smoking in middle- and low-income countries, this method has been extended to less developed countries (Ezzati and Lopez, 2003, 2004; Pampel, 2006).

For estimating ASAF using the Peto-Lopez method, we need first to estimate age- and cause-of-death-specific SAF. The age groups used for estimation are 0-34, 35-59, 60-64, 65-69, 70-74, 75-79, and 80+. For each age group, annual death counts of the following nine categories of causes of death are obtained from the five raw datasets of WHO Mortality Database: lung cancer, upper aero-digestive cancer, other cancers, COPD, other respiratory diseases, vascular diseases, liver cirrhosis, non-medical causes, and all other medical causes. A detailed list of codes from ICD 7, 8, 9, and 10 for these nine categories is provided in Table 2.1.

The ICD categorizes death count data according to availability using so-called sublists, which can be one of A-list or several others; see Table 2.1. The sublists we use are those satisfying the minimum requirements for ASAF calculation. More specifically, for ICD 7 and 8, only countries whose ICD sublist is A-list are used. For ICD 9, only those countries whose ICD sublist is 09A-, 09B-, or 09N-list are used. For ICD 10, countries whose ICD sublist is one of 101-, 103-, 104-, 10M-list are used. In addition, we only calculate age-specific SAF for countries whose age group breakdown is finer than the following age group breakdown: 0-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75+. This corresponds to the age group format number 00, 01, 02, 03, 04 in the raw datasets.

To estimate the proportion of a population exposed to smoking, i.e., p in Eq. 2.1, the method compares the observed lung cancer mortality rate with the lung cancer mortality rate of smokers estimated from CPS-II. The estimated proportion, indexed by country c, age group a, gender s, and year t, is estimated by

$$p_{c,a,s,t} = \frac{d_{c,a,s,t} - d^S_{a,s}}{d^S_{a,s} - d^{NS}_{a,s}},$$

where $d_{c,a,s,t}$ is the observed country-age-gender-year-specific lung cancer mortality rate, and $d^S_{a,s}$ and $d^{NS}_{a,s}$ are age-gender-specific lung cancer mortality rates for smokers and nonsmokers

from the CPS-II respectively. Here the observed lung cancer mortality rate $d_{c,a,s,t}$ is the observed lung cancer death count divided by the population estimated from the 2017 Revision of the World Population Prospects for country c, age group a, gender s, and year t.

The Peto-Lopez method uses the CPS-II to estimate the relative risk of dying for each cause of death for smokers and nonsmokers, i.e., r in Eq. 2.1. Specifically, the Cochran-Mantel-Haenszel method is used to estimate the relative risk for age group 35-59 by combining five sub-age groups (35-39, 40-44, 45-49, 50-54, 55-59). The relative risk is indexed by cause-of-death k, age group a, and gender s. Here k takes integer values 1-9 corresponding to the nine categories mentioned above.

The excess mortality rate attributable to smoking is denoted by $er_{k,a,s}$ for cause-of-death k, age group a, and gender s. For lung cancer, the excess mortality rate attributable to smoking is calculated as $er_{1,a,s} = r_{1,a,s} - 1$. For all other categories except liver cirrhosis ($k = 7$) and non-medical causes ($k = 8$), the excess risk is discounted by 50%, i.e., $er_{k,a,s} = 0.5(r_{k,a,s} - 1)$ for $k = 2, 3, 4, 5, 6, 9$, so as to control for confounding factors. The excess risks for liver cirrhosis and non-medical causes are set to 0, i.e., $er_{7,a,s} = er_{8,a,s} = 0$. The country-cause-age-gender-year-specific SAF, denoted by $y_{c,k,a,s,t}$, is then

$$y_{c,k,a,s,t} = \frac{p_{c,a,s,t} \times er_{k,a,s}}{p_{c,a,s,t} \times er_{k,a,s} + 1}.$$

Any estimated negative values are set to zero.

Since the hazard due to smoking is accumulated across years and mostly causes deaths at older ages, the fraction of deaths due to smoking for ages 0-34 is typically very small and is set to 0. In addition, the SAF for ages 80+ is set to the same value as that for ages 75-79 since smoking data are unreliable for very old ages. Finally, the country-gender-year-specific ASAF, denoted by $y_{c,s,t}$, is a weighted average of the age-specific smoking attributable fractions $y_{c,k,a,s,t}$. Thus

$$y_{c,s,t} = \sum_a \sum_k y_{c,k,a,s,t} \times d_{c,k,a,s,t},$$

where $d_{c,k,a,s,t}$ is the country-cause-age-gender-year-specific mortality rate.

We chose the Peto-Lopez method to estimate the ASAF because it has been validated and widely used (Preston et al., 2009; Bongaarts, 2014; Tachfouti et al., 2014; Kong et al., 2016). Also, the data required for the estimation are cause- and age-specific death counts and population, which are provided with high quality by the WHO Mortality Database and the 2017 Revision of the World Population Prospects.

There are some variants of the Peto-Lopez method, which also assume that the lung cancer mortality rate is a good indicator for measuring smoking exposure. Some of the modifications include using different relative risk estimation instead of the CPS-II to extend the method to developing countries (Ezzati and Lopez, 2003) or using a regression-based approach (Preston et al., 2009). Section 2.5.1 contains more detailed discussion and comparison of these methods.

2.2.4 Model

We develop a four-level Bayesian hierarchical framework to model male and female ASAF jointly for multiple regions simultaneously.

Random walk with drift for the true ASAF The observed ASAF data show a strong and consistent pattern of increasing, then leveling, and then declining again for both genderes (Stoeldraijer et al., 2015) (see Figure 2.1 for the example of United States). This pattern can be captured by the following five-parameter double logistic curve:

$$g(t|\theta) = \frac{k}{1+\exp\{-a_1(t-1950-a_2)\}} - \frac{k}{1+\exp\{-a_3(t-1950-a_2-a_4)\}}, \qquad (2.2)$$

where t is the year of observation and θ is the double-logistic parameter vector, $\theta = (a_1, a_2, a_3, a_4, k)$.

Models based on the double logistic curve have been used quite widely for human population measures such as life expectancy and total fertility rates (Marchetti et al., 1996; Raftery et al., 2013; Alkema et al., 2011)). Due to its natural scientific interpretability, the double logistic curve has also been used in other scientific fields such as hematology (Head and McCarty, 1987; Head et al., 2004), phenology (Yang et al., 2012), and agricultural science

(Shabani et al., 2018). This function has also been used to describe social change, diffusion, and substitution processes (Grübler et al., 1999; Fokas, 2007; Kucharavy and De Guio, 2011).

Most developed countries have had male smoking prevalence that started before 1950, and peaked around the 1950s or 1960s when the adverse impacts of smoking on health became known and tobacco control measures started being put in place. This led to a peak in smoking-related mortality a generation or so later, followed by a continuous decline since then. Pampel (2005) argued that the smoking epidemic involves diffusion from males to females, and from more developed countries to less developed ones. Hence, the strong increasing-peaking-decreasing trend of ASAF observed in most countries is a consequence of the smoking epidemic diffusion process, and the double logistic curve can naturally describe its dynamics.

For the five-parameter double logistic function in Eq. 2.2, a_2 controls the first (left) inflection point of the curve and a_4 controls the distance between the first (left) and the second (right) inflection points. The rates of change at these inflection points are controlled by a_1 and a_3 respectively. The parameter k is an upper bound for the maximum value of the curve. See the left panel of Figure 2.2 for an illustration.

To represent this and also take account of the observed pattern of variability, we model changes in the true ASAF between adjacent time points using a random walk with drift given by the difference between the double logistic curve at the two points. This takes the form

$$h_{c,s,t} = h_{c,s,t-1} + g(t|\theta_{c,s}) - g(t-1|\theta_{c,s}) + \varepsilon^h_{c,s,t}, \tag{2.3}$$

where $g(\cdot|\theta_{c,s})$ (i.e., Eq. 2.2) quantifies the expected change of the true ASAF governed by the country- and gender-specific parameters $\theta_{c,s} = (a_1^{c,s}, a_2^{c,s}, a_3^{c,s}, a_4^{c,s}, k^{c,s})$, and $\varepsilon^h_{c,s,t}$ are independent Gaussian noises. This random walk with drift model is designed to capture the variability of the true ASAF and allows the uncertainty of the forecast to increase when projecting further into the future.

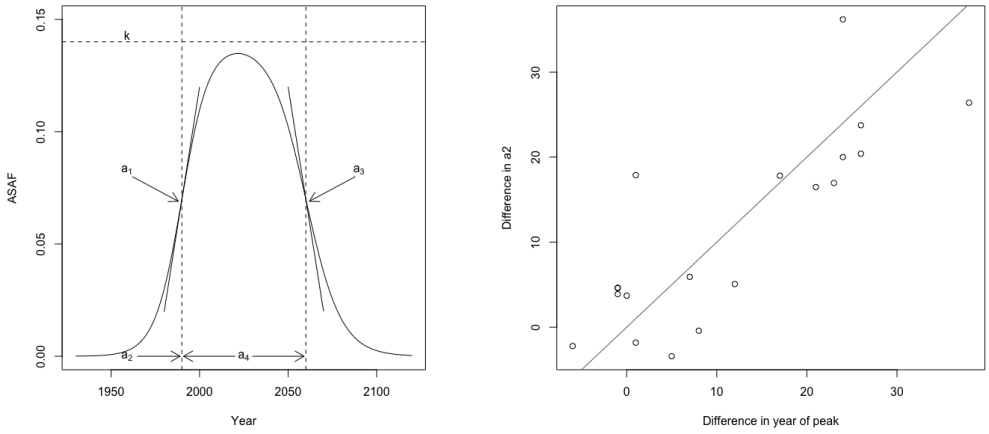


Figure 2.2: Left: The five-parameter double logistic curve. a_2 controls the left inflection point, a_4 controls the distance between left and right inflections points, a_1, a_3 determine the rate of change at left and right inflection points, and k approximates the maximum value. Right: The difference of country-specific a_2^m and a_2^f plotted against the difference between the country-specific peaks for males and females. The peak and a_2 are estimated from the countries whose male and female ASAF have all passed the maximum by 2015, according to the results of the non-linear least squares estimation. The solid line is the 45 degree line.

Male-female joint model Since the female smoking epidemic usually starts one to two decades after the male one, the start of the increase in the female ASAF is also later than that of the male ASAF. For most countries, the observed female ASAF is still in the increasing or leveling phase up to 2015. However, as the smoking epidemic diffuses from the male to the female population, it is reasonable to assume that the female ASAF will follow the same trend of increasing-leveling-declining as that of the male ASAF. This has already been observed for several countries with early smoking epidemics, such as the United Kingdom, Denmark, and Japan (Pampel, 2005; Peto et al., 2006; Janssen et al., 2013; Bongaarts, 2014; Stoeldraijer et al., 2015). For these countries, the female ASAF follows the same trend as

that of the male ASAF, but differs mainly in terms of the rate of increase or decrease, the number of years taken to reach the peak, and the peak ASAF value.

For males, we need only estimate the rate of decline of the ASAF. For females, especially for those countries whose observed ASAF data have not levelled yet, one needs first to determine the time and value of leveling. By modeling male and female data jointly, the right panel of Figure 2.2 shows that for countries whose male and female ASAF both passed the leveling period, the difference between the years of maximum of male and female is approximately the same as the difference in the a_2 parameter estimated from Eq. 2.2. The a_2 parameter represents the time point where the speed of the increasing part of the double logistic curve begins to slow down.

The difference between the times-to-peak of male and female ASAF also differs among countries. For example, the time-to-peak of the female ASAF in the United States is about 15 years later than that of the male ASAF, while the time-to-peak of the ASAF happened at about the same time for both genders in Hong Kong. To incorporate these observations, we model the difference between male and female country-specific a_2^c using a Gaussian distribution:

$$a_2^{c,f} = a_2^{c,m} + \triangle_{a_2}^c, \qquad \triangle_{a_2}^c | \triangle_{a_2}, \sigma^2_{\triangle_{a_2}} \sim \mathcal{N}(\triangle_{a_2}, \sigma^2_{\triangle_{a_2}}), \tag{2.4}$$

where $a_2^{c,m}$ and $a_2^{c,f}$ are the country- and gender-specific values of a_2, and $\triangle_{a_2}^c$ is the country-specific difference between these two parameters with prior mean $\triangle_{a_2}$ and variance $\sigma^2_{\triangle_{a_2}}$.

Moreover, since there are very few countries whose female ASAF have begun to decline by 2015, while the male ASAF has been declining for many years in most countries, we set the same global parameters for the gender-specific parameters $a_4^{c,m}$ and $a_4^{c,f}$ for each country, namely,

$$a_4^{c,m}, a_4^{c,f} | a_4, \sigma^2_{a_4} \overset{\text{ind}}{\sim} \mathcal{N}(a_4, \sigma^2_{a_4}). \tag{2.5}$$

Except for a_4^c, the other four country-specific parameters of the double logistic curve are conditioned on their own gender-specific global parameters.

Measurement error model for observed ASAF The observed country-gender-year-specific ASAF $y_{c,s,t}$ are modeled based on the true (unobserved) ASAF $h_{c,s,t}$ by incorporating measurement error due to the variability of data quality across different countries:

$$y_{c,s,t}|h_{c,s,t},\sigma_c^2 \sim_{ind} \mathcal{N}(h_{c,t,s},\sigma_c^2). \tag{2.6}$$

We assume that the variance of the observed ASAF for each country is time- and gender-invariant based on exploratory analyses that indicate that the data quality is consistent across time and between genders within the same country.

Summary of model We combine the Bayesian hierarchical model and measurement error model into a four-level Bayesian hierarchical model. We model the observed ASAF estimates using the measurement error model in Level 1, conditional on the true (unobserved) ASAF data which are modeled with a random walk with drift in Level 2, conditional on the country-specific parameters. Country-specific parameters are modeled in Level 3, where parameters for male and female ASAF are modelled jointly conditional on the global parameters, whose prior distributions are specified in Level 4.

The overall model is specified as follows:

$$\begin{aligned}
&\text{Level 1: } y_{c,s,t}|h_{c,s,t} \sim \mathcal{N}(h_{c,s,t},\sigma_c^2);\\
&\text{Level 2: } h_{c,s,t_{0,c}} = g(t_{0,c}|\theta_{c,s}) + \varepsilon^h_{c,s,t_{0,c}},\\
&\qquad\qquad h_{c,s,t} = h_{c,s,t-1} + g(t|\theta_{c,s}) - g(t-1|\theta_{c,s}) + \varepsilon^h_{c,s,t} \text{ for } t > t_{0,c},\\
&\qquad\qquad \varepsilon^h_{c,s,t} \overset{\text{ind}}{\sim} \mathcal{N}(0,\sigma_h^2);\\
&\text{Level 3: } \theta_{c,s} \sim f(\cdot|\psi),\\
&\qquad\qquad \sigma_c^2 \sim \text{Lognormal}(\nu,\rho^2);\\
&\text{Level 4: } \psi,\nu,\rho^2,\sigma_h^2 \sim \pi(\cdot).
\end{aligned}$$

Here, $t_{0,c}$ is the year of the first available ASAF data for country c, g denotes the five-parameter double logistic curve in Eq. 2.2, f denotes the conditional distribution of the

country-specific parameters $\theta_{c,s}$, and π denotes the hyperpriors for the global parameters $\psi, \nu, \rho^2, \sigma_h^2$. The country-specific parameters $\theta_{c,s} = (a_1^{c,s}, a_2^{c,s}, a_3^{c,s}, a_4^{c,s}, k^{c,s})$ are gender-specific and the interaction between male and female parameters are governed by Eq. 2.4 and 2.5. The global parameters $\psi = (a_1^m, a_2^m, a_3^m, a_4, k^m, a_1^f, a_3^f, k^f, \triangle_{a_2}, \sigma^2_{a_2^m}, \sigma^2_{a_4}, \sigma^2_{k^m}, \sigma^2_{k^f}, \sigma^2_{\triangle_{a_2}})$ are also gender-specific except for $\triangle_{a_2}, \sigma^2_{\triangle_{a_2}}, a_4, \sigma^2_{a_4}$. More information about the specification of the full model is given in the Appendix A.1.

Estimation and prediction Statistical analysis of the model is carried out in two phases, estimation and prediction. The goal of the estimation phase is to obtain the joint posterior distribution of the true ASAF $h_{c,s,t}$ during the estimation period 1950–2015 and the country-specific parameters for the underlying double-logistic curve. The aim of the prediction phase is to forecast the future ASAF of both genders for the prediction period 2015–2050 based on the observed ASAF for over 60 countries whose male ASAF data are classifed as clear-pattern (see Section 2.2.5 for the definition of clear-pattern).

The functional form of the prior distribution $\pi(\cdot)$ is assessed using results from non-linear least squares estimation based on clear-pattern countries (see Section 2.2.5 for details). Specifically, the priors for $(a_1^m, a_2^m, a_3^m, a_4, k^m, \sigma^2_{a_2^m}, \sigma^2_{a_4}, \sigma^2_{k^m}, \sigma^2_{a_2^m}, \sigma^2_{a_4}, \sigma^2_{k^m})$ are based on non-linear least squares results from the male ASAF of over 60 clear-pattern countries, the prior for a_1^f is estimated based on non-linear least squares results from the female ASAF of 52 clear-pattern countries, the priors for $(a_3^f, k^f, \sigma^2_{a_3^f})$ are set to the same priors as their counterparts for males, while the priors for $(\triangle_{a_2}, \sigma^2_{\triangle_{a_2}})$ are estimated based on 19 countries for which both male and female ASAF have passed the leveling stage by 2015. The priors for ν, ρ^2, σ_h^2 are estimated by pooling male and female ASAF from all clear-pattern countries. A complete specification of the model is given in the Appendix A.1.

2.2.5 ASAF Categorization

We categorize estimated ASAF for 127 countries and regions into two categories according to the data availability and quality: clear-pattern and non-clear-pattern. On one hand,

the Peto-Lopez method is not guaranteed to produce reliable ASAF estimates for some less developed countries because of poor data quality. On the other hand, modeling only with clear-pattern countries can improve estimation and projection accuracy without introducing too much random noise.

The classification is based on non-linear least squares estimation of the following model for each country and gender separately:

$$y_t = g(t|\theta) + \varepsilon_t,$$

where $g(t|\theta)$ is as in Eq. 2.2 and ε_t are independent standard Gaussian errors. Its fit to the data in a given country provides an indication of data quality for that country.

Our categorization is based on the number of observations, maximum of observed values, and the R^2 value of the non-linear least squares fit. Due to the differences between the diffusion processes of smoking in the male and female populations (Pampel, 2006), we use different criteria for male and female data. For male data, we require that (1) the number of available annual observations up to 2015 be greater than 10; (2) at least one of the observations be greater than 0.05; and (3) that the R^2 value be greater than 0.5.

For female data, since the smoking epidemic in general started one to two decades later than the male one, the onset and the value of the ASAF is later and smaller than that of the male epidemic (Pampel, 2005; Preston and Wang, 2006). The criteria for female data are that (1) the number of observations up to 2015 be greater than 10; (2) at least one of the observations be greater than 0.01; and (3) that the R^2 value be greater than 0.6.

By these rules, there are over 60 countries whose male data are classifed as clear-pattern (2 in Africa, 16 in the Americas, 9 in Asia, 40 in Europe and 2 in Oceania), and 52 countries whose female data are classified as clear-pattern (12 in the Americas, 7 in Asia, 31 in Europe and 2 in Oceania).

2.2.6 Estimation

Estimation is based on the male and female ASAF data from over 60 countries whose male ASAF is classified as clear-pattern for the period 1950–2015. The reason why we chose clear-pattern ASAF data is that non-clear-pattern data either have too few observations, very low values, or their shapes are not identifiable.

We used the `Rstan` package (Version 2.18.2) in R to obtain the joint posterior distributions of the parameters of interest (Carpenter et al., 2017). Rstan uses a No-U-turn sampler, which is an adaptive variant of Hamiltonian Monte Carlo (Neal, 2011; Hoffman and Gelman, 2014). We ran 3 chains with different initial values, each of length 10,000 iterations with a burn-in of 2,000 without thining. This yielded a final, approximately independent sample of size 8,000 for each chain. We monitored convergence by inspecting trace plots and using standard convergence diagnostics.

We also conducted a sensitivity analysis on the hyperparameters that specify the priors $\pi(\cdot)$ for the global parameters ψ, and concluded that the proposed model is not sensitive to the choice of hyperparameters. More information about the convergence diagnostics and the sensitivity analysis is given in the Appendix A.2 and A.3.

2.2.7 Projection

We produce projections of future ASAF for the period 2015–2050 for over 60 countries whose male ASAF is classified as clear-pattern. The prediction of future ASAF for each country is based on past and present ASAF. We sample from the joint posterior distribution of the country-specific parameters $\theta_{c,s}$ and of the past, and present true ASAF $h_{c,s,t}$. We then use Eq. 2.3 and 2.6 to generate a sample of trajectories of future true and observed ASAF respectively from their joint posterior predictive distribution. It is possible that the quantity generated by Eq. 2.3 and Eq. 2.6 is negative, and we set such values to zero. This yields a sample from the joint posterior predictive distribution of the future ASAF for over 60 countries, for both genders, taking account of uncertainty about the past observations as

well as the future evolution. We include the plots of ASAF projections for over 60 countries and both genders in the Appendix A.4.

2.3 Results

We assess the predictive performance of our model using out-of-sample predictive validation.

2.3.1 Study Design

The data we used for out-of-sample validation cover the period 1950–2015. We assess the quality of our model based on different choices of estimation and validation data from the observed data. Since the trend of increasing-leveling-declining pattern plays an important role for estimation and projection, assessing how the model works when only part of the trend has been observed is crucial. We consider different choices for estimation and validation periods, namely (1) 1950– 2000 for estimation and 2000–2015 for validation; (2) 1950–2005 for estimation and 2005–2015 and for validation; and (3) 1950–2010 for estimation, 2010–2015 for validation. The countries used for validation in each time-split scenario are required to be clear-pattern countries based on the male ASAF, to contain more than 10 observations in the estimation period, and to have at least one observation in the prediction period. This results in 63, 66 and 66 countries used for validation under choices (1), (2) and (3), respectively.

Since we are making probabilistic projections, our evaluation is based on both accuracy of point prediction and calibration of prediction intervals. Our goal is not only to produce accurate point predictions, but also to account for variability of future predictions based on historic data, especially for those countries whose data in the estimation period reveal only part of the pattern. If the proposed model works well, we would expect the point predictor to have small gender-specific mean absolute error (MAE), which is defined as

$$\text{MAE}_s = \frac{1}{N} \sum_{c \in \mathcal{C}} \sum_{t \in \mathcal{T}_c} |\hat{y}_{c,s,t} - y_{c,s,t}|, \tag{2.7}$$

where $\mathcal{C}$ is the set of countries considered in the validation, $\mathcal{T}_c$ is the set of country-year combinations used for validation, $\hat{y}_{c,s,t}$ is the posterior median of the predictive distribution of ASAF at year t for country c and gender s, and N is the total number of data used for validation.

We wish the prediction to be well calibrated and sharp, i.e., the coverage of the prediction interval to be close to the nominal level with its half-width as short as possible. Thus, we include the empirical coverage and the half-width of the prediction interval in the validation. To assess the overall predictive performance, we also calculate the gender-specific continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007), which is defined as

$$\text{CRPS}_s = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\left[\frac{1}{|\mathcal{T}_c|}\sum_{t\in\mathcal{T}_c}\int_{-\infty}^{\infty}\{F_{c,s,t}(y) - \mathbf{1}(y_{c,s,t}\leq y)\}^2 dy\right], \tag{2.8}$$

where $F_{c,s,t}(y)$ is the predictive distribution of the future ASAF for country c, gender s, and time t, and $\mathbf{1}(\cdot)$ is equal to 1 if the condition in the parenthesis is satisfied and 0 otherwise. CRPS is a summary statistic measuring the quality of the probabilistic forecast, which evaluates model calibration and sharpness simultaneously. The smaller the CRPS, the closer the predictive distribution to the true data-generating distribution.

2.3.2 Out-of-sample Validation Results

To our knowledge, no other method is available in the literature to produce probabilistic forecasts for male and female ASAF for developed and developing countries jointly. Janssen et al. (2013) and Stoeldraijer et al. (2015) developed methods for projection of age-specific SAF and age-standardized SAF, and their methods are based on age-period-cohort analysis, which cannot be trivially extended to ASAF. See Section 2.5.2 for more discussion of their procedures and comparison to the present ones.

As benchmarks against which to compare our method, we consider four other forecast procedures. The first one is the persistence forecast, which takes the last observed value as the forecast for the prediction period. The second method is the Bayesian thin plate regression spline method (Wood, 2003), implemented in the `mgcv` package (Version 1.8-27)

in R. The third method is the Bayesian structural time series model (Harvey, 1990; Durbin and Koopman, 2012), implemented in the `bsts` package (Version 0.8.0) in R. Here we choose to use two state components — local linear trend and autocorrelation with lag 1 — to build the structural time series model. Our fourth comparison method is a non-hierarchical version of our proposed model, namely our proposed model without Level 4 (i.e., the global parameters). This is included to see whether the hierarchical structure is necessary.

We summarize the validation results in Table 2.2 for males and females separately. This shows the MAE, the coverage and half-width of the prediction intervals, and the continuous ranked probability score (CRPS). For males, our method improved the prediction accuracy for all three scenarios over the persistence forecast. For forecasting one and two five-year periods ahead, our method improved the MAE by 30% and 21% respectively. Since most male ASAF series had passed the peak by 2005 and had experienced declines for several years, the double logistic curve captures this trend well. For predictions three five-year periods into the future, during which the male ASAF series for some countries were just reaching the peak, our method still improved the MAE by 6%. For females, we observed similar improvements. Our method decreased the MAE by 22%, 17%, and 27% for predictions one, two, and, three five-year periods ahead compared to those of the persistence forecast.

Also, compared with other probabilistic forecast methods, our method produced shorter prediction intervals with empirical coverages close to the nominal level for one and two five-year predictions, while it produced predictive intervals with reasonably close to nomial for the three five-year predictions for the male ASAF. On the other hand, since most female ASAF series have not yet reached the peak, capturing the variability of future female ASAF is essential. The coverage of our method is close to the nominal level, indicating that our method is well calibrated.

Overall, our proposed BHM yielded the smallest CRPS among all methods in most cases for both the male and female epidemics. Among all five methods compared in the validation exercise, the Bayesian spline method was worst in terms of forecast accuracy, and tended to underestimate the variability of future values. The Bayesian structural time series model

produced predictive interval close to the nominal level with slightly larger average half-width than our method. However, a significant drawback of the persistence forecast, the Bayesian spline method, and the Bayesian structural time series model is that they tend to produce unrealistic forecasts when all the observed data are before the peak, since they do not incorporate the increasing-peaking-decreasing information in the model. The left panel of Figure 2.3 indicates that the Bayesian thin plate spline method projected a monotonically increasing ASAF for United States female based on data before 2000, where the entire prediction interval missed the observed data after 2000. The right panel of Figure 2.3 shows that the Bayesian structural time series model did cover the data but with an unrealistically wide prediction interval.

The Bayesian model without the global level parameters produced results similar to those from our BHM for projecting short term male ASAF. When forecasting three five-year periods ahead, or the female ASAF, in both of which cases the peak has often not been reached, the Bayesian model without the global level parameters was worse in accuracy and CRPS. This indicates that the hierarchical structure did indeed improve the overall forecast when only part of the trend has been observed, by sharing information among all the countries.

Table 2.3 gives validation results for subgroups of countries, categorized by membership of the Organization for Economic Cooperation and Development (OECD). Most of the countries in the OECD are regarded as developed countries with high GDP and human development index (HDI). For male ASAF, our BHM improved most of the forecasts for OECD countries, especially the longer term projections. For OECD countries, the increasing-peaking-decreasing pattern is clearer and stronger, which fits with our modeling well. In contrast, our BHM performed less well among non-OECD countries.

Table 2.2: Predictive validation results for all-age smoking attributable fraction (ASAF). The first and second columns indicate the estimation and validation periods . The "Gender" and "n" columns indicate the gender and the number of countries used for the validation. In the "Model" column, "Bayes" represents the Bayesian hierarchcial model with measurement error and random walk with drift, "Bayes(S)" represents the same model as "Bayes" without the global parameters, "Persistence" represents the persistence forecast, "Spline" represents the Bayesian thin plate regression spline method, and "BSTS" represents the Bayesian structural time series method. The "MAE" column contains the mean absolute prediction error defined by Eq. 2.7. The "Coverage" columns show the proportion of validation observations contained in the 80%, 90%, 95% prediction intervals with their average half-widths in parentheses. The "CRPS" column contains the continuous ranked probability score defined by Eq. 2.8.

Training	Test	n	Gender	Model	MAE	Coverage			CRPS
						80%	90%	95%	
1950–2010	2010–2015	66	Male	Persistence	0.010	-	-	-	-
				Bayes	0.007	0.78 (0.011)	0.86 (0.014)	0.90 (0.017)	0.00523
				Bayes(S)	0.007	0.86 (0.014)	0.94 (0.018)	0.97 (0.022)	0.00505
				Spline	0.008	0.58 (0.009)	0.65 (0.011)	0.72 (0.013)	0.00648
				BSTS	0.008	0.85 (0.015)	0.94 (0.020)	0.94 (0.025)	0.00570
			Female	Persistence	0.009	-	-	-	-
				Bayes	0.007	0.83 (0.012)	0.93 (0.015)	0.96 (0.018)	0.00507
				Bayes(S)	0.008	0.88 (0.014)	0.94 (0.018)	0.97 (0.022)	0.00538
				Spline	0.010	0.42 (0.007)	0.52 (0.009)	0.61 (0.011)	0.00763
				BSTS	0.008	0.80 (0.013)	0.89 (0.016)	0.94 (0.020)	0.00562

Continued on next page

Table 2.2 – *Continued from previous page*

Training	Test	n	Gender	Model	MAE	Coverage			CRPS
						80%	90%	95%	
1950–2005	2005–2015	66	Male	Persistence	0.014	-	-	-	-
				Bayes	0.011	0.72 (0.014)	0.83 (0.018)	0.89 (0.022)	0.00797
				Bayes(S)	0.010	0.85 (0.020)	0.93 (0.027)	0.97 (0.033)	0.00795
				Spline	0.014	0.54 (0.014)	0.65 (0.018)	0.72 (0.021)	0.01096
				BSTS	0.013	0.83 (0.026)	0.90 (0.035)	0.95 (0.043)	0.00989
			Female	Persistence	0.012	-	-	-	-
				Bayes	0.010	0.80 (0.015)	0.90 (0.020)	0.92 (0.025)	0.00721
				Bayes(S)	0.011	0.88 (0.021)	0.93 (0.028)	0.95 (0.035)	0.00808
				Spline	0.014	0.44 (0.011)	0.51 (0.014)	0.58 (0.016)	0.01133
				BSTS	0.011	0.77 (0.017)	0.88 (0.023)	0.93 (0.029)	0.00802
1950–2000	2000–2015	63	Male	Persistence	0.017	-	-	-	-
				Bayes	0.016	0.65 (0.020)	0.76 (0.026)	0.84 (0.031)	0.01214
				Bayes(S)	0.018	0.84 (0.031)	0.92 (0.042)	0.95 (0.052)	0.01278
				Spline	0.018	0.59 (0.019)	0.69 (0.024)	0.76 (0.029)	0.01335
				BSTS	0.016	0.85 (0.039)	0.93 (0.053)	0.98 (0.068)	0.01281
			Female	Persistence	0.015	-	-	-	-
				Bayes	0.011	0.81 (0.021)	0.90 (0.029)	0.95 (0.037)	0.00817
				Bayes(S)	0.012	0.88 (0.027)	0.96 (0.039)	0.98 (0.050)	0.00887
				Spline	0.016	0.48 (0.014)	0.59 (0.018)	0.70 (0.022)	0.01151
				BSTS	0.012	0.79 (0.022)	0.89 (0.030)	0.94 (0.039)	0.00831

Table 2.3: Predictive validation results for all-age smoking attributable fraction (ASAF) for categories of countries. The "OECD" column represents whether the countries in the subgroup belong to the OECD. The number of countries in the subgroup used for the validation is in parentheses. All the other columns are the same as those in Table 2.2.

Training	Test	Gender	OECD	Model	MAE	Coverage			ACRPS
						80%	90%	95%	
1950–2010	2010–2015	Male	Y(34)	Persistence	0.011	-	-	-	-
				Bayes	0.006	0.81 (0.011)	0.90 (0.014)	0.95 (0.016)	0.00448
				Bayes(S)	0.006	0.88 (0.013)	0.94 (0.017)	0.99 (0.021)	0.00459
				Spline	0.007	0.60 (0.008)	0.67 (0.010)	0.73 (0.012)	0.00565
				BSTS	0.007	0.86 (0.014)	0.95 (0.018)	0.98 (0.022)	0.00529
			N(32)	Persistence	0.008	-	-	-	-
				Bayes	0.009	0.75 (0.011)	0.81 (0.015)	0.84 (0.018)	0.00601
				Bayes(S)	0.008	0.85 (0.015)	0.92 (0.019)	0.94 (0.023)	0.00554
				Spline	0.010	0.56(0.010)	0.63 (0.012)	0.70 (0.015)	0.00736
				BSTS	0.009	0.86 (0.017)	0.95 (0.023)	0.98 (0.028)	0.00629
		Female	Y(34)	Persistence	0.009	-	-	-	-
				Bayes	0.007	0.82 (0.011)	0.92 (0.015)	0.94 (0.018)	0.00505
				Bayes(S)	0.008	0.86 (0.013)	0.93 (0.017)	0.96 (0.021)	0.00560
				Spline	0.010	0.42 (0.007)	0.51 (0.008)	0.58 (0.010)	0.00762
				BSTS	0.009	0.78 (0.012)	0.85 (0.015)	0.91 (0.019)	0.00616
			N(32)	Persistence	0.008	-	-	-	-
				Bayes	0.008	0.83 (0.012)	0.95 (0.015)	0.95 (0.018)	0.00507
				Bayes(S)	0.007	0.89 (0.015)	0.95 (0.019)	0.98 (0.023)	0.00516
				Spline	0.011	0.42(0.008)	0.54 (0.010)	0.63 (0.012)	0.00764
				BSTS	0.007	0.82 (0.013)	0.89 (0.017)	0.94 (0.021)	0.00506

Continued on next page

Table 2.3 – *Continued from previous page*

Training	Test	Gender	OECD	Model	MAE	Coverage			ACRPS
						80%	90%	95%	
1950–2005	2005–2015	Male	Y(34)	Persistence	0.016	-	-	-	-
				Bayes	0.010	0.73 (0.014)	0.85 (0.018)	0.90 (0.021)	0.00676
				Bayes(S)	0.010	0.84 (0.019)	0.93 (0.025)	0.97 (0.032)	0.00717
				Spline	0.013	0.52 (0.012)	0.61 (0.015)	0.69 (0.018)	0.01008
				BSTS	0.012	0.85 (0.028)	0.91 (0.039)	0.97 (0.049)	0.01000
			N(32)	Persistence	0.011	-	-	-	-
				Bayes	0.012	0.70 (0.014)	0.81 (0.019)	0.88 (0.022)	0.00928
				Bayes(S)	0.011	0.87 (0.021)	0.93 (0.029)	0.96 (0.035)	0.00879
				Spline	0.015	0.57 (0.016)	0.68 (0.020)	0.76 (0.024)	0.01189
				BSTS	0.013	0.83 (0.026)	0.90 (0.035)	0.95 (0.043)	0.00989
		Female	Y(34)	Persistence	0.012	-	-	-	-
				Bayes	0.009	0.82 (0.015)	0.92 (0.020)	0.95 (0.025)	0.00669
				Bayes(S)	0.010	0.88 (0.019)	0.95 (0.025)	0.96 (0.032)	0.00736
				Spline	0.012	0.38 (0.008)	0.45 (0.011)	0.52 (0.013)	0.00945
				BSTS	0.012	0.82 (0.019)	0.90 (0.026)	0.92 (0.033)	0.00851
			N(32)	Persistence	0.013	-	-	-	-
				Bayes	0.011	0.78 (0.015)	0.88 (0.020)	0.90 (0.025)	0.00780
				Bayes(S)	0.012	0.88 (0.023)	0.91 (0.031)	0.93 (0.039)	0.00885
				Spline	0.017	0.51 (0.013)	0.58 (0.017)	0.66 (0.020)	0.01333
				BSTS	0.011	0.77 (0.017)	0.88 (0.023)	0.93 (0.029)	0.00802

Continued on next page

Table 2.3 – *Continued from previous page*

Training	Test	Gender	OECD	Model	MAE	Coverage			ACRPS
						80%	90%	95%	
1950–2000	2000–2015	Male	Y(33)	Persistence	0.018	-	-	-	-
				Bayes	0.014	0.67 (0.020)	0.79 (0.026)	0.88 (0.032)	0.01063
				Bayes(S)	0.017	0.83 (0.030)	0.90 (0.040)	0.95 (0.050)	0.01221
				Spline	0.017	0.58 (0.015)	0.68 (0.020)	0.74 (0.023)	0.01338
				BSTS	0.018	0.88 (0.035)	0.93 (0.047)	0.97 (0.060)	0.01308
			N(30)	Persistence	0.017	-	-	-	-
				Bayes	0.019	0.63 (0.020)	0.72 (0.026)	0.80 (0.031)	0.01377
				Bayes(S)	0.018	0.86 (0.032)	0.93 (0.042)	0.95 (0.053)	0.01341
				Spline	0.018	0.60 (0.022)	0.71 (0.029)	0.79 (0.034)	0.01331
				BSTS	0.016	0.85 (0.045)	0.94 (0.063)	0.98 (0.082)	0.01308
		Female	Y(33)	Persistence	0.016	-	-	-	-
				Bayes	0.011	0.80 (0.021)	0.89 (0.029)	0.95 (0.037)	0.00817
				Bayes(S)	0.013	0.84 (0.028)	0.94 (0.038)	0.97 (0.048)	0.00981
				Spline	0.016	0.41 (0.012)	0.54 (0.015)	0.62 (0.018)	0.01230
				BSTS	0.011	0.73 (0.016)	0.86 (0.022)	0.92 (0.027)	0.00777
			N(30)	Persistence	0.013	-	-	-	-
				Bayes	0.010	0.82 (0.017)	0.92 (0.023)	0.95 (0.030)	0.00699
				Bayes(S)	0.010	0.93 (0.028)	0.98 (0.040)	0.99 (0.052)	0.00784
				Spline	0.016	0.56 (0.017)	0.66 (0.022)	0.79 (0.026)	0.01066
				BSTS	0.010	0.86 (0.022)	0.94 (0.030)	0.95 (0.039)	0.00735

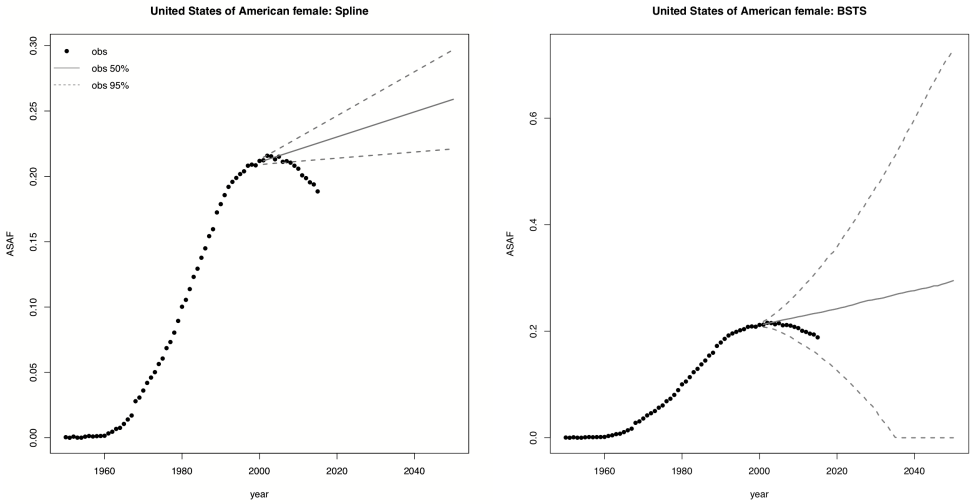


Figure 2.3: Forecast of United States female ASAF based on data before 2000 using Bayesian spline method (left) and Bayesian structured time series method (right). Black dots are observed ASAF. The solid and dashed line represent posterior median and 95% predictive interval, respectively.

Figure 2.4 shows validation results for the male ASAF of four countries or regions for predictions three five-year periods ahead. We see that our method works quite well for the United States and Hong Kong, and the prediction interval captures the variability of the male ASAF of Chile. Figure 2.5 shows the results from Scenario (1) where most female ASAF of countries among the examples have not reached the peak by the year 2000. We see that the posterior median of the predictive distribution captures the general trend of future female ASAF of the United States, the Netherlands, and Chile reasonably well. For countries or regions like Hong Kong whose female ASAF already passed the peak, our method also accurately estimates the rate of decline.

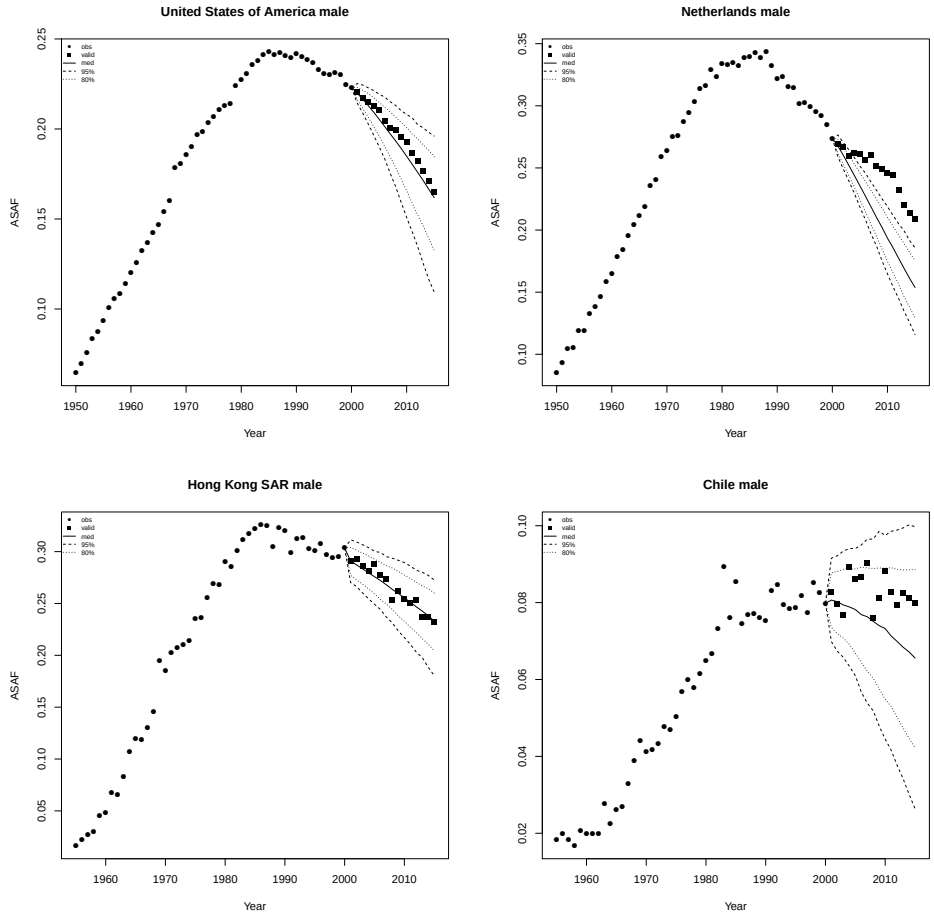


Figure 2.4: Validation of male all-age smoking attributable fraction for the United States, Netherlands, Hong Kong, and Chile. Past observed ASAF values are shown by black dots for 1950–2000 and by black squares for 2000–2015. The posterior median for 2000–2015 is shown by the solid line, and the 80% and 95% prediction intervals are shown by dotted and dashed lines respectively.

2.4 Case Studies

Probabilistic forecasts of ASAF to 2050 are given in the Appendix A.4 for over 60 countries. Broadly, the patterns in the OECD countries are similar, with male ASAF having declined

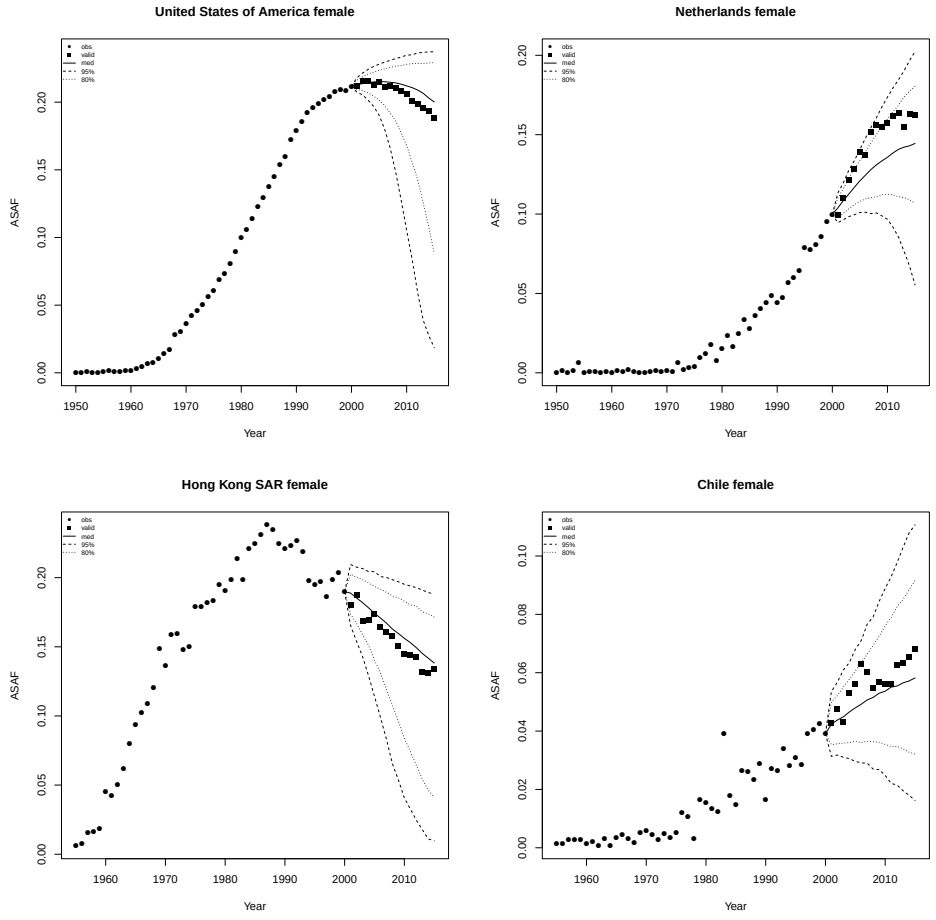


Figure 2.5: Validation of female all-age smoking attributable fraction for the United States, Netherlands, Hong Kong, and Chile. Past observed ASAF values are shown by black dots for 1950–2000 and by black squares for 2000–2015. The posterior median for 2000–2015 is shown by the solid line, and the 80% and 95% prediction intervals are shown by dotted and dashed lines respectively.

from about 30% in the 1990s to around 15% in 2015, with further declines projected to 2050, reaching around 5%. The patterns vary more for females in OECD, and for both males and females in non-OECD countries because they are currently at different stages of the

epidemic.

We now give four cases studies which illustrate various aspects of the proposed method for estimating and forecasting ASAF.

2.4.1 United States

The annual ASAF for both male and female for the time period 1950–2015 is shown in Figure 2.1. The very clear pattern is due to the high quality of the data, reflecting the fact that the United States has one of the the best vital registration systems in the world.

The smoking epidemic in the male population in the United States started in the earlier 1900s, and there was a substantial decrease of smoking prevalence and lung cancer mortality rate after the 1950s. Smoking prevalence among US male adults was approximately 60% in 1950s, and went down to about 20% in the 1990s, and the general decline is still continuing (Burns et al., 1997; Islami et al., 2015). The observed ASAF levelled around the 1990s and declined afterwards. We forecast that by 2050, the median observed ASAF for US males will be around 4.3% (with 95% prediction interval [0.0%, 8.3%]). Because the measurement error for the US is tiny, the projected true ASAF (long dashed line for posterior mean and dotted line for 95% predictive interval in Figure 2.6) for US males is almost equal to that of the observed ASAF.

The female smoking epidemic started two decades later than the male one and the maximum prevalence was around 30% in the 1960s, and then declined to about 20% in the 1990s (Burns et al., 1997). The pattern of smoking prevalence among US females is similar to that for males, but around 20 years behind (Burns et al., 1997; Islami et al., 2015). The female ASAF started to rise around the 1960s and reached its peak of 23% around 2005. We forecast that by 2050, the median observed ASAF for US females will be around 2.7% (with 95% prediction interval [0.0%, 9.3%]). Similarly, the projected US female true ASAF follows closely with that of the observed ASAF. Figure 2.6 shows the historical records of the observed male and female ASAF during the time period 1950–2015, along with projections up to 2050 with posterior median and prediction intervals.

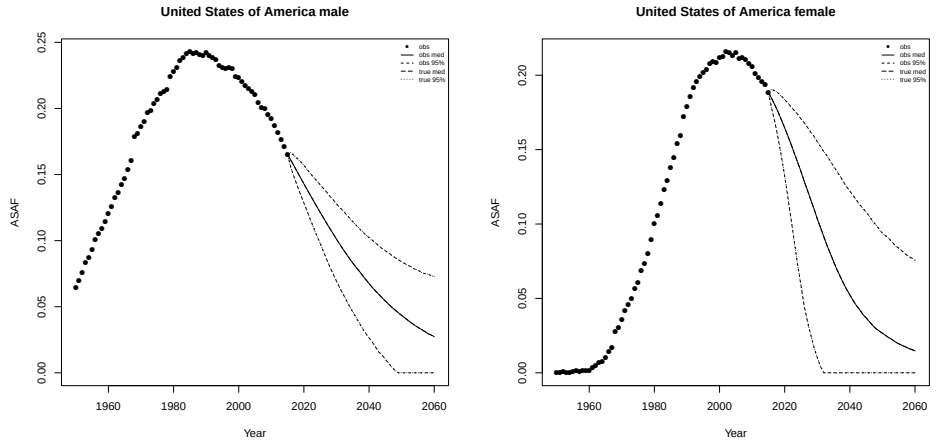


Figure 2.6: ASAF projection of the United States. The left and right panels show the projection of ASAF up to 2050 under the proposed model for male and female respectively. The solid and long dashed lines show the posterior median of projected observed ASAF and true ASAF respectively. The dashed and dotted lines represent 95% prediction intervals for observed ASAF and true ASAF respectively.

2.4.2 The Netherlands

The Netherlands is a high-income western Europe country whose smoking epidemic started relatively early. Smoking prevalence reached 90% in the 1950s and dropped to 30% in the 2010s. The male observed ASAF in Netherlands passed its maximum ASAF around the 1990s and we project that it will go down to around 5.7% (with 95% prediction interval [1.4%, 9.7%]) in 2050.

For females, smoking prevalence is also relatively high, and reached its peak of about 40% in the 1970s and dropped to 24% in the 2010s (Stoeldraijer et al., 2015). The female ASAF in Netherlands is among the few that is already experiencing the leveling stage. By our projection, the median year-to-peak for the female ASAF will be around 2020, which is about 30 years after the male peak, and will reach 16.6% (with 95% prediction interval [12.4%,

18.5%]). By 2050, the median observed female ASAF will be 4.7% (with 95% prediction interval [0.0%, 19.3%]). Similarly to the case of US, the projected true ASAF follows that of the observed ASAF closely, due to the small measurement error. Figure 2.7 shows the historical records of the observed male and female ASAF during time period 1950–2015, and projections are given up to 2050 with posterior median and prediction intervals for both observed and true ASAF.

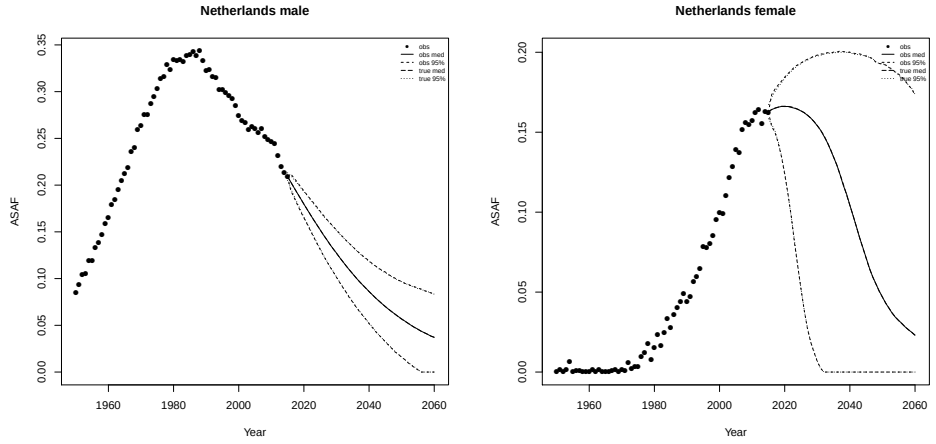


Figure 2.7: ASAF projection of the Netherlands. The left and right panels show the projection of ASAF up to 2050 under the proposed model for male and female respectively. The solid and long dashed lines show the posterior median of projected observed ASAF and true ASAF respectively. The dashed and dotted lines represent 95% prediction intervals for observed ASAF and true ASAF respectively.

2.4.3 Hong Kong

Hong Kong has an advanced smoking epidemic, but had a decrease in male smoking prevalence from about 40% in the 1980s to 22% in 2000. A decline has also been observed in female smoking prevalence, from 5.6% to 3.3% (Au et al., 2004). Like Japan, Singapore,

and South Korea, both male and female ASAF have passed the leveling stage and have been declining for two decades. Unlike in most western developed countries, the time trend of the ASAF has been almost identical for males and females in Hong Kong, with similar times of onset and times-to-peak. Au et al. (2004) showed that the time trends of lung cancer incidence were similar for both genders.

By our projection, the observed ASAF will reach 9.7% for males (with 95% prediction interval [4.9%, 14.3%]) and 4.1% for females (with 95% prediction interval [0.0%, 8.1%]) by 2050. Compared with US and the Netherlands, the projected true ASAF of Hong Kong will have narrower predictor intervals than those of the observed ASAF due to larger measurement error exhibited in the historical data. However, the difference becomes less and less since the majority uncertainty of the future ASAF will be account mainly by the variance from the random walk model of the true ASAF.

As discussed by Lam et al. (2001), Hong Kong may be a good indicator for the future development of the smoking epidemic and its impact on mortality in mainland China and other developing countries. Figure 2.8 shows the historical records of the observed male and female ASAF during time period 1950–2015, along with projections up to 2050 with posterior median and prediction intervals.

2.4.4 Chile

Chile is one of the South America countries that have clear-pattern ASAF data for both males and females. It also has relatively high smoking prevalence. A decline in prevalence among males and females has been observed in recent years but is modest compared to the decline in the United States (Islami et al., 2015). Also, female smoking prevalence is far behind that of males.

Our method projects that the male ASAF will decline gradually. By 2050, the projected median observed ASAF for the male population will be 4.3% (with 95% prediction interval [0.0%, 9.1%]). For females, we expect an increase for another 10 years with the median observed ASAF reaching the maximum 7.6% (with 95% prediction interval [2.0%, 11.8%]) by

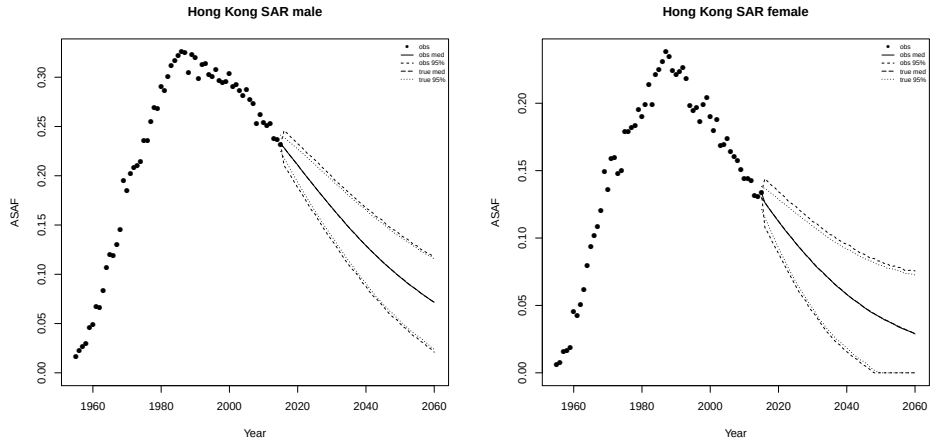


Figure 2.8: ASAF projection of Hong Kong. The left and right panels show the projection of ASAF up to 2050 under the proposed model for male and female respectively. The solid and long dashed lines show the posterior median of projected observed ASAF and true ASAF respectively. The dashed and dotted lines represent 95% prediction intervals for observed ASAF and true ASAF respectively.

2030. By 2050, the median observed female ASAF be 5.36% (with 95% prediction interval [0.0%, 15.2%]); see Figure 2.9. Similarly to Hong Kong, Chile also has larger measurement error and the pattern is less clear, so that the projected true ASAF has wider predictive intervals compared with previous cases and the difference between true and observed projections also appears in the short term.

2.5 Discussion

2.5.1 Comparison between SAF Estimation Methods

In Section 2.1, we briefly described three categories of estimation methods for SAF. Prevalence-based methods depend heavily on smoking prevalence history. Since the lag between smoking prevalence and SAF is usually around two or three decades, in order to use smoking preva-

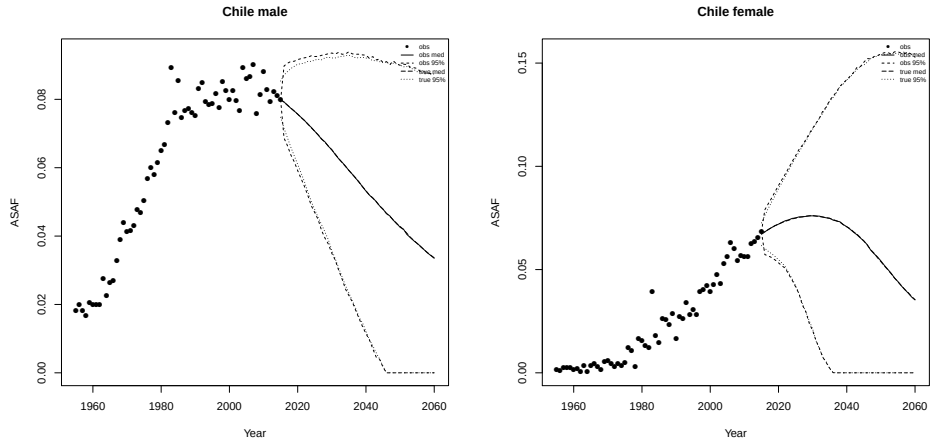


Figure 2.9: ASAF projection of Chile. The left and right panels show the projection of ASAF up to 2050 under the proposed model for male and female respectively. The solid and long dashed lines show the posterior median of projected observed ASAF and true ASAF respectively. The dashed and dotted lines represent 95% prediction intervals for observed ASAF and true ASAF respectively.

lence to estimate and predict SAF, especially for those countries whose onset of SAF is before 1950, one needs data at least back to the 1920s or 1930s. However, such smoking prevalence history is not available for most countries, and reconstruction of such data is challenging. Ng et al. (2014) provided estimates of smoking prevalence for many countries only from 1980 onwards.

Insufficient historical data is a major obstacle to using smoking prevalence for estimation and projection of SAF, and with currently available historical data, the predictive power using smoking prevalence data is not very high. In addition, smoking prevalence only reveals one aspect of the smoking epidemic, which cannot capture other aspects such as smoking intensity and duration and thus has been argued to be a poor indicator of the smoking exposure of the population (Shibuya et al., 2005; Luo et al., 2018). Prevalence-based estimation and projection have generally been applied only to specific countries on an individual basis, and

examples include Taiwan (Wen et al., 2005) and the United States (Ma et al., 2018).

There are two main indirect methods used widely in the literature, which both use the lung cancer mortality rate as an indicator for the accumulated hazard of smoking. The first one is the Peto-Lopez method which we have used here. This has been widely used in the demographic literature, in part because its data requirements are relatively modest. It has been validated in many studies (Preston et al., 2009; Kong et al., 2016).

One drawback of the Peto-Lopez method is that it uses the CPS-II to estimate the relative risk. Since the CPS-II was conducted in 1982 with volunteer participants only from the United States and most of them were middle-class, the CPS-II might not be fully representative and may potentially underestimate lung cancer mortality in nonsmokers (Tachfouti et al., 2014). Moreover, the Peto-Lopez method assumes that the relative risk is constant over time and homogeneous across nations. Mehta and Preston (2012), Teng et al. (2017), and Lariscy et al. (2018) have shown that the risks from smoking are changing over time. Also, in China and India, the lung cancer mortality rate among nonsmokers is higher than that of the developed countries such as that in the CPS-II (Liu et al., 1998; Gajalakshmi et al., 2003). Another issue is that the original Peto-Lopez paper reduced the smoking excess risk of each cause-of-death except lung cancer by 50% to control for other confounders. As stated in their paper, this reduction is somewhat arbitrary. To avoid some of these issues, here we have used only data from clear-pattern countries, which avoids some countries for which the Peto-Lopez method may not give good estimates.

Some variants of the Peto-Lopez method have been proposed. For example, Ezzati and Lopez (2003) reduced the correction factor for excess risk from 50% to 30% for all countries and extended this method to less developed countries by estimating the non-smoker lung cancer mortality rate based on household use of coal in poorly-vented stoves. They also provided an analysis of uncertainty. Mackenbach et al. (2004) used a simplified version which only used the all-cause relative risk in the CPS-II study and avoided calculations for the nine disease categories separately. Janssen et al. (2013) used this version to calculate age-specific SAF to partition mortality into smoking and non-smoking attributable parts,

and projected them separately.

Muszyńska et al. (2014) and Stoeldraijer et al. (2015) used the same method to calculate an age-standardized SAF, whose purposes are to compare the role of smoking in different regions of Poland, and to estimate and compare smoking attributable fraction of mortality among England & Wales, Denmark and the Netherlands, respectively. While age-standardization is used mainly to compare SAF among different populations, ASAF provides the all-cause SAF with all age-groups aggregated and is the main quantity reported in the iterature, e.g., Peto et al. (1992, 1994, 2006); Preston et al. (2009).

Based on these concerns about the Peto-Lopez method, Preston et al. (2009) and Preston et al. (2011) came up with the PGW method, which used a regression-based method to connect lung cancer mortality rate with other causes of death mortality rate instead of using the CPS-II. The PGW method avoids the relative risk problem faced by the Peto-Lopez method and provides estimates of uncentainty. However, its authors stated that the Peto-Lopez method might work better for countries where the cause-of-death structure is very different from that observed in developed countries, such as tropical African countries. They also pointed out that both methods would not work well for countries whose lung cancer mortality rate is also influenced largely by some other factors such as air pollution. As discussed by Preston et al. (2009), the PGW method produces similar estimates to the Peto-Lopez method in general for both males and females.

2.5.2 Projection Methodology

To our knowledge, there are only two other methods available for projecting SAF based on the Peto-Lopez method. Janssen et al. (2013) proposed the first method to forecast age-specific SAF and to our knowledge it has so far been applied only to the Netherlands. For projecting male age-specific SAF, a constant decline rate (-1.5%) based on the current trend of all-age combined SAF is applied for each age group. For females, it first estimates the time-to-peak and value of peak of female SAF. It uses age-period-cohort (APC) analysis to find the cohort with the highest lung cancer mortality rate and then adds 68, which is the

average age of dying from lung cancer, to that cohort to estimate the year which the all-age combined female SAF would reach the maximum. Then the difference between year-to-peak of male and female SAF with all ages combined is estimated and applied to get the time-to-peak and thus the age-specific female SAF. Finally, the rate of decline of female age-specific SAF is set to the same as that of the male.

The other method proposed for projecting SAF is to first estimate and project lung cancer mortality rate by considering the cohort effect, and use it to calculate the age-specific SAF. Stoeldraijer et al. (2015) used an APC model to estimate and forecast the lung cancer mortality rate of three countries: England & Wales, Denmark, and the Netherlands. For female data, they first estimated the time-to-peak for each age group by assuming that the time-to-peak of age-specific lung cancer mortality rate for females is when it reaches the corresponding rate for males. By assuming that the female lung cancer mortality will follow the same increasing-leveling-declining time trend as that for males for each age group, the authors argued that their method could provide long-term projections of lung cancer mortality rate, while previous work which only used historic trends in APC analysis could only provide short-term projections.

APC analysis is widely used, but it is also plagued by the unidentifiability issue resulting from the perfect linear relationship between the three effects. To resolve this requires extra constraints on the parameter space, many of which are not desirable (Luo, 2013; Smith and Wakefield, 2016). Also, projection of the future lung cancer mortality rate also requires the projection of age, period, and cohort effects, which introduces additional projection error, even more so for young cohorts for which historical data are not available.

Another way to resolve the unidentifiability issue in APC analysis is by introducing cohort explanatory variables (Smith and Wakefield, 2016). Cohort smoking history is one such powerful tool for estimating and projecting mortality. Preston and Wang (2006) and Wang and Preston (2009) used the average year of smoking before 40 of a cohort as a covariate to explain the mortality differences between genders and forecasted mortality of United States for both genders up to 2035. Shibuya et al. (2005) and Luo et al. (2018) used

APC anlaysis with selected smoking covariates such as cigarette tar exposure to estimate and project the lung cancer mortality rate. Cohort smoking history is a powerful tool, but it requires additional data (Burns et al., 1997) that are not available for many of the countries we considered.

2.5.3 China and India

According to Reitsma et al. (2017), China and India are the two countries that have seen the largest percentage increase in smoking prevalence. As a result, the ASAF for these two countries is important for understanding and projecting the world trend of the effect of smoking on mortality since the diffusion of the smoking epidemic from developed countries to developing countries has already started.

Parascandola and Xiao (2019) found that smoking-related health issues in China have increased over the past two decades, and the trend resembles the early pattern observed in high income countries such as the US and Japan. Smoking prevalence among Chinese men has remained high (around 60%) since the 1980s, with a modest decrease to 52% by 2015. Smokers born after 1970 tended to start smoking earlier and more intensely than those born before 1970.

Chen et al. (2015) analyzed two nationwide prospective cohort studies on smoking conducted in China during 1991-99 and 2006-14. They found that the excess risk among smokers almost doubled over the 15-year period. They reported that the SAF of males aged 40-79 increased from 11% in the first study to 18% in the second study, and they predicted that it would be over 20% in the mid-2010s.

In contrast, female smoking prevalence decreased from 7% in the 1980s to 3% in 2015 (Parascandola and Xiao, 2019). However, second-hand smoking remains high among Chinese females. Zheng et al. (2018) estimated that 65% of Chinese female non-smokers were exposed to second hand smoking in 2012. Nonetheless, the SAF for Chinese females aged 40-79 years was around 3% in 2006-14.

There are also substantial geographic differences in smoking prevalence. In big cities like

Beijing and Shanghai, smoking control measures have developed more rapidly than in other areas.

India has become the country with the second largest cigarette consumption in the world, after China. Smoking, including manufactured cigarettes, bidis, and chewing tobacco is one of the major causes of death for middle-aged Indians. Mishra et al. (2016) estimated that smoking prevalence among male Indians aged 15-69 years declined modestly from 27% in 1998 to 24% in 2010, while smoking prevalence among young adults aged 15-29 years rose.

We have not included these two countries in our analysis for the following two reasons. Firstly, we do not have enough data to estimate the ASAF for China and India. Even though there are some records of lung cancer death count data in the WHO Mortality Database for China (World Health Organization, 2017), these are only regional data and so could be biased. India has a reasonably good vital registration system but it also has lung cancer mortality data only for selected regions and locations.

Secondly, as pointed out by Preston et al. (2009), neither the Peto-Lopez original method nor the PGW method will provide reliable estimates of SAF for countries like China since smoking is not the only major factor that can cause lung cancer. The main assumptions of the Peto-Lopez and PGW methods are that lung cancer mortality is primarily caused by smoking and that the lung cancer mortailty rate is very low among nonsmokers. Therefore, as proposed by Ezzati and Lopez (2003) and others, some extra covariates such as household use of coal in poorly-vented stoves are used to adjust the estimates. Incorporating China and India in the joint model could be feasible in the future if better ASAF estimation methods and related data become available.

2.5.4 Decision-making and covariates

A main purpose of our method is to help improve mortality forecasts. One could also ask whether our approach could be used directly for policy-making. One possible use would be to provide a baseline forecast of what would happen with a continuation of current trends in general health, development and tobacco control measures. This could help to assess the

effectiveness of additional policies in accelerating the decline of smoking-related mortality. This could be done retrospectively, by considering a time point in the past at which a new tobacco control policy was introduced, and then comparing the probabilistic forecast based on data up to that point with what actually happened.

To do this prospectively would require the addition of covariates to the model. This is challenging, and would be a good topic for further research. A difficulty with forecasting using covariates is that the covariates themselves need to be forecast, and the covariates can be harder to forecast than the quantity being forecast. This is especially the case when, as here, the quantity being forecast has a strong time trend, and thus may well itself be easier to forecast than the covariates. In this situation, adding covariates can lead to forecasts that are noisier. This is one reason why, after decades of research, the majority of demographic studies do not use covariates in forecasting demographic quantities.

Chapter 3

ACCOUNTING FOR SMOKING IN FORECASTING MORTALITY AND LIFE EXPECTANCY

3.1 Introduction

Forecasting human mortality and life expectancy is of considerable importance for public health policy, planning social security systems, life insurance, and other areas, particularly as the world's population continues to age. It is also a major component of population projections, as it impacts the number of people alive and their distribution by age and gender. Population projection are themselves a major input to government planning at all levels, as well as private sector planning, monitoring international development and environmental goals, and research in the health and social sciences.

Many methods for forecasting mortality have been developed. The Lee-Carter method (Lee and Carter, 1992) for forecasting age-specific mortality rates was a milestone and has developed rapidly since it was proposed. Lee and Miller (2001) modified the Lee-Carter method by matching estimated life expectancy with the observed value. Other variations of the Lee-Carter method include adding a cohort effect (Renshaw and Haberman, 2006), applying a functional data approach (Hyndman and Ullah, 2007; Shang, 2016), and incorporating biomedical information (Janssen et al., 2013). Bayesian Lee-Carter methods have also been proposed (Pedroza, 2006; King and Soneji, 2011; Wiśniowski et al., 2015). See Booth et al. (2006) for a review.

The main organization that produces regularly updated mortality and population forecasts for all countries is the United Nations, which publishes these forecasts every two years in the *World Population Prospects* (United Nations, 2017). Traditionally since the 1940s, population projections have been done using deterministic methods that do not primarily

use statistical estimation methods or assess uncertainty in a statistical way (Whelpton, 1936; Preston et al., 2000a). In 2015, in a major advance, the UN changed the method for producing their official mortality and population forecasts from the traditional deterministic method to a Bayesian approach that estimates and assesses uncertainty about future trends in a principled statistical way using Bayesian hierarchical models for life expectancy and fertility (Raftery et al., 2012, 2013, 2014a; United Nations, 2015).

The basic approach of these methods is to extrapolate past trends in observed mortality rates, which have been dominated by a monotone increasing trend in life expectancy for over a century. However, it may also be helpful to include risk factors that can impact health, and hence mortality (Janssen, 2018). This has been done, for example, for the HIV/AIDS epidemic (Godwin and Raftery, 2017), alcohol consumption (Trias Llimós and Janssen, 2019), and the obesity epidemic (Vidra et al., 2017). Another major factor is smoking, which is mainly responsible for lung cancer and is a risk factor for many other fatal diseases, and causes about 6 million deaths per year (Britton, 2017). Smoking can account for some nonlinear trends, cohort effects, and between-country and between-gender differentials observed in mortality, suggesting that it could be used to improve mortality and life expectancy projections (Bongaarts, 2014).

Here we propose a Bayesian method for doing this for both genders and multiple countries jointly. It uses the smoking attributable fraction (SAF) of mortality, estimated by the Peto-Lopez method (Peto et al., 1992; Bongaarts, 2006; Janssen et al., 2013; Stoeldraijer et al., 2015). The proposed method consists of two main components, one to forecast the age-specific SAF (ASSAF), and the other to forecast non-smoking life expectancy. Our method develops male and female forecasts jointly, since the female smoking epidemic tends to resemble the male one, but with a lag, and possibly a different maximum level, a fact that can be used to improve forecasts. The female advantage in life expectancy is partly due to smoking effects, and our method quantifies this and uses it to forecast the future life expectancy gap between females and males. We apply our method to over 60 countries with high quality data on the historical impact of smoking on mortality.

The chapter is organized as follows. The methodology is described in Section 3.2. Section 3.2.3 describes the method for estimating and forecasting the ASSAF. Section 3.2.4 presents the estimation and forecasting method for non-smoking life expectancy. Section 3.2.5 describes our model for the gap between male and female life expectancy to complete the coherent projection. An out-of-sample validation experiment is reported in Section 3.3 to evaluate and compare the projection accuracy and calibration of our model with several benchmark methods. We then study the details of the forecast results for four selected countries in Section 3.4. We conclude with a discussion in Section 3.5.

3.2 Method

3.2.1 Notation

We use indices ℓ for country (always as a superscript unless otherwise indicated), s for gender, t for time (usually in terms of the year), and c for cohort (usually in terms of the year of birth). We use x to denote the left end of an age group, i.e., x represents the a-year age group $[x, x+a)$, and $x+$ represents the age group $[x, +\infty)$.

A key general concept in our approach is the smoking attributable fraction (SAF) of mortality for a population of interest. This is defined as the proportion by which mortality would be reduced if the population were not exposed to smoking. We focus on the age-specific SAF (ASSAF) of mortality for age group x in country ℓ and time period t, denoted by $y^{\ell}_{x,t}$. The all-age smoking attributable fraction (ASAF) of mortality is defined as a weighted average of the ASSAF over all age groups, where the weights are the age-specific mortality rates. We use the symbols m, e_0, and e_0^{NS} to denote the mortality rate, the life expectancy at birth, and the non-smoking life expectancy at birth, respectively.

We denote by $\mathcal{N}_{[u,v]}(\lambda, \kappa)$ the truncated normal distribution with mean λ and variance κ on the support $[u, v]$ (the subscript $[u, v]$ is omitted if supported on the whole real line), by $\mathcal{G}(\lambda, \kappa)$ the Gamma distribution with mean λ/κ and shape parameter κ, by $\mathcal{IG}(\lambda, \kappa)$ the inverse-Gamma distribution with mean $\kappa/(\lambda - 1)$ and shape parameter κ, and by $\mathcal{U}_{[u,v]}$ the

continuous uniform distribution on the support $[u, v]$. We denote the cardinality of a set $\mathcal{A}$ by $|\mathcal{A}|$ and the absolute value of a number b by $|b|$. A truncated function is written as $b_+ := \max\{b, 0\}$.

3.2.2 Data

To calculate the ASAF and ASSAF, we need annual death counts by country, age group, gender, and cause of death from the WHO Mortality Database (World Health Organization, 2017), which covers data from 1950 to 2015 for more than 130 countries and regions around the world. This dataset comprises death counts registered in national vital registration systems and is coded under the rules of the International Classification of Diseases (ICD). Quinquennial population, mortality rates, and life expectancy at birth were obtained from the 2017 Revision of the *World Population Prospects* (United Nations, 2017) for each country, gender, and age group.

3.2.3 Age-specific Smoking Attributable Fraction

We use estimates of the smoking attributable fraction (SAF) obtained with the Peto-Lopez method, an indirect method based on the observed lung cancer count data (Peto et al., 1992; Kong et al., 2016; Li and Raftery, 2019). Here we use a modified version of the Peto-Lopez method proposed by Rostron and Wilmoth (2011) to estimate the ASSAF. The modified method calculates the ASSAF for all 5-year age groups from 35 to 100, which is finer than the original Peto-Lopez method. Also, the reference lung cancer mortality rates used in the original Peto-Lopez method were underestimated because of selection bias, and the modified method addresses this by introducing an inflation factor. Because of data quality issues, we set ASSAF for age groups less than 40 to 0, and ASSAF for age groups 85 and older to the same value as that for the 80–84 age group. These rules follow the guidelines in Peto et al. (1992) and Rostron and Wilmoth (2011) with minor modifications, and result in nine age groups with non-zero ASSAF. The left panel of Figure 3.1 shows the estimated quinquennial ASSAF of US males for all nine age groups (shown in different colors) from 1953 to 2013.

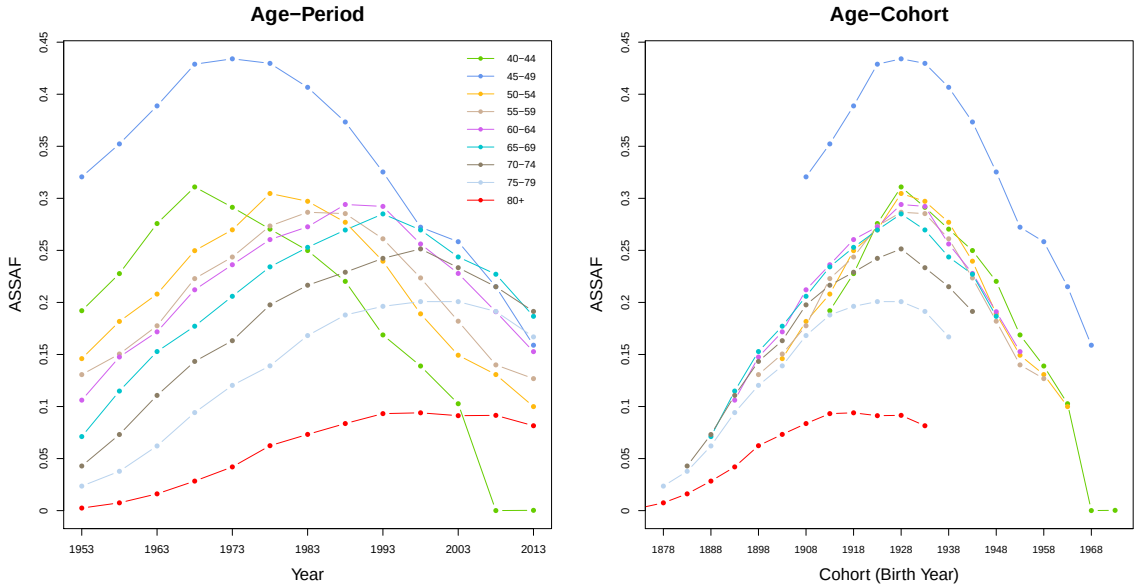


Figure 3.1: Age-specific smoking attributable fractions (ASSAF) for the male population in the United States from 1950-2015. Left: Age-period plot. The horizontal axis is the year of observation and colors differentiate age groups. Right: Age-cohort plot. The horizontal axis is the year of birth for all cohorts, where the values for each age group are shown by a different color.

Estimation and Forecasting: Age-cohort Modeling

We propose a probabilistic age-cohort approach to estimate and forecast the ASSAF for the male population. The age-cohort plot of the US male ASSAF (right panel) in Figure 3.1 has two main features that lead to our modeling. First, the ASSAF can be well approximated by the product of an age effect and a cohort effect. The ASSAF of age group 80+ tends to shift horizontally from other age groups for most of the countries (e.g., see the red dashed line in the age-cohort plot of Figure 3.1 for the case of US males). Hence, we apply a cohort

effect τ for all age groups less than 80, and a separate cohort effect $\tilde{\tau}$ for the 80+ age group.

The probabilistic model of ASSAF in country ℓ is

$$y_{x,t}^{\ell} \overset{\text{ind}}{\sim} \mathcal{N}(\xi_x^{\ell}\tau_{t-x}^{\ell}\mathbf{1}_{x\neq 80} + \xi_x^{\ell}\tilde{\tau}_{t-x}^{\ell}\mathbf{1}_{x=80},\ \sigma_{\ell}^2), \tag{3.1}$$

where x takes values in $\{40, 45, 50, 55, 60, 65, 70, 75, 80\}$. To ensure identifiability, we set $\xi_{40}^{\ell} = 1$ for all countries. Eq. 3.1 is also closely related to a low-rank matrix completion method. The age-cohort matrix based on the observed period ASSAF inevitably contains missing values since we do not observe the ASSAF of early cohorts at young ages or that of late cohorts at old ages (see Figure 3.2).

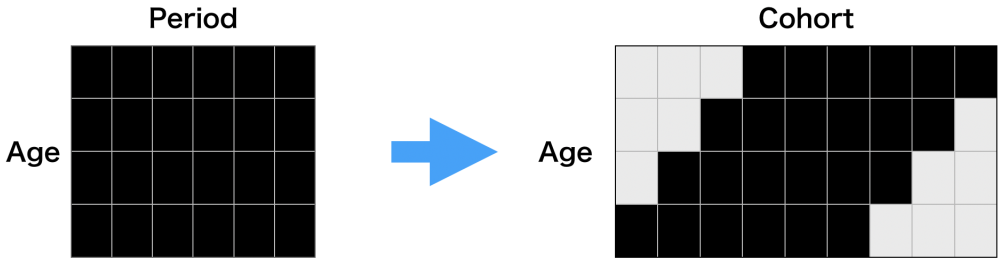


Figure 3.2: Transformation from age-period matrix (left) to age-cohort matrix (right). Black and grey cells represent observed and missing values, respectively.

Second, the cohort pattern of the male ASSAF has a strong increasing-peaking-declining pattern. This trend can be well captured by a five-parameter double logistic function (Meyer, 1994):

$$g(c|\theta) := \frac{k}{1+\exp\{-\triangle_1(c-1873-\triangle_2)\}} - \frac{k}{1+\exp\{-\triangle_3(c-1873-\triangle_2-\triangle_4)\}}, \tag{3.2}$$

where $\theta := (\triangle_1, \triangle_2, \triangle_3, \triangle_4, k)$. The double logistic curve is a flexible parametric curve, which has been used in many scientific fields such as hematology, phenology, and agricultural

science. Due to its scientific interpretability, it is often used to describe social change, diffusion, and substitution processes (Grübler et al., 1999; Fokas, 2007; Kucharavy and De Guio, 2011). Examples of the use of a double logistic curve to describe dynamics in human demography include mortality rates (Marchetti et al., 1996), life expectancy at birth (Raftery et al., 2013), and total fertility rates (Alkema et al., 2011).

Most developed countries have already entered the declining stage of the smoking epidemic. The epidemic started in the early 1900s with a steady increase until the 1950s-60s when the adverse impact of smoking became widely known and anti-smoking measures started to be put in place. Since then, the smoking epidemic has continued to decline. Thus the cohort effect of smoking exhibits a similar increasing-peaking-decreasing trend, which can be captured naturally by the double logistic curve.

The cohort effect $\tilde{\tau}$ for ages 80+ is just a horizontal shift of the cohort effect τ for younger ages, so we use two related double logistic curves to bridge them:

$$\tau_c^\ell|\theta^\ell, \sigma^{2[\tau]} \overset{\text{ind}}{\sim} \mathcal{N}(g(c|\theta^\ell),\ \sigma^{2[\tau]}), \qquad \tilde{\tau}_c^\ell|\tilde{\theta}^\ell, \sigma^{2[\tau]} \overset{\text{ind}}{\sim} \mathcal{N}(g(c|\tilde{\theta}^\ell), \sigma^{2[\tau]}), \tag{3.3}$$

where $c := t - x$, $\theta^\ell := (\triangle_1^\ell, \triangle_2^\ell, \triangle_3^\ell, \triangle_4^\ell, k^\ell)$, and $\tilde{\theta}^\ell := (\triangle_1^\ell, \triangle_2^\ell, \triangle_3^\ell, \triangle_4^\ell + \delta^\ell, k^\ell)$. Here δ^ℓ is a shift parameter controlling the amount of horizontal translation $\tilde{\tau}$ can make with respect to τ.

We use a three-level Bayesian hierarchical model (BHM) to estimate and forecast male ASSAF for all countries of interest jointly. Level 1 models the observed male ASSAF in terms of the tensor product of the age effect and the cohort effect (i.e., Eq. 3.1). Level 2 models the distributions (conditioning on the global parameters) of the country-specific age effect ξ_x^ℓ, the country-specific cohort effects τ_c^ℓ and $\tilde{\tau}_c^\ell$ in Eq. 3.3, the country-specific parameters θ^ℓ and $\tilde{\theta}^\ell$ of the double logistic function, and the country-specific measurement variance σ_ℓ^2. Level 3 sets hyperpriors on the global parameters
$\psi := (\{\mu_x^{[\xi]}\}_{x\neq 40},\ \{\sigma_x^{2[\xi]}\}_{x\neq 40},\ \sigma^2,\ \sigma^{2[\tau]},\ \mu_{\triangle_1},\ \mu_{\triangle_2},\ \sigma_{\triangle_2}^2,\ \mu_{\triangle_3},\ \mu_{\triangle_4},\ \sigma_{\triangle_4}^2,\ \mu_k,\ \sigma_k^2,\ \mu_\delta,\ \sigma_\delta^2)$. More details of the specification of the full model are given in the Appendix B.1.

The left and right panels of Figure 3.3 plot the cohort effects and age effect of US male

ASSAF, respectively. The estimated cohort effect τ for the age groups 45-79 shows a clear increasing-peaking-decreasing trend as observed in Figure 3.1. The estimated cohort effect $\tilde{\tau}$ for the 80+ age group shows the same trend for the 13 cohorts reaching age 80 by 2015. We could forecast any cohort effects based on the posterior distribution of the double logistic function. The estimated age effect indicates that the smoking-attributed fraction of mortality is higher among middle-aged males (aged 40–69) in the US than among older males (70 and over). Figure 3.4 plots the posterior distributions of the means of the US male ASSAF for all 9 age groups and all 21 cohorts.

To project the future ASSAF, we first generate future cohort effects by plugging samples drawn from the posterior distributions of country-specific parameters θ^ℓ and $\tilde{\theta}^\ell$ in Eq. 3.2 and 3.3. Then, we apply Eq. 3.1 using samples drawn from posterior distributions of the future cohort effects, age effect, and country-specific variance σ_ℓ^2 to get projections of ASSAF.

3.2.4 Non-smoking Life Expectancy

The non-smoking life expectancy at birth, e_0^{NS}, is the life expectancy at birth that a population would have if no one smoked, but all mortality risks were otherwise the same (Bongaarts, 2006). To estimate e_0^{NS}, we need the age-specific mortality rates d_x and the ASSAF y_x described in Section 3.2.3. As in the last section, all quantities described in this section are specific to the male population, and the gender index s is omitted unless otherwise specified.

The calculation of e_0^{NS} consists of two steps. First, the age-specific non-smoking attributable mortality rate for a given country ℓ, age group x, and period t (denoted by $m_{\ell,x,t}^{NS}$) is calculated as

$$m_{\ell,x,t}^{NS} := (1 - y_{x,t}^{\ell}) \cdot m_{x,t}^{\ell}. \tag{3.4}$$

Second, we convert the set of $m_{\ell,x,t}^{NS}$ to e_0^{NS} using the standard period life table method (Preston et al., 2000a, Chapter 3), as implemented in the *life.table* function in the R package `MortCast` (Ševčíková et al., 2019a). Figure 3.5 shows the relationship between quinquennial e_0 and e_0^{NS} for US males and Netherlands males from 1950 to 2015, respectively. The vertical

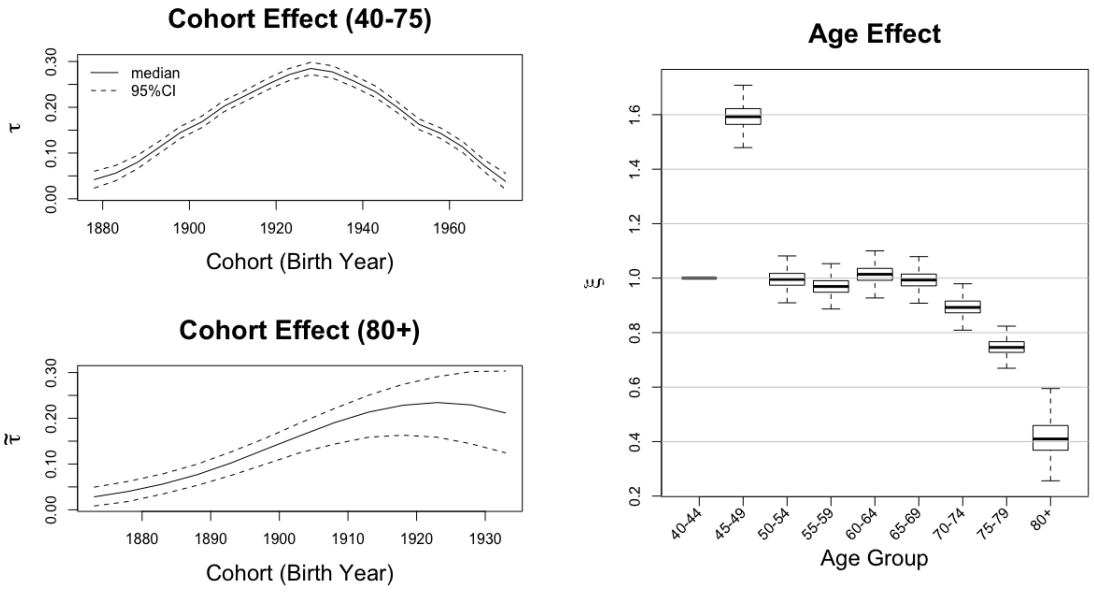


Figure 3.3: Posterior distributions of cohort and age effects of United States male ASSAF. Top Left: posterior median and 95% credible intervals of the cohort effects τ for the 40–79 age groups. Bottom Left: posterior median and 95% credible intervals of the cohort effect $\tilde{\tau}$ for the 80+ age groups. Right: boxplot of posterior distribution of the age effect.

gap between e_0 and e_0^{NS} at each time point presents the years of life expectancy lost due to smoking. The changes in the gaps also follow a similar increasing-peaking-decreasing trend over the period 1950 to 2015.

Estimation and Forecasting: Non-linear Life Expectancy Gain Model

We forecast e_0^{NS} by investigating the nonlinear five-year gains of e_0^{NS}. As discussed by Raftery et al. (2013), the improvement of gains on e_0 for most of the countries has experienced a slow-rapid-slow increasing pattern and a six-parameter double logistic function is used to

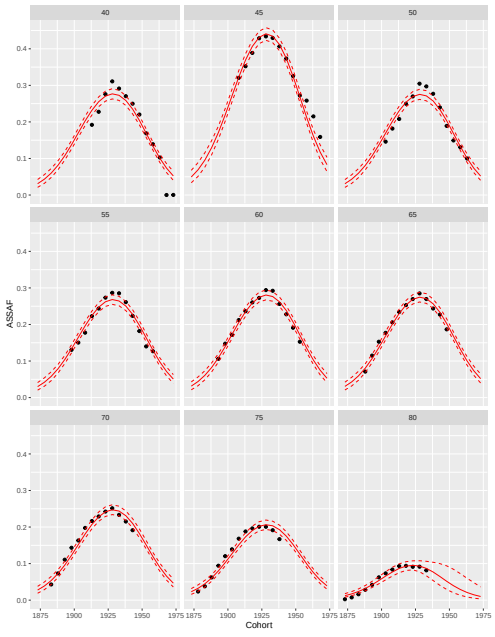


Figure 3.4: Posterior distributions of the means of US male ASSAF for all 9 age groups. The observed ASSAF is shown by black dots. The posterior median and 95% credible intervals of the means are shown by solid and dashed red lines, respectively.

capture the non-linearity of five-year gains of e_0:

$$\tilde{g}(e_0|\zeta) := \frac{w}{1+\exp\{-\frac{4.4}{a_2}(e_0 - a_1 - 0.5a_2)\}} + \frac{z-w}{1+\exp\{-\frac{4.4}{a_4}(e_0 - \sum_{i=1}^{3} a_i - 0.5a_4)\}}, \quad (3.5)$$

where $\zeta := (a_1, a_2, a_3, a_4, w, z)$ and z is the asymptotic average rate of increase in e_0. We assume that z is nonnegative, implying that life expectancy will continue to increase on average (Oeppen and Vaupel, 2002; Bongaarts, 2006).

The five-year gains in e_0^{NS} exhibit this nonlinear pattern as well. The left panel of Figure 3.6 plots the observed five-year gains of e_0 (in grey dots) and e_0^{NS} (in red dots) for over 60 countries with data of high enough quality from 1950 to 2015. The five-year gains in e_0^{NS} have nearly the same shape as the five-year gains in e_0, which supports using the same double logistic function to model the gains. Also, e_0^{NS} has almost the same five-year gain at the

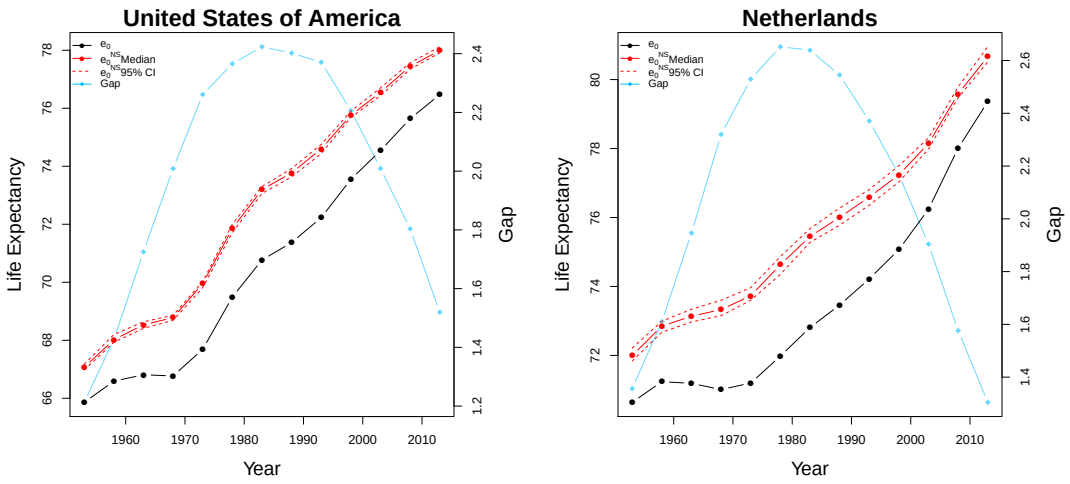


Figure 3.5: Male life expectancy at birth, e_0, and male non-smoking life expectancy at birth, e_0^{NS}, for the United States (left) and the Netherlands (right). The black line shows e_0. The solid red line and the dashed red lines show the posterior median and the 95% credible interval of e_0^{NS}. The blue line represents the gap between e_0 and e_0^{NS}.

highest age as e_0, suggesting that the asymptotic average rate of increase z for e_0^{NS} should be similar to that of e_0. Further, the variability of the five-year gains of e_0^{NS} changes from a low level to a high level of e_0^{NS}, which suggests including a nonconstant variance component in the model.

We use a three-level Bayesian hierarchical model for e_0^{NS}. Level 1 models $e_{0,\ell,t}^{NS}$ for country ℓ and period t by

$$e_{0,\ell,t}^{NS} \overset{\text{ind}}{\sim} \mathcal{N}(e_{0,\ell,t-1}^{NS} + \tilde{g}(e_{0,\ell,t-1}^{NS}|\zeta^{\ell}),\ (\omega^{\ell} \cdot \phi(e_{0,\ell,t-1}^{NS}))^2), \tag{3.6}$$

with country-specific parameters $\zeta^{\ell} := (a_1^{\ell}, a_2^{\ell}, a_3^{\ell}, a_4^{\ell}, w^{\ell}, z^{\ell})$. Here $\phi(\cdot)$ is a regression spline fitted to the absolute residuals resulting from the model with constant variance in Eq. 3.6 with the same estimation method described later. The regression spline is used to account for the changing variability of the observed data. The right panel of Figure 3.6 illustrates the

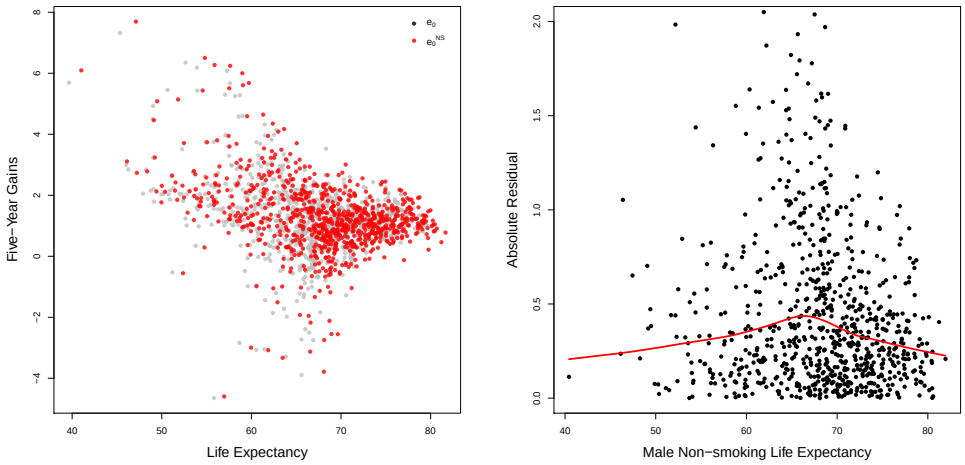


Figure 3.6: Left: Five-year gains of e_0 and e_0^{NS} for over 60 countries from 1950 to 2015. The gains in e_0 and e_0^{NS} are represented using grey and red dots, respectively. Right: Plot of absolute residuals estimated from the constant variance model against life expectancy shown by black dots, with fitted regression spline shown by the red line.

varying absolute residuals with the fitted spline in red. Level 2 specifies the conditional distribution for all country-specific parameters including ζ^ℓ and ω^ℓ. Level 3 sets the hyperpriors for the global parameters $\tilde{\psi} := (\{\mu_{a_i}\}_{i=1}^4, \{\sigma_{a_i}^2\}_{i=1}^4, \mu_w, \sigma_w^2, \mu_z, \sigma_z^2)$. The full specification of the model is given in the Appendix B.1.

To produce a probabilistic forecast, we sample from the joint posterior distributions of the country-specific parameters ζ^ℓ to calculate the five-year gains $\tilde{g}(e_0^{NS})$ together with the posterior distributions of ω^ℓ. For the variance component, we evaluate $\phi(e_{0,\ell,t-1}^{NS})$ if $e_{0,\ell,t-1}^{NS}$ is within the range of the fitted data; otherwise, it is set equal to the spline value evaluated at the largest observed e_0^{NS}. We then use Eq. 3.5 and 3.6 to generate samples from the posterior predictive distribution for future country-specific $e_{0,\ell,t}^{NS}$. The set of samples approximates the

posterior predictive distribution.

3.2.5 Male-Female Joint Forecast

Male e_0 Forecast

First, we use the coherent Lee-Carter method (Li and Lee, 2005; Ševčíková et al., 2016) to convert the projected $e^{NS}_{0,\ell,t}$ back to $m^{NS}_{\ell,x,t}$ for all age groups x at period t of country ℓ. Then, we invert Eq. 3.4 to get the projected age-specific all-cause mortality, i.e., $m^{\ell}_{x,t} = m^{NS}_{\ell,x,t}/(1-y^{\ell}_{x,t})$ for any age groups x, period t, and country ℓ. Finally, applying the same life table method described in Section 3.2.4 to the forecast $m^{\ell}_{x,t}$, we obtain the forecast life expectancy at birth for period t and country ℓ. Figure 3.7 illustrates the projections of e^{NS}_0 and e_0 for US and the Netherlands males to 2060. The projected e_0 converges to the projected e^{NS}_0 as ASSAF decreases towards 0 for all age groups of US and the Netherlands males.

Female e_0 Forecast: Gap Model

We propose a gap model similar to that of Raftery et al. (2014b) to produce a coherent projection of male-female life expectancy at birth. It has been argued that differences in smoking largely account for the life expectancy gap between males and females (Preston and Wang, 2006; Wang and Preston, 2009). Here we explore the relationship between the between-gender gap in life expectancy and the between-gender gap in the all-age smoking attributable fraction (ASAF). The ASAF is a single statistic summarizing the smoking effect on mortality and is defined as a weighted average of the ASSAF values as calculated in Section 3.2.3, where the weights are the age-specific mortality rates. Li and Raftery (2019) describe the estimation of ASAF, as well as a method for forecasting it using a four-level Bayesian hierarchical model.

We modify the gap model of Raftery et al. (2014b) by adding the country-specific between-

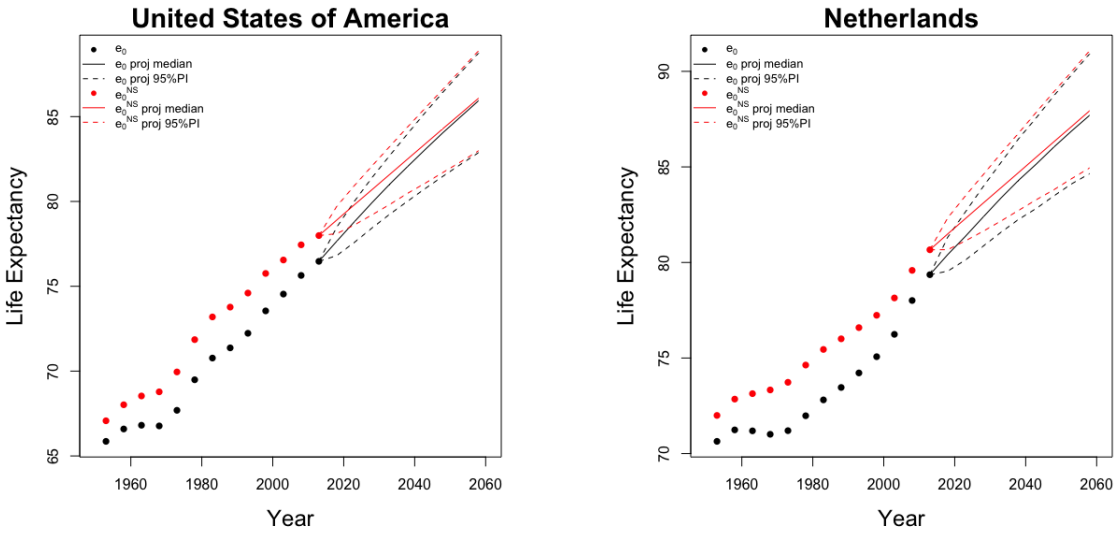


Figure 3.7: Projections of e_0^{NS} and e_0 of US (left) and the Netherlands (right) males to 2060. The posterior medians and the 95% predictive intervals of projected e_0^{NS} are shown by solid and dashed red lines, respectively. The posterior medians and the 95% predictive intervals of projected e_0 are shown by solid and dashed black lines, respectively.

gender ASAF gap as a covariate. The proposed gap model is as follows:

$$G_t^\ell := \min\{\max\{\tilde{G}_t^\ell, L\}, U\} \tag{3.7}$$
$$\tilde{G}_t^\ell \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 e_{0,m,1953}^\ell + \beta_2 G_{t-1}^\ell + \beta_3 e_{0,m,t}^\ell + \beta_4 (e_{0,m,t}^\ell - \varpi)_+ + \beta_5 h_t^\ell,\ \sigma_G^2),$$

where U and L are the observed historical maximum and minimum of the between-gender gap in e_0, ϖ is the level of male e_0 at which the gap is expected to stop widening, and h_t is the between-gender gap (male minus female) of the posterior median of ASAF in period t.

The estimated parameters of the model based on the data for over 60 countries for 1950–2015 are reported in Table 3.1. Our estimates indicate that the e_0 gender gap has a strong positive association with the ASAF gap after adjusting for other factors ($\hat{\beta}_5 = 1.180$ with

Table 3.1: Estimated gap model coefficients with standard errors in parentheses, if available.

Variable	Parameter	Estimate	Variable	Parameter	Estimate
Intercept	β_0	-2.173 (0.627)	h_t^ℓ	β_5	1.180 (0.384)
$e_{0,m,1953}^\ell$	β_1	0.012 (0.003)		σ_G	0.496
G_{t-1}^ℓ	β_2	0.901 (0.010)		ϖ	61
$e_{0,m,t}^\ell$	β_3	0.043 (0.011)		L	0.03
$(e_{0,m,t}^\ell - \varpi)_+$	β_4	-0.107 (0.012)		U	13.35
	R^2			0.933	

p-value < 0.01). Since the estimated lower bound of the life expectancy gap L is positive, our model guarantees that no crossover of male and female life expectancy forecasts will happen for all trajectories. The other coefficients have similar estimates and significance as in Raftery et al. (2014b), which accounts for the remaining variability in the between-gender life expectancy gap, possibly due to biological and other social factors (Janssen and van Poppel, 2015).

When performing projection, we forecast all terms in Eq. 3.7 forward. Instead of using a random walk as in Raftery et al. (2014b), we make use of the ASAF gap to guide our projection. However, we constrain the quantity $(e_{0,m,t}^\ell - \varpi)_+$ to be 20 when $e_{0,m,t}^\ell$ is greater than 81 years, which is the largest male e_0 observed in countries of interest up to 2015, since there is not enough information to determine whether the gap will continue to shrink for higher e_0. After the gender gap has been forecast, we add the gap to each posterior trajectory of the forecast male e_0 to get the full posterior predictive distribution of female e_0.

3.2.6 Estimation and Projection of the Full Model

We use data from over 60 countries for which the data on the male smoking-attributable mortality was of good enough quality. The precise data quality criteria and thresholds used are described in Li and Raftery (2019). Of these countries, two are in Africa, 16 are in the Americas, nine are in Asia, 40 are in Europe and two in Oceania. Estimation of the full model makes uses of male ASSAF, male age-specific mortality rates, both genders e_0, and both genders ASAF of all clear-pattern countries over 13 five-year periods during 1950–2015. Future e_0 of the same set of countries over 9 five-year periods from 2015 to 2060 is projected based on the joint posterior predictive distribution of the full model. The full procedure is described in the Appendix.

We use Markov Chain Monte Carlo (MCMC) to sample from the joint posterior distributions of the parameters of interest. For the BHM of the ASSAF, we ran three chains, each of length 100,000 iterations thinned by 20 iterations with a burn-in of 2,000. This yielded a final, approximately independent sample of size 3,000 for each chain. For the BHM of each of the 30 samples of e_0^{NS}, we ran one chain with length 100,000 iterations thinned by 50 with a burn-in of 1,000. This yielded a final, approximately independent sample of size 1,000 for each chain. We monitored convergence by inspecting trace plots and using standard convergence diagnostics, details of which are given in Appendix B.2. We include the plots of e_0 projections for over 60 countries and both genders in Appendix B.3.

3.3 Results

We assess the predictive performance of our model using out-of-sample predictive validation.

3.3.1 Study design

The data we used for out-of-sample validation cover the period 1950–2015, dividing it into an earlier training period and a later test period. We fit the model using only data from the training period, and then generated probabilistic forecasts for the training period. We

finally compared the probabilistic forecasts with the observations for the training period. We used two different choices of test period: 2000–2015, and 2010–2015. The former allows us to assess longer-term forecasts, while the latter focuses on shorter-term forecasts.

To assess the accuracy of the probabilistic forecasts, we define the gender-specific mean absolute error (MAE) as

$$\text{MAE}_s = \frac{1}{|\mathcal{L}||\mathcal{T}|} \sum_{\ell \in \mathcal{L}} \sum_{t \in \mathcal{T}} |\hat{e}^{\ell}_{0,s,t} - e^{\ell}_{0,s,t}|, \tag{3.8}$$

where $\mathcal{L}$ is the set of countries considered in the validation, $\mathcal{T}$ is the set of training periods, and $\hat{e}^{\ell}_{0,s,t}$ is the posterior median of the predictive distribution of life expectancy at birth at year t for country ℓ and gender s. To assess the calibration and sharpness of the model, we calculated the average empirical coverage of the prediction interval over the validation period, which we hope to be close to its nominal level with as short a halfwidth of the interval as possible (Gneiting and Raftery, 2007).

3.3.2 Out-of-sample validation

We evaluated and compared the performance of the proposed model with four commonly used methods for forecasting e_0: the Lee-Carter method (Lee and Carter, 1992), the Lee-Miller method (Lee and Miller, 2001), the Hyndman-Ullah functional data method (Hyndman and Ullah, 2007), and the Bayesian hierarchical model as implemented in the `bayesLife` R package (Raftery et al., 2013). We refer to the last as the bayesLife method. The first three methods were implemented using the corresponding functions with default settings in the `demography` R package (Booth et al., 2006; Hyndman et al., 2019). The bayesLife method was implemented under default settings using the R package `bayesLife` (Raftery et al., 2013, 2014b; Ševčíková et al., 2019b).

Table 3.2 gives the out-of-sample validation results for the four methods described above as well as our proposed method. Our method had the smallest MAE for both genders and both choices of test period among the five methods. For predicting one five-year period ahead, our method improved accuracy over the Lee-Carter method by 70% (60%), and over

the bayesLife method by 24% (11%) for males (females). For predicting three five-year periods ahead, the new method improved accuracy over the Lee-Carter method by 53% (40%), and over the bayesLife method by 24% (17%) for males (females).

For model calibration, the Lee-Carter-type models produced predictive intervals that are too narrow, thus underestimating the predictive uncertainty in the testing period. The bayesLife method and the new method produced predictive intervals with coverage close to the nominal level. We assess the sharpness of the forecast method using the 80% predictive interval halfwidth. For male data under the three five-year periods prediction, the 80% predictive interval of the new method was 30% shorter on average, but yielded the same empirical coverage as the bayesLife method. Under the one five-year out-of-sample predictions, the 80% predictive interval of the new method was 30% shorter on average but yielded even higher empirical coverage than the bayesLife method. For female data, the predictive intervals of our method overcovered the observations slightly for each choice of test period, but their median halfwidths were not much wider than those of the bayesLife method (e.g., the largest increment was less than 18%). The major source of variability in the female projections of the new method comes from the gap model.

3.4 Case studies

On average, smoking results in 1.4 years lost of male life expectancy at birth for over 60 countries over 1950-2015. The trend in years lost due to smoking also follows the pattern of the smoking epidemic. The average years lost due to smoking among males increased from 0.9 in 1953 to a maximum of 1.7 in 1993, and decreased to 1.3 in 2013.

For male populations of most countries, the ASSAF has already passed the peak for most age groups. When this is the case, accounting for the smoking effect leads to higher forecasts of life expectancy at birth. On average, our proposed method gives forecasts of male life expectancy at birth that are 1.1 years higher than the bayesLife method used by the UN for over 60 countries over the period 2015–2060.

Most female populations are still at the increasing or peaking stage of the smoking epi-

Table 3.2: Out-of-sample validation results for forecasting life expectancy at birth of males and females one and three five-year periods ahead. "Num" is the number of countries used in the validation. In the "Method" column, "H-U FDA" is the Hyndman-Ullah functional data analysis method, "bayesLife" represents the method described in Raftery et al. (2013), and "smokeLife" is the our proposed method. "Halfwidth" represents the median of the halfwidth of the prediction interval.

Period	Num	Gender	Method	MAE	Coverage		Halfwidth	
					80%	95%	80%	95%
Train:1950–2000	67	M	Lee-Carter	2.043	0.144	0.199	0.368	0.568
			Lee-Miller	1.536	0.318	0.418	0.831	1.239
			H-U FDA	2.206	0.189	0.274	0.808	1.259
			bayesLife	1.273	**0.741**	**0.950**	1.722	2.714
			smokeLife	**0.962**	**0.741**	0.896	1.197	1.943
Test: 2000–2015		F	Lee-Carter	1.210	0.199	0.294	0.391	0.599
			Lee-Miller	0.748	0.602	0.756	0.612	0.940
			H-U FDA	1.430	0.114	0.299	0.412	0.633
			bayesLife	0.876	**0.816**	**0.955**	1.312	1.985
			smokeLife	**0.718**	0.891	1.000	1.380	2.173
Train:1950–2010	68	M	Lee-Carter	1.741	0.103	0.118	0.306	0.448
			Lee-Miller	0.853	0.544	0.721	0.581	0.931
			H-U FDA	1.364	0.191	0.324	0.548	0.791
			bayesLife	0.688	**0.824**	0.897	1.098	1.748
			smokeLife	**0.523**	0.912	**0.985**	0.773	1.250
Test: 2010–2015		F	Lee-Carter	1.025	0.118	0.221	0.279	0.436
			Lee-Miller	0.486	0.662	0.779	0.476	0.708
			H-U FDA	0.895	0.250	0.368	0.373	0.573
			bayesLife	0.464	**0.868**	**0.941**	0.853	1.291
			smokeLife	**0.413**	0.971	1.000	0.974	1.517

demic. However, for 2055-2060, we expect to see an increment of 1.0 in female life expectancy compared to the forecast result from the bayesLife method, since the female smoking epidemic will be following the same decreasing trend as that of males by then.

We now study four countries in detail, representing different patterns of the smoking epidemic.

3.4.1 United States

The United States of America has one of the best vital registration systems in the world and also high quality data on cause of death. It thus has high quality data on the SAF. The smoking epidemic started in the early 1900s among the male population and rose to the historical maximum of around 60% in the 1950s. At that point, government programs and social movements against smoking began to develop, and the US public became increasingly aware of the adverse impacts of smoking. Since then, there has been a substantial decrease in smoking prevalence, going down to about 20% in the 1990s, and 17.5% in 2016 (Burns et al., 1997; Islami et al., 2015).

The female smoking epidemic started two decades later than the male one with a maximum prevalence of around 30% in the 1960s. Female smoking prevalence decline to about 20% in 1990s and 13.5% in 2016 (Burns et al., 1997; Islami et al., 2015). Figure 3.8a shows projections of the US male and female ASAF to 2060. Figure 3.8b predicts a continuously narrowing gap of the between-gender life expectancy due to the shrinking gap between male and female ASAF up to 2060.

Figures 3.8c and 3.8d show projections of male and female life expectancy for the period 2015-2060. The bayesLife method projects male life expectancy in 2055-2060 to be 84.0 years, with 95% predictive interval (79.2, 87.6). We project male life expectancy to be 86.1 in 2060, with 95% predictive interval (83.0, 88.9). The bayesLife method projects US female life expectancy for 2055-2060 to be 86.5 with 95% predictive interval (82.9, 90.0). We project female life expectancy to be 88.6 with interval (84.8, 92.4).

Our method gives forecasts of life expectancy that are about two years higher than

those from the bayesLife method for both males and females, because of accounting for the smoking effect. Our predictive interval for male life expectancy at birth is 29% shorter than the bayesLife one, while our female interval is comparable with that of the bayesLife method. Both of our 95% predictive intervals cover the posterior medians from the bayesLife method.

3.4.2 The Netherlands

The Netherlands is a western European country with a long history of the smoking epidemic, which can be dated back to the 1880s when the cigarette industry began there. Male smoking prevalence reached 90% in most age groups in the 1950s, but dropped rapidly to 30% in the 2010s. In contrast, smoking was more prevalent among females in the 1970s, when about 40% of female smoked, and after 1975 there was a sustained drop to 24% in the 2010s (Stoeldraijer et al., 2015).

Figure 3.9a shows that the female ASAF is forecast to surpass the male ASAF for the next two decades and by 2060, both male and female ASAF will be at about the same level. Figure 3.9b shows that the turning point in the between-gender gap of life expectancy happened around the 1990s, when the male ASAF had passed its peak and the female ASAF started to climb. With the shrinking of the ASAF gap, the projected life expectancy gap is forecast to continue to shrink and plateau around 2.8, due to biological and social factors (Janssen and van Poppel, 2015).

Both Dutch males and females experienced a period of stagnation in life expectancy gains—in the 1960s for males and the 1990s for females. Smoking is a major reason for this stagnation. The right panel of Figure 3.5 indicates that the forecast Dutch male life expectancy gain is more linear and sustained after removing the smoking effect. Figures 3.9c and 3.9d show projections of male and female life expectancy for 2015–2060. We project male life expectancy for the period 2055-2060 to be 88.0 years, with a 95% prediction interval of (85.0, 91.1), while the bayesLife method projects 86.1, with interval of (82.3, 89.7). We project female life expectancy for the period 2055-2060 to be 90.8, with a 95% prediction interval of (86.6, 95.0), while the bayesLife method projects 88.4 years, with interval of (85.1,

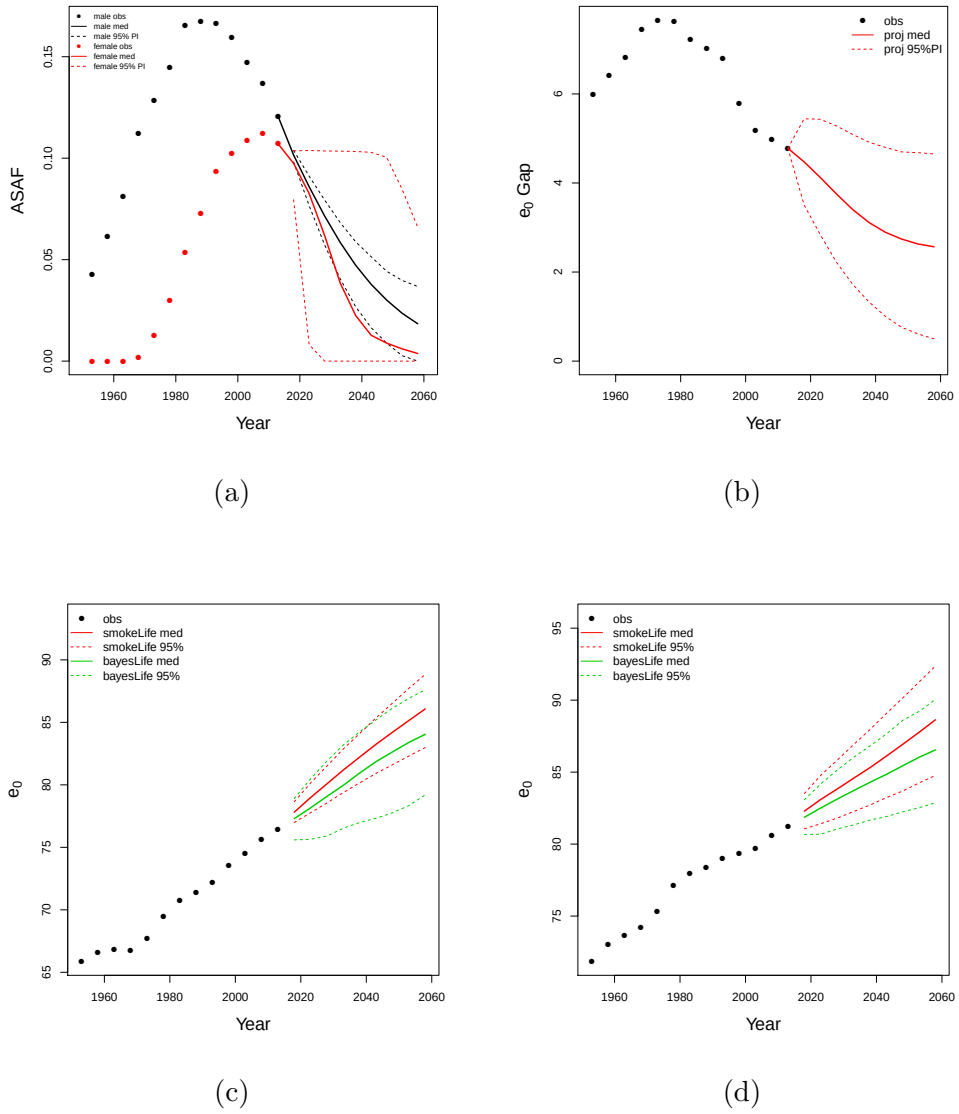


Figure 3.8: United States of America. (a) All-age smoking attributable fraction (ASAF) for male (black) and female (red) with median and 95% PI of posterior predictive distributions. (b) Between-gender gap of life expectancy at birth with posterior predictive median (red solid) and 95% PI (red dotted). (c) Forecasts of male life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted). (d) Forecasts of female life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted).

91.9).

Similarly to the US, our forecast of life expectancy in 2060 is about two years higher than a forecast that does not take account of smoking. By considering the decreasing trend of the smoking epidemic, our forecast is 1.9 years higher for males and 2.3 years higher for females expectancy compared with the bayesLife method. Janssen et al. (2013) forecast the Dutch male and female life expectancy in 2040 to be 84.6 years and 87.2 years respectively, taking account of smoking. This agrees well with our forecasts —85.0 for males and 87.3 for female—in 2040.

3.4.3 Chile

Chile is a South American country where the smoking epidemic had a late start, and it is currently one of the countries with the highest smoking prevalence in the Americas. Smoking prevalence decreased from 50% in 2000 to 40% in 2016 among males, and from 44% to 36% among females. This decline is modest compared to that in the United States (Islami et al., 2015).

Figure 3.10a shows the projections of male and female ASAF. Chilean male ASAF has been at the peaking stage for a long time, with high prevalence and no sign of a decline. Female ASAF is predicted to grow to approach the male level. The narrowing of the ASAF gap is forecast to lead to a sustained closing of the life expectancy between-gender gap (Figure 3.10b).

Figures 3.10c and 3.10d show projections of male and female life expectancy for 2015–2060. We project male life expectancy for the period 2055-2060 be 83.2, with a 95% predictive interval of (80.9, 86.3). In contrast with the USA and the Netherlands, our median projection is 1.8 years *less* than that from bayesLife method. This is due to the fact that the epidemic has not yet clearly peaked. We project female life expectancy to be 84.5, with a 95% predictive interval of (81.7, 88.5), which is again substantially smaller than that from the bayesLife method with forecast median 87.6 years and 95% prediction interval (84.1, 91.0). This is due to the increasing impact of smoking on the Chilean female population.

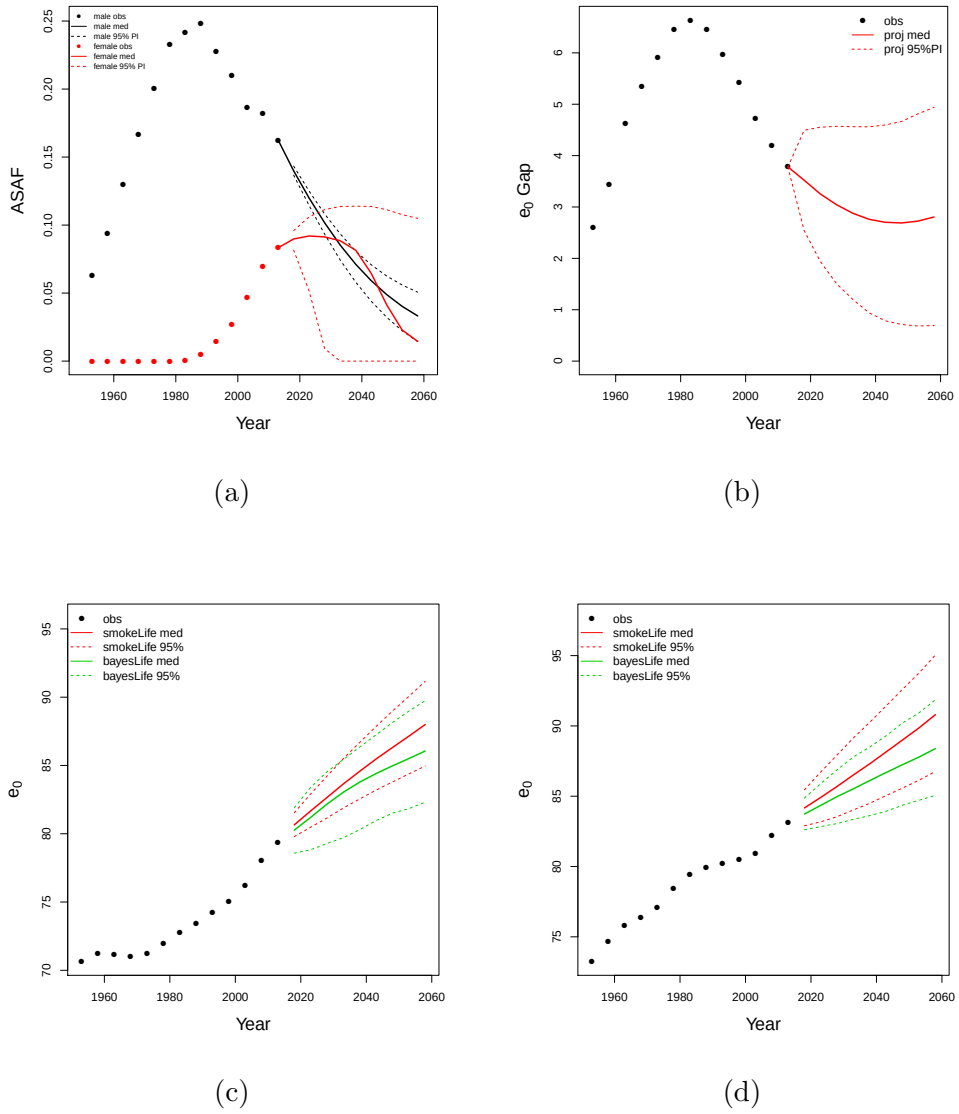


Figure 3.9: The Netherlands. (a) All-age smoking attributable fraction (ASAF) for male (black) and female (red) with median and 95% PI of posterior predictive distributions. (b) Between-gender gap of life expectancy at birth with posterior predictive median (red solid) and 95% PI (red dotted). (c) Forecasts of male life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted). (d) Forecasts of female life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted).

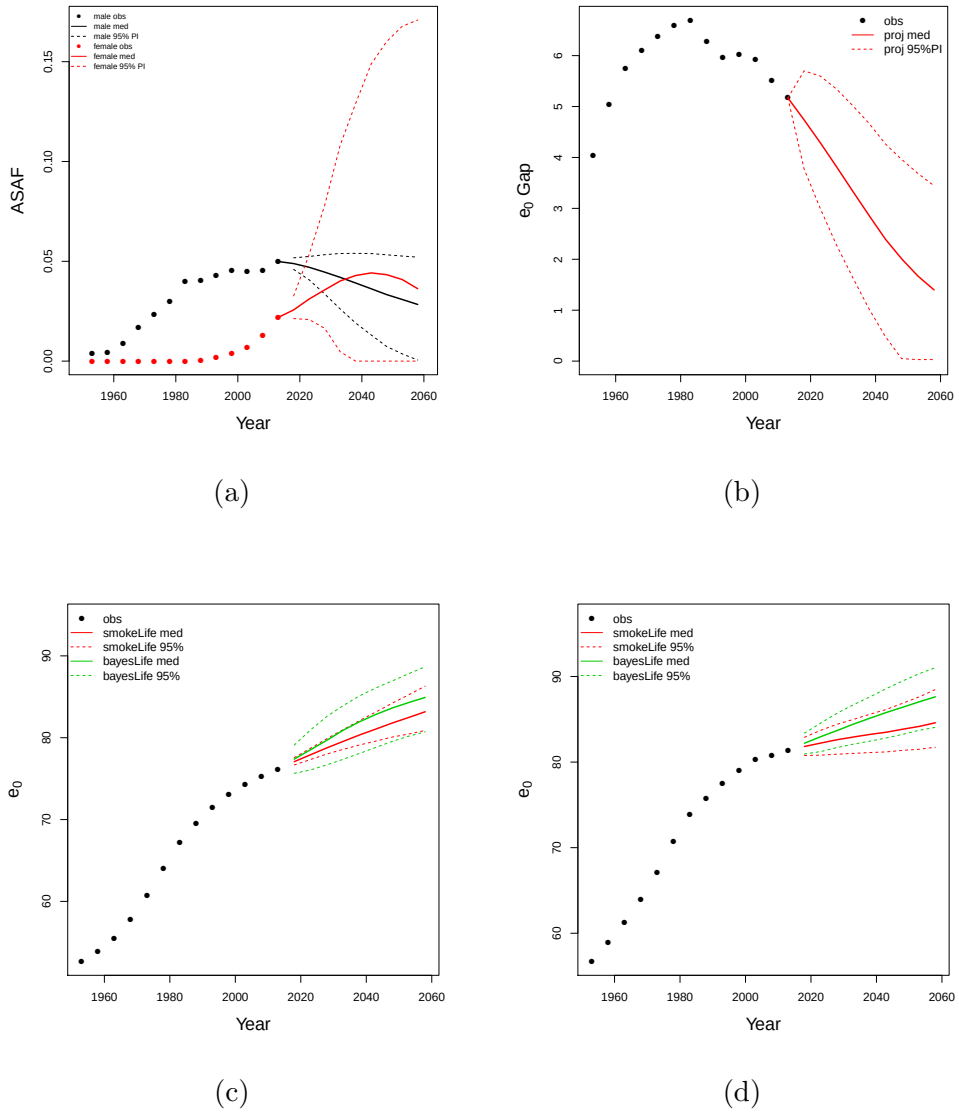


Figure 3.10: Chile. (a) All-age smoking attributable fraction (ASAF) for male (black) and female (red) with median and 95% PI of posterior predictive distributions. (b) Between-gender gap of life expectancy at birth with posterior predictive median (red solid) and 95% PI (red dotted). (c) Forecasts of male life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted). (d) Forecasts of female life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted).

3.4.4 Japan

Japan has been a leading country in life expectancy for a long period, while it also has a long history of smoking and is one of the largest tobacco consumers. Male smoking prevalence reached 83.7% in 1966. That number dropped to 36% in the 1990s and halved again by 2018. Female smoking prevalence is far lower and changes less dramatically than that of males. Female smoking prevalence reached 16% in the 1970s and decreased to 9.7% in 2015. The significant changes result mainly from government regulations and anti-smoking movements starting in the 1980s. Figure 3.11a shows the forecast male and female ASAF. Figure 3.11b shows the narrowing of the life expectancy gap as a result.

Figures 3.11c and 3.11d show projections of life expectancy for males and females. We project male life expectancy for the period 2055-2060 to be 88.8, with a 95% predictive interval of (85.8, 91.5). The bayesLife method forecasts 85.6, with a projection interval (81.6, 89.7). Notice that our median forecast is 3.2 years higher than that of bayesLife, while its interval is 1.4 years narrower. We project female life expectancy to be 92.2 with a 95% prediction interval of (88.3, 96.1). Our forecast shows a noticeable slowdown of the growth of female life expectancy due to the smoking effect. The bayesLife method projects 92.0 years with interval (88.8, 95.3). Though both methods produce comparable forecast results for 2055-2060, the bayesLife method forecasts a more linear increase while ours reflects the nonlinear smoking effect on the life expectancy forecast.

3.5 Discussion

We have proposed a method for probabilistic forecasting of mortality and life expectancy that takes account of the smoking epidemic. The method is based on the idea of the smoking attributable fraction of mortality, as estimated by the Peto-Lopez method using data on lung cancer mortality. The age-specific smoking attiributable fraction (ASSAF) of mortality is estimated and used to infer the non-smoking life expectancy at birth, e_0^{NS}. Both the ASSAF and e_0^{NS} are then forecast using a Bayesian hierarchical models for all countries with

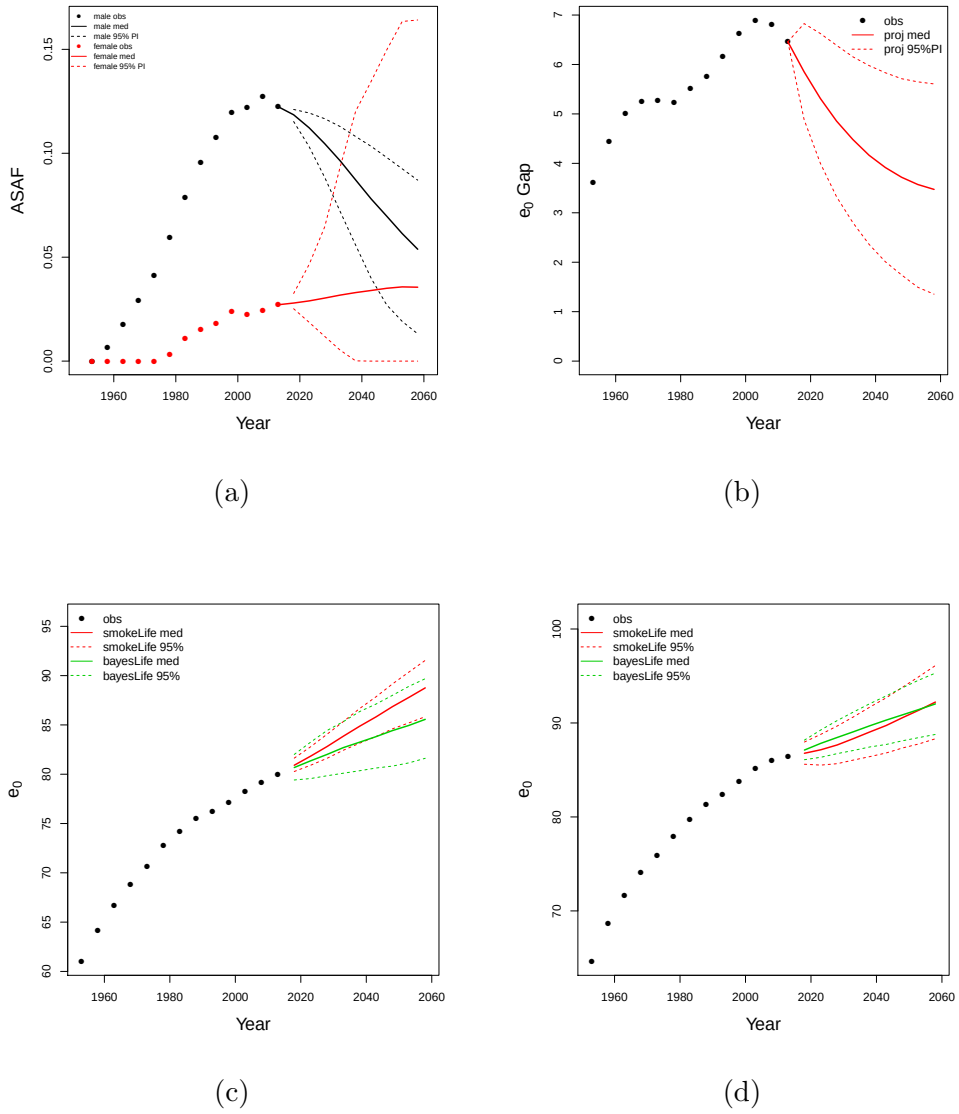


Figure 3.11: Japan. (a) All-age smoking attributable fraction (ASAF) for male (black) and female (red) with median and 95% PI of posterior predictive distributions. (b) Between-gender gap of life expectancy at birth with posterior predictive median (red solid) and 95% PI (red dotted). (c) Forecasts of male life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted). (d) Forecasts of female life expectancy at birth to 2060 using bayesLife method (green) and our proposed method (red) with posterior predictive medians (dashed) and 95% PI (dotted).

sufficiently good data. This in turn yields posterior predictive distributions of mortality rates and life expectancy at birth. The method performed well in an out-of-sample validation study.

The strength of the method derives from the fact that the smoking attributable fraction of mortality follows a very strong increasing-peaking-decreasing trend over time in all countries where the smoking epidemic has been going for long enough. This pattern is strong, broadly the same across countries, is to a large extent socially determined, and is also not highly correlated over time with the life expectancy at birth itself, which follows a broadly increasing pattern over time. However, smoking does impact mortality. Thus smoking mortality can be predicted with considerable accuracy, and accurate predictions improve mortality forecasts.

Another strength of the method is its use of a hierarchical model, which greatly facilitates forecasting, particularly for countries where the smoking epidemic is at an early stage. This allows forecasts for such countries to be informed by information from other countries, especially those where the epidemic is more advanced. It also makes it easier to incorporate all major sources of uncertainty.

The results indicate that for country-gender combinations where the smoking epidemic is advanced enough that we can expect it to be declining by 2060, incorporating smoking increases forecasts of life expectancy by about two years. When the epidemic is at an earlier stage, though, incorporating smoking tends to reduce forecasts of life expectancy. The results also indicate that much of change over time in the female-male gap in life expectancy is due to relative changes in smoking related mortality.

The biggest limitation of our method is that it relies on the availability of high-quality data on cause of death, particularly lung cancer, which are available for only around 70 countries of the 201 or so countries in the world. Thus the biggest improvement in the method would come from improvements in data quality. In particular, China and India are missing from our study, because national data on cause of death of high enough quality are not available. Producing such data should be a focus of future data collection and research. This is very important because, not only are China and India the two most populous countries

in the world, but they also have high smoking rates and are likely to experience high smoking mortality in the coming decades.

Several other approaches to the problem have been proposed. Bongaarts (2006) introduced the concept of non-smoking life expectancy, and proposed modeling it in a linear way. However, the time evolution of non-smoking life expectancy appears generally to follow a non-linear pattern, with gains that broadly follow a non-monotonic increasing-peaking-declining patter. This is modeled in our method by a random walk with a the double logistic drift.

Janssen et al. (2013) proposed directly modeling the ASSAF and the age-specific non-smoking attributable mortality rates. They observed that non-smoking mortality rates decline more linearly than overall mortality rates, making the data fit a Lee-Carter model better. They conducted an age-period-cohort analysis, while we found an age-cohort model to be sufficient. There are well-known identifiability issues with age-period-cohort analysis that our approach avoids. They used a coherent Lee-Carter method. This assumes linear progress in log mortality rates, while in fact progress tends to be nonlinear, and also tends to be more linear on the scale of life expectancy than of log mortality rates, which our double logistic random walk attempts to represent.

The mortality component of the UN's population projections for all countries is based on the Bayesian hierarchical model of Raftery et al. (2013), which does not take account of smoking. We have shown that this could be improved significantly by taking account of smoking. However, the data to do this are available for only around 70 countries currently, and the UN aims to use a unified approach for all the 230 countries and territories that they analyze. Thus extending the UN's method to take account of smoking in this way might not be feasible in the short term. To do this would likely require a major improvement in data availability for many countries. However, it could be useful for national population and mortality projections for individual countries, for example for planning health services, and also for the private sector, for example for actuarial and insurance analyses.

Chapter 4

MOMENT BOUNDS FOR AUTOCOVARIANCE MATRICES UNDER DEPENDENCE

4.1 Introduction

Consider a sequence of p-dimensional mean-zero random vectors $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ and a size-n fraction $\{\boldsymbol{Y}_i\}_{i=1}^n$ of it. This chapter aims to establish moment bounds for the spectral norm deviation of lag-m autocovariances of $\{\boldsymbol{Y}_i\}_{i=1}^n$, $\hat{\boldsymbol{\Sigma}}_m := (n-m)^{-1}\sum_{i=1}^{n-m}\boldsymbol{Y}_i\boldsymbol{Y}_{i+m}^\top$, from their mean values.

A first result at the origin of such problems concerns product measures, with $m=0$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ independent and identically distributed (i.i.d.). For this, Rudelson (1999) derived a bound on $\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_0 - \mathbb{E}\hat{\boldsymbol{\Sigma}}_0\|$, where $\|\cdot\|$ represents the spectral norm for matrices. The technique is based on symmetrization and the derived maximal inequality is a consequence of a concentration inequality on a "symmetrized" version of $p\times p$ symmetric and deterministic matrices, $\mathbf{A}_1,\ldots,\mathbf{A}_n$ (cf. Oliveira (2010)). That is, for any $x\geq 0$,

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^n \epsilon_i\mathbf{A}_i\Big\| \geq x\Big) \leq 2p\exp\{-x^2/(2\sigma^2)\}, \quad \sigma^2 := \Big\|\sum_{i=1}^n \mathbf{A}_i^2\Big\|, \tag{4.1}$$

where $\{\epsilon_i\}_{i=1}^n$ are independent and taking values $\{-1,1\}$ with equal probability. The applicability of this technique then hinges on the assumption that the data are i.i.d..

Later, Vershynin (2012), Srivastava and Vershynin (2013), Mendelson and Paouris (2014), Lounici (2014), Bunea and Xiao (2015), Tikhomirov (2017), among many others, derived different types of deviation bounds for $\hat{\boldsymbol{\Sigma}}_0$ under different distributional assumptions. For example, Lounici (2014) and Bunea and Xiao (2015) showed that, for such $\{\boldsymbol{Y}_i\}_{i=1}^n$ that are subgaussian and i.i.d.,

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\| \leq C\|\boldsymbol{\Sigma}_0\|\Big\{\sqrt{\frac{r(\boldsymbol{\Sigma}_0)\log(ep)}{n}} + \frac{r(\boldsymbol{\Sigma}_0)\log(ep)}{n}\Big\}. \tag{4.2}$$

Here $C > 0$ is a universal constant, $\mathbf{\Sigma}_0 := \mathbb{E}\boldsymbol{Y}_1\boldsymbol{Y}_1^\mathsf{T}$, and $r(\mathbf{\Sigma}_0) := \mathrm{tr}(\mathbf{\Sigma}_0)/\|\mathbf{\Sigma}_0\|$ is termed the "effective rank" (Vershynin, 2012) where $\mathrm{tr}(\mathbf{X}) := \sum_{i=1}^p \mathbf{X}_{i,i}$ for any real $p \times p$ matrix $\mathbf{X}$.

Statistically speaking, Equation (4.2) is of rich implications. For example, combining (4.2) with Davis-Kahan inequality (Davis and Kahan, 1970) suggests that the principal component analysis (PCA), a core statistical method whose aim is to recover the leading eigenvectors of $\mathbf{\Sigma}_0$, could still produce consistent estimators even if the dimension p is much larger than the sample size n, as long as the "intrinsic dimension" of the data, quantified by $r(\mathbf{\Sigma}_0)$, is small enough. See Section 1 in Han and Liu (2018) for more discussions on the statistical performance of PCA in high dimensions.

The main goal of this chapter is to give extensions of the deviation inequality (4.2) to large autocovariance matrices, where the matrices are constructed from a high dimensional structural time series. Examples of such time series include linear vector autoregressive model of lag d (VAR(d)), vector-valued autoregressive conditionally heteroscedastic (ARCH) model, and a model used in Banna et al. (2016). The main result appears below as Theorem 3, and is nonasymptotic in its nature. This result will have important consequences in high dimensional time series analysis. For example, it immediately yields new analysis for estimating large covariance matrix (Chen et al., 2013), a new proof of consistency for Brillinger's PCA in the frequency domain (cf. Chapter 9 in Brillinger (2001)), and we envision that it could facilitate a new proof of consistency for the PCA procedure proposed in Chang et al. (2018).

The rest of this chapter is organized as follows. Section 4.2 characterizes the settings and gives the main concentration inequality for large autocovariance matrices. In Section 4.3, we present applications of our results to some specific time series models. Proofs of the main results are given in Section 4.4, with more relegated to an appendix.

4.2 Main results

We first introduce the notation that will be used in this chapter. Without further specification, we use bold, italic lower case alphabets to denote vectors, e.g., $\boldsymbol{u} = (u_1, \cdots, u_p)^\mathsf{T}$ as a p-dimensional real vector, and $\|\boldsymbol{u}\|_2$ as its vector L_2 norm. We use bold, upper case

alphabets to denote matrices, e.g., $\mathbf{X} = (X_{i,j})$ as a $p \times p$ real matrix, and $\mathbf{I}_p$ as the $p \times p$ identity matrix. Throughout the chapter, let c, c', C, C', C'' be generic universal positive constants, whose actual values may vary at different locations. For any two sequences of positive numbers $\{a_n\}, \{b_n\}$, we denote $a_n = O(b_n)$ if there exists an universal constant C such that $a_n \leq C b_n$ for all n large enough. We write $a_n \asymp b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$ hold.

Consider a time series $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ of p-dimensional real entries $\boldsymbol{Y}_t \in \mathbb{R}^p$ with $\mathbb{R}, \mathbb{Z}$ denoting the sets of real and integer numbers respectively. In the sequel, the considered time series does not need to be stationary nor centered, and we are focused on a size-n fraction of it. Without loss of generality, we denote this fraction to be $\{\boldsymbol{Y}_i\}_{i=1}^n$.

As described in the introduction, the case of independent $\{\boldsymbol{Y}_i\}_{i=1}^n$ has been discussed in depth in recent years. We are interested here in the time series setting, and our main emphasis will be to describe nontrivial but easy to verify cases for which Inequality (4.2) still holds. The following four assumptions are accordingly made, with the notations that

$$\mathbb{S}^{p-1} := \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 = 1\}, \quad \bar{\mathbb{S}}^{p-1} := \{\boldsymbol{x} \in \mathbb{R}^p : |x_1| = \cdots = |x_p| = 1\},$$

and

$$\|X\|_{L(p)} := (\mathbb{E}|X|^p)^{1/p}, \quad \|X\|_{\psi_2} := \inf\{k \in (0,\infty) : \mathbb{E}[\exp\{(|X|/k)^2\} - 1] \leq 1\}$$

for any random variable X.

(A1) Define

$$\kappa_1 := \sup_{t\in\mathbb{Z}} \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\|_{\psi_2} < \infty, \quad \kappa_* := \sup_{t\in\mathbb{Z}} \sup_{\boldsymbol{v}\in\bar{\mathbb{S}}^{p-1}} \|\boldsymbol{v}^\mathsf{T}\boldsymbol{Y}_t\|_{\psi_2} < \infty.$$

Note that κ_1 is the supremum taken over vectors in the unit hypersphere, while κ_* is the supremum taken over vectors in the discrete hypercube.

(A2) Assume that there exist some constants $\gamma_1, \gamma_2, \epsilon > 0$ such that for any integer j, there exists a sequence of random vectors $\{\tilde{\boldsymbol{Y}}_t\}_{t>j}$ which is independent of $\sigma(\{\boldsymbol{Y}_t\}_{t\leq j})$, identically distributed as $\{\boldsymbol{Y}_t\}_{t>j}$, and for any integer $k \geq j+1$,

$$\|\|\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k\|_2\|_{L(1+\epsilon)} \leq \gamma_1\kappa_1 \exp\{-\gamma_2(k-j-1)\}.$$

(A3) Assume that there exist some constants $\gamma_3, \gamma_4, \epsilon > 0$ such that for any integer j, there exists a sequence of random vectors $\{\tilde{\boldsymbol{Y}}_t\}_{t>j}$ which is independent of $\sigma(\{\boldsymbol{Y}_t\}_{t\le j})$, identically distributed as $\{\boldsymbol{Y}_t\}_{t>j}$, and for any integer $k \ge j+1$,

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|(\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k)^\top \boldsymbol{u}\|_{L(1+\epsilon)} \le \gamma_3\kappa_1 \exp\{-\gamma_4(k-j-1)\}.$$

(A4) Assume there exists an universal constant $c > 0$ such that, for all $t \in \mathbb{Z}$ and for all $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{u}^\top \boldsymbol{Y}_t\|_{\psi_2}^2 \le c\mathbb{E}(\boldsymbol{u}^\top \boldsymbol{Y}_t)^2$.

Two observations are in order. We first define a generalized "effective rank" as follows:

$$r_* := \kappa_*^2/\kappa_1^2.$$

It is easy to see the close relationship between r_* and the effective rank highlighted in (4.2). As $\boldsymbol{Y}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$, κ_1^2 and κ_*^2 scale at the same orders of $\|\boldsymbol{\Sigma}_0\|$ and $\mathrm{tr}(\boldsymbol{\Sigma}_0)$, and the same observation applies to all subgaussian distributions with the additional condition **(A4)**, which is identical to Assumption 1 in Lounici (2014). As a matter of fact, r_* could be considered as a natural generalized version of $r(\boldsymbol{\Sigma}_0)$ without these additional assumptions, and is used in our main theorem.

Secondly, we note that Assumptions **(A2)** and **(A3)** are characterizing the intrinsic coupling property of the sequence. In practice, such couples can be constructed from time to time. Consider, for example, the following causal shift model,

$$\boldsymbol{Y}_t = H_t(\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots),$$

where $\{\xi_t\}_{t\in\mathbb{Z}}$ consists of independent elements with values in a measurable space $\mathcal{X}$ and $H_t : \mathcal{X}^{\mathbb{Z}^+} \to \mathbb{R}^p$ is a vector-valued function. Then it is natural to consider

$$\tilde{\boldsymbol{Y}}_t = H_t(\xi_t, \ldots \xi_{j+1}, \tilde{\xi}_j, \tilde{\xi}_{j-1}, \ldots)$$

for an independent copy $\{\tilde{\xi}_t\}_{t\in\mathbb{Z}}$ of $\{\xi_t\}_{t\in\mathbb{Z}}$.

The following is the main result of this chapter.

Theorem 3 (Proof in Section 4.4.1)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a sequence of random vectors satisfying Assumptions **(A1)-(A3)** and recall $r_* = \kappa_*^2/\kappa_1^2$. Assume $\gamma_1 = O(\sqrt{r_*})$ and $\gamma_3 = O(1)$. Then, for any integer $n \geq 2$ and $0 \leq m \leq n-1$, we have*

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_m - \mathbb{E}\hat{\boldsymbol{\Sigma}}_m\| \leq C\kappa_1^2\Big\{\sqrt{\frac{r_* \log ep}{n-m}} + \frac{r_* \log ep(\log np)^3}{n-m}\Big\} \tag{4.3}$$

*for some constant C only depending on $\epsilon, m, \gamma_2, \gamma_4$. If in addition, $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ is a second-order stationary sequence of mean-zero random vectors and Assumption **(A4)** holds, then*

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_m - \mathbb{E}\hat{\boldsymbol{\Sigma}}_m\| \leq C'\|\boldsymbol{\Sigma}_0\|\Big\{\sqrt{\frac{r(\boldsymbol{\Sigma}_0) \log ep}{n-m}} + \frac{r(\boldsymbol{\Sigma}_0) \log ep(\log np)^3}{n-m}\Big\}$$

for some constant C' only depending on $\epsilon, c, m, \gamma_2, \gamma_4$.

We first comment on the temporal correlatedness conditions, Assumptions **(A2)** and **(A3)**. We note that they correspond exactly to the δ-measure of dependence introduced in Chapter 3 of Dedecker et al. (2007), for the sequence $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ and $\{\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ respectively. In addition, as will be seen soon, our measure of dependence is also very related to the τ-measure introduced in Dedecker and Prieur (2004). In particular, ours is usually stronger than, but as $\epsilon \to 0$, reduces to the τ-measure. Lastly, our conditions are also quite connected to the functional dependence measure in Wu (2005), on which many moment inequalities in real space have been established (cf. Liu et al. (2013) and Wu and Wu (2016)). However, it is still unclear if a similar matrix Bernstein inequality could be developed under Weibiao Wu's functional dependence condition.

Secondly, we note that one is ready to verify that Inequality (4.3) gives the exact control of the deviation from the mean. Actually, Inequality (4.3) is nearly a strict extension of the results in Lounici (Lounici, 2014) and Bunea and Xiao (Bunea and Xiao, 2015) to weak data dependence up to some logarithmic terms. This extension is achieved by applying Theorem 10, a concentration inequality for a sequence of weakly dependent random matrices. Theorem 10 is an extension of the Bernstein-type inequality for real-valued weakly dependent random variables derived in Merlevède et al. (2011) to dependent random matrices, and is a slight extension of the Bernstein-type inequality for a sequence of β-mixing random

matrices derived in Banna et al. (2016). In some applications, especially those in high dimensions, verifying the weak dependence condition in Theorem 10 is more straightforward than verifying the β-mixing condition in Theorem 1 in Banna et al. (2016). The details of the weak dependence condition will be introduced in Section 4.4.1, and Theorem 10 will be proved in the Appendix.

Admittedly, it is still unclear if Inequality (4.3) could be further improved under the given conditions. Recently, in a remarkable series of papers (Koltchinskii and Lounici, 2017a,b,c), Koltchinskii and Lounici showed that, for subgaussian independent data, the extra multiplicative p term on the righthand side of Inequality (4.3) could be further removed. The proof rests on Talagrand's majorizing measures (Talagrand, 2014) and a corresponding maximal inequality due to Mendelson (Mendelson, 2010). In the most general case, to the authors' knowledge, it is still unknown if Talagrand's approach could be extent to weakly dependent data, although we conjecture that, under stronger temporal dependence (e.g., geometrically ϕ-mixing) conditions, it is possible to recover Koltchinskii and Lounici's result without resorting to the matrix Bernstein inequality in the proof of Theorem 3.

Nevertheless, we make a first step towards eliminating these logarithmic terms via the following theorem. It shows, when assuming a Gaussian sequence is observed, one could further tighten the upper bound in Inequality (4.3) by removing all logarithm factors. The obtained bound is thus tight in view of Theorem 2 in Lounici (2014) and Theorem 4 in Koltchinskii and Lounici (2017a).

Theorem 4 (Proof in Section 4.4.2)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a stationary mean-zero Gaussian sequence that satisfies Assumptions **(A2)**-**(A3)** with $\gamma_1 = O(\sqrt{r(\boldsymbol{\Sigma}_0)})$ and $\gamma_3 = O(1)$. Then, for any integer $n \geq 2$ and $0 \leq m \leq n-1$,*

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_m\| \leq C\|\boldsymbol{\Sigma}_0\|\Big(\sqrt{\frac{r(\boldsymbol{\Sigma}_0)}{n-m}} + \frac{r(\boldsymbol{\Sigma}_0)}{n-m}\Big)$$

for some constant $C > 0$ only depending on $\epsilon, m, \gamma_2, \gamma_4$.

In a related track of studies, Bai and Yin (1993), Srivastava and Vershynin (2013), Mendelson and Paouris (2014), and Tikhomirov (2017), among many others, explored the

optimal scaling requirement in approximating a large covariance matrix for heavy-tailed data. For instance, for i.i.d. data and as $\mathbf{\Sigma}_0$ is identity, Bai and Yin (Bai and Yin, 1993) showed that $\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|$ will converge to zero in probability as long as $p/n \to 0$ and 4-th moments exist. Some recent developments further strengthen the moment requirement. These results cannot be compared to ours. In particular, our analysis is focused on characterizing the role of "effective rank", a term of strong meanings in statistical implications and a feature that cannot be captured using these alternative procedures.

4.3 Applications

In this section, we examine the validity of Assumptions **(A1)**-**(A4)** in Section 4.2 under three models, a stable VAR(d) model, a model proposed by Banna et al. (2016), and an ARCH-type model. One shall be aware of examples that are of VAR(d) or ARCH-type structures but are not α- or β-mixing (cf. Andrews (1984)).

We first consider such $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ that is a random sequence generated from VAR(d) model, i.e.,

$$\boldsymbol{Y}_t = \mathbf{A}_1\boldsymbol{Y}_{t-1} + \cdots + \mathbf{A}_d\boldsymbol{Y}_{t-d} + \boldsymbol{E}_t,$$

where $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}}$ is a sequence of independent vectors such that for all $t \in \mathbb{Z}$ and $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{u}^\top\boldsymbol{E}_t\|_{\psi_2} \le c'\|\boldsymbol{u}^\top\boldsymbol{E}_t\|_{L(2)}$ for some universal constant $c' > 0$. In addition, assume $\sup_{t\in\mathbb{Z}}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\|\boldsymbol{u}^\top\boldsymbol{E}_t\|_{\psi_2} < D_1$ for some universal positive constant $D_1 < \infty$, $\|\mathbf{A}_k\| \le a_k < 1$ for all $1 \le k \le d$, and $\sum_{k=1}^d a_k < 1$, where $\{a_k\}_{k=1}^d, d$ are some universal constants.

Under these conditions, we have the following theorem.

Theorem 5 (Proof in Section 4.4.4)**.** *The above* $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ *satisfies Assumptions* ***(A1)***-***(A4)*** *with*

$$\gamma_1 = C(\kappa_*/\kappa_1)(\|\bar{\mathbf{A}}\|/\rho_1)^K, \gamma_2 = \log(\rho_1^{-1}), \gamma_3 = C'd(\|\bar{\mathbf{A}}\|/\rho_1)^K, \gamma_4 = \log(\rho_1^{-1}).$$

Here we denote

$$\bar{\mathbf{A}} := \begin{bmatrix} a_1 & a_2 & \dots & a_{d-1} & a_d \\ 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix},$$

ρ_1 is a universal constant such that $\rho(\bar{\mathbf{A}}) < \rho_1 < 1$ whose existence is guaranteed by the assumption that $\sum_{k=1}^{d} a_k < 1$ (cf. Lemma 17 in Section 4.4), K is some constant only depending on ρ_1, and $C, C' > 0$ are some constants only depending on ϵ.

We secondly consider the following time series generation scheme whose corresponding matrix version has been considered by Banna, Merlevède, and Youssef (Banna et al., 2016). In detail, let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a random sequence generated by

$$\boldsymbol{Y}_t = W_t \boldsymbol{E}_t,$$

where $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}}$ is a sequence of independent random vectors independent of $\{W_t\}_{t\in\mathbb{Z}}$ such that for all $t \in \mathbb{Z}$ and $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{u}^\top \boldsymbol{E}_t\|_{\psi_2} \le c' \|\boldsymbol{u}^\top \boldsymbol{E}_t\|_{L(2)}$ for some universal constant $c' > 0$. In addition, we assume

$$\sup_{t\in\mathbb{Z}} \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|\boldsymbol{u}^\top \boldsymbol{E}_t\|_{\psi_2} \le \kappa_1' \quad \text{and} \quad \sup_{t\in\mathbb{Z}} \sup_{\boldsymbol{v}\in\bar{\mathbb{S}}^{p-1}} \|\boldsymbol{v}^\top \boldsymbol{E}_t\|_{\psi_2} \le \kappa_*'$$

for some constants $0 < \kappa_1', \kappa_*' < \infty$, $\{W_t\}_{t\in\mathbb{Z}}$ is a sequence of uniformly bounded τ-mixing random variables such that $\max_{t\in\mathbb{Z}} |W_t| \le \kappa_W$, and

$$\tau(k; \{W_t\}_{t\in\mathbb{Z}}, |\cdot|) \le \kappa_W \gamma_5 \exp\{-\gamma_6 (k-1)\}$$

for some constants $0 < \gamma_5, \gamma_6, \kappa_W < \infty$ (see, Appendix Section C.1 for a detailed introduction to the τ-mixing random variables).

Theorem 6 (Proof in Section 4.4.4)**.** *The above $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ satisfies Assumptions **(A1)-(A4)** with*

$$\gamma_1 = C\kappa_*' \kappa_W \gamma_5^{\frac{1}{1+\epsilon}} / \kappa_1, \gamma_2 = \gamma_6/(1+\epsilon), \gamma_3 = C'\kappa_1' \kappa_W \gamma_5^{\frac{1}{1+\epsilon}} / \kappa_1, \gamma_4 = \gamma_6/(1+\epsilon)$$

for some constants $C, C' > 0$ only depending on ϵ.

Lastly, we consider an vector-valued ARCH-model with $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ being a random sequence generated by

$$\boldsymbol{Y}_t = \mathbf{A}\boldsymbol{Y}_{t-1} + H(\boldsymbol{Y}_{t-1})\boldsymbol{E}_t,$$

where $H : \mathbb{R}^p \to \mathbb{R}^{p\times p}$ is a matrix-valued function and $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}}$ is a sequence of independent random vectors such that

$$\sup_{t\in\mathbb{Z}}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|\boldsymbol{u}^\mathsf{T}\boldsymbol{E}_t\|_{\psi_2} \le \kappa_1' \quad \text{and} \quad \sup_{t\in\mathbb{Z}}\sup_{\boldsymbol{v}\in\bar{\mathbb{S}}^{p-1}} \|\boldsymbol{v}^\mathsf{T}\boldsymbol{E}_t\|_{\psi_2} \le \kappa_*'$$

for some constants $0 < \kappa_1', \kappa_*' < \infty$. Assume further that $\|\mathbf{A}\| \le a_1$ and the function $H(\cdot)$ satisfies

$$\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^p} \|H(\boldsymbol{u}) - H(\boldsymbol{v})\| \le \frac{a_2}{\kappa_*'}\|\boldsymbol{u}-\boldsymbol{v}\|_2$$

for some universal constant $a_1 < 1, a_2 > 0$ such that $a_1 + a_2 < 1$.

Theorem 7 (Proof in Section 4.4.4)**.** *If the above $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ satisfies Assumption **(A1)**, it satisfies Assumptions **(A2)**-**(A3)** with*

$$\gamma_1 = C\kappa_*/\kappa_1, \gamma_2 = -\log(a_1+a_2), \gamma_3 = C'\max(\kappa_*\kappa_1'/\kappa_1\kappa_*', 1), \gamma_4 = \log(a_1+a_2)^{-1}$$

*for some constants $C, C' > 0$ only depending on ϵ. If we further assume the above $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ to be a stationary sequence and $\sup_{\boldsymbol{u}\in\mathbb{R}^p}\|H(\boldsymbol{u})\| < D_2$ for some universal constant $D_2 < \infty$, then $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ satisfies Assumption **(A1)**.*

4.4 Proofs

4.4.1 Proof of Theorem 3

Proof of Theorem 3. The proof depends mainly on the following tail probability bound of deviation of the sample covariance from its mean.

Proposition 8 (Proof in Section 4.4.1)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a sequence of random vectors satisfying **(A1)**-**(A3)**. For any integer $n \ge 2$, integer $0 \le m \le n-2$ and real number $0 < \delta \le 1$, define*

$$M_\delta := C\max\left\{\left(\frac{\kappa_*}{\kappa_1}\right)^2 \log\frac{n-m}{\delta}, \left(\frac{\kappa_*}{\kappa_1}\right)^2, \frac{2\kappa_*\gamma_1}{\kappa_1}\right\}.$$

Then for any $x \geq 0$,

$$\mathbb{P}[\|\hat{\mathbf{\Sigma}}_m - \mathbb{E}\hat{\mathbf{\Sigma}}_m\| \geq \kappa_1^2\{x + \sqrt{\delta/(n-m)}\}] \leq 2p \exp\left\{ - \frac{C'(n-m)^2x^2}{A_1(n-m) + A_2M_\delta^2 + A_3(n-m)xM_\delta} \right\} + \delta,$$

with

$$A_1 := \frac{\{\kappa_*\gamma_1/\kappa_1 + (\kappa_*/\kappa_1)^2(\gamma_3 + 2m + 1) + 2m + 1\}}{1 - \exp\{-\min(\frac{5+\epsilon}{6\epsilon+10}\gamma_2, \gamma_4)\}}, \qquad A_2 := \frac{453^2}{\gamma_2},$$
$$A_3 := \frac{2\log(n-m)}{\log 2} \max\left\{1, 8m + \frac{48\log(n-m)p}{\gamma_2}\right\}$$

for some constants $C, C' > 0$ *only depending on* ϵ.

Without loss of generality, let $m = 0$. Taking $x = \sqrt{\frac{r_*\log ep}{n}}t$, $\delta = x^{-\gamma}$ for some $\gamma > 1$, $\gamma_1 = O(\sqrt{r_*})$, and $\gamma_3 = O(1)$ in Proposition 8, we obtain

$$\mathbb{P}\Big(\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\| \geq C_1\kappa_1^2\sqrt{\frac{r_*\log ep}{n}}t\Big) \leq 2p\exp\left[- \frac{C_2(\log ep)t/\{\log(\sqrt{\frac{r_*\log ep}{n}}t)\}^2}{1 + \frac{r_*(\log n)^2}{n} + \sqrt{\frac{r_*\log ep}{n}}t(\log np)^3}\right] + x^{-\gamma}$$

for some constants $C_1, C_2 > 0$ only depending on $\epsilon, \gamma_2, \gamma_4$.

If $1 + \frac{r_*(\log n)^2}{n} \geq \frac{r_*\log ep(\log np)^6}{n}$, we have

$$\begin{aligned}
\frac{\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\|^2}{(C_1\kappa_1^2\sqrt{\frac{r_*\log ep}{n}})^2} \leq & 1 + \frac{r_*(\log n)^2}{n} + \int_{1+\frac{r_*(\log n)^2}{n}}^{\frac{\{1+\frac{r_*(\log n)^2}{n}\}^2}{\frac{r_*\log ep(\log np)^6}{n}}} 2p\exp\left[- \frac{C_2(\log ep)t/\{\log(\sqrt{\frac{r_*\log ep}{n}}t)\}^2}{1 + \frac{r_*(\log n)^2}{n}}\right]\mathrm{d}t \\
& + \int_{\frac{\{1+\frac{r_*(\log n)^2}{n}\}^2}{\frac{r_*\log ep(\log np)^6}{n}}}^{\infty} 2p\exp\left[- \frac{C_2(\log ep)\sqrt{t}/\{\log(\sqrt{\frac{r_*\log ep}{n}}t)\}^2}{\sqrt{\frac{r_*\log ep(\log np)^6}{n}}}\right]\mathrm{d}t \\
\leq & C_3\Big(1 + \frac{r_*(\log n)^2}{n} + \frac{r_*\log ep(\log np)^6}{n}\Big).
\end{aligned}$$

This gives that

$$\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\|^2 \leq C_4\kappa_1^4\Big\{\frac{r_*\log ep}{n} + \frac{r_*^2(\log ep)^2(\log np)^6}{n^2}\Big\}.$$

On the other hand, if $1 + \frac{r_*(\log n)^2}{n} \leq \frac{r_*\log ep(\log np)^6}{n}$,

$$\begin{aligned}
\frac{\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\|^2}{(C_1\kappa_1^2\sqrt{\frac{r_*\log ep}{n}})^2} \leq & \frac{r_*\log ep(\log np)^6}{n} + \int_{\frac{r_*\log ep(\log np)^6}{n}}^{\infty} 2p\exp\left[- \frac{C_2(\log ep)\sqrt{t}/\{\log(\sqrt{\frac{r_*\log ep}{n}}t)\}^2}{\sqrt{\frac{r_*\log ep(\log np)^6}{n}}}\right]\mathrm{d}t \\
\leq & C_5\frac{r_*\log ep(\log np)^6}{n}.
\end{aligned}$$

This renders

$$\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\|^2 \leq C_5\kappa_1^4\Big\{\frac{r_*^2(\log ep)^2(\log np)^6}{n^2}\Big\}.$$

Combining two cases gives us the final result by using the simple fact that $\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\| \leq (\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\|^2)^{\frac{1}{2}}$. This completes the proof of the first part of Theorem 3.

Notice that under Assumptions **(A1)**, **(A4)**, zero-mean, and second-order stationarity, we have $\kappa_1^2 \asymp \|\mathbf{\Sigma}_0\|$ and $\kappa_*^2 \asymp \mathrm{tr}(\mathbf{\Sigma}_0)$. Thus plugging in the first part of Theorem 3 finishes the proof. □

Now we prove Proposition 8 under Assumptions **(A1)**-**(A3)**. In the proof, the cases for covariance and autocovariance matrices are treated separately. In the following we give a roadmap. The proof of Proposition 8 is based on combining a Bernstein-type inequality for the almost surely (a.s.) bounded matrices and a truncation method. The probability bound for the a.s. bounded part (a.k.a., the truncated part) of the random matrix is obtained by employing a Bernstein-type inequality for τ-mixing random matrices, which is presented in Theorem 10, and some related lemmas (Lemmas 11 and 12), whose proofs are presented later. The tail part of the random matrix is controlled under the sub-Gaussian Assumption **(A1)**, which uses Lemma 9 that will be presented soon.

In more detail, given a sequence of random vectors $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$, denote $\mathbf{X}_t := \boldsymbol{Y}_t\boldsymbol{Y}_t^\mathsf{T}$ for all $t \in \mathbb{Z}$. Then for any constant $M > 0$, we introduce the following "truncated" version of $\mathbf{X}_t$:

$$\mathbf{X}_t^M := \frac{M \wedge \|\mathbf{X}_t\|}{\|\mathbf{X}_t\|}\mathbf{X}_t,$$

where $a \wedge b := \min(a, b)$ for any two real numbers a, b.

For any integer $m > 0$, we denote $\mathbf{Z}_t^{(m)} := \boldsymbol{Y}_t\boldsymbol{Y}_{t+m}^\mathsf{T}$ for all $t \in \mathbb{Z}$. For the sake of clarification, the superscript "(m)" is dropped when no confusion is possible. Then the truncated version is

$$\mathbf{Z}_t^M := \frac{M \wedge \|\mathbf{Z}_t\|}{\|\mathbf{Z}_t\|}\mathbf{Z}_t$$

for any $M > 0$.

We further define the "variances" for $\{\mathbf{X}_i^M\}_{i=1}^n$ and $\{\mathbf{Z}_i^M\}_{i=1}^{n-m}$ as

$$\nu_{\mathbf{X}^M}^2 := \sup_{K\subseteq\{1,\ldots,n\}} \frac{1}{\mathrm{card}\,(K)}\lambda_{\max}\left\{\mathbb{E}\left(\sum_{i\in K}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M\right)^2\right\},$$
$$\nu_{\mathbf{Z}^M}^2 := \sup_{K\subseteq\{1,\ldots,n-m\}} \frac{1}{\mathrm{card}\,(K)}\left\|\mathbb{E}\left(\sum_{i\in K}\mathbf{Z}_i^M - \mathbb{E}\mathbf{Z}_i^M\right)^2\right\|.$$

Here $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ denote the largest and smallest eigenvalues of $\mathbf{X}$ respectively.

Proof of Proposition 8. We first assume $\kappa_1 = 1$. We consider two cases.

Case I: When $m = 0$, $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ is a sequence of symmetric random matrices. We have,

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\left\|\sum_{i=1}^n(\mathbf{X}_i - \mathbb{E}\mathbf{X}_i)\right\| \geq x\right\}\\
=&\mathbb{P}\left\{\frac{1}{n}\left\|\sum_{i=1}^n(\mathbf{X}_i - \mathbf{X}_i^M + \mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M + \mathbb{E}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i)\right\| \geq x\right\}\\
\leq&\mathbb{P}\left\{\frac{1}{n}\left\|\sum_{i=1}^n(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M + \mathbb{E}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i)\right\| + \frac{1}{n}\left\|\sum_{i=1}^n(\mathbf{X}_i - \mathbf{X}_i^M)\right\| \geq x\right\}\\
\leq&\mathbb{P}\left\{\left\|\sum_{i=1}^n(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M + \mathbb{E}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i)\right\| \geq nx\right\} + \mathbb{P}\left\{\left\|\sum_{i=1}^n(\mathbf{X}_i - \mathbf{X}_i^M)\right\| > 0\right\}\\
\leq&\mathbb{P}\left\{\left\|\sum_{i=1}^n(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M)\right\| \geq nx - \sum_{i=1}^n\|\mathbb{E}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i\|\right\} + \sum_{i=1}^n\mathbb{P}(\mathbf{X}_i \neq \mathbf{X}_i^M)\\
\leq&\mathbb{P}\left[\lambda_{\max}\left\{\sum_{i=1}^n(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M)\right\} \geq nx - \sum_{i=1}^n\|\mathbb{E}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i\|\right] +\\
&\mathbb{P}\left[\lambda_{\min}\left\{\sum_{i=1}^n(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M)\right\} \leq -nx + \sum_{i=1}^n\|\mathbb{E}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i\|\right] + \sum_{i=1}^n\mathbb{P}(\mathbf{X}_i \neq \mathbf{X}_i^M). \qquad (4.4)
\end{aligned}$$

We first show that the difference in expectation between the "truncated" $\mathbf{X}_t^{M_\delta}$ and original one $\mathbf{X}_t$ can be controlled with the chosen truncation level M_δ. For this, we need the following lemma.

Lemma 9 (Proof in Section 4.4.3)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a sequence of p-dimensional random vectors under Assumption **(A1)**. Then for all $t \in \mathbb{Z}$ and for all $x \geq 0$,*

$$\mathbb{P}\{\|\boldsymbol{Y}_t\|_2^2 \geq 2\kappa_*^2 + 8\kappa_*^2(x + \sqrt{x})\} \leq \exp(-Cx)$$

for some arbitary constant $C > 0$.

By applying Lemma 9, we obtain that for all $i \in \{1, \ldots, n\}$,

$$\begin{aligned}\|\mathbb{E}\mathbf{X}_i^{M_\delta} - \mathbb{E}\mathbf{X}_i\| =& \left\|\mathbb{E}\Big(1 - \frac{M_\delta}{\|\mathbf{X}_i\|}\Big)\mathbf{X}_i \mathbf{1}_{\{\|\mathbf{X}_i\| > M_\delta\}}\right\| \\ \leq& \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{u}^\mathsf{T}\mathbf{X}_i\boldsymbol{v}|\mathbf{1}_{\{\|\mathbf{X}_i\| > M_\delta\}} \\ \leq& \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \{\mathbb{E}(\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_i\boldsymbol{Y}_i^\mathsf{T}\boldsymbol{v})^2\}^{\frac{1}{2}}\{\mathbb{P}(\|\mathbf{X}_i\| > M_\delta)\}^{\frac{1}{2}} \\ \leq& \sqrt{\delta/n},\end{aligned}$$

where the last line followed by Assumption **(A1)**, Lemma 9, and the chosen M_δ.

The second step heavily depends on a Bernstein-type inequality for τ-mixing random matrices. The theorem slightly extends the main theorem of Banna et al. (2016) in which the random matrix sequence is assumed to be β-mixing. Its proof is relegated to the Appendix C.

Theorem 10 (Proof in Appendix)**.** *Consider a sequence of real, mean-zero, symmetric* $p \times p$ *random matrices* $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ *with* $\|\mathbf{X}_t\| \leq M$ *for some positive constant* M*. In addition, assume that this sequence is* τ*-mixing (see, Appendix Section C.1 for a detailed introduction to the* τ*-mixing coefficient) with geometric decay, i.e.,*

$$\tau(k; \{\mathbf{X}_t\}_{t\in\mathbb{Z}}, \|\cdot\|) \leq M\psi_1 \exp\{-\psi_2(k-1)\}$$

for some constants $\psi_1, \psi_2 > 0$*. Denote* $\tilde{\psi}_1 := \max\{p^{-1}, \psi_1\}$*. Then for any* $x \geq 0$ *and any integer* $n \geq 2$*, we have*

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{i=1}^n \mathbf{X}_i\right) \geq x\right\} \leq p\exp\left\{-\frac{x^2}{8(15^2 n\nu^2 + 60^2 M^2/\psi_2) + 2xM\tilde{\psi}(\tilde{\psi}_1, \psi_2, n, p)}\right\},$$

where

$$\nu^2 := \sup_{K\subseteq\{1,\ldots,n\}} \frac{1}{card(K)}\lambda_{\max}\left\{\mathbb{E}\left(\sum_{i\in K}\mathbf{X}_i\right)^2\right\} \text{ and } \tilde{\psi}(\tilde{\psi}_1, \psi_2, n, p) := \frac{\log n}{\log 2}\max\left\{1, \frac{8\log(\tilde{\psi}_1 n^6 p)}{\psi_2}\right\}.$$

In order to apply Theorem 10, we need the following two lemmas. Lemma 11 is to show that the sequence of "truncated" matrices $\{\mathbf{X}_t^M\}$ under Assumptions **(A1)-(A2)** is a τ-mixing random sequence with geometric decay. Lemma 12 calculates the upper bound for ν^2 term in Theorem 10 for $\{\mathbf{X}_t^M\}_{t\in\mathbb{Z}}$.

Lemma 11 (Proof in Section 4.4.3)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a sequence of random vectors under Assumptions **(A1)-(A2)**. Then $\{\mathbf{X}_t^M\}_{t\in\mathbb{Z}}$, $\{\mathbf{X}_t^M-\mathbb{E}\mathbf{X}_t^M\}_{t\in\mathbb{Z}}$, $\{\mathbf{Z}_t^M\}_{t\in\mathbb{Z}}$, and $\{\mathbf{Z}_t^M-\mathbb{E}\mathbf{Z}_t^M\}_{t\in\mathbb{Z}}$ are all τ-mixing random sequences. Moreover,*

$$\begin{aligned}
&\tau(k;\{\mathbf{X}_t^M\}_{t\in\mathbb{Z}},\|\cdot\|)\le C\gamma_1\kappa_1\kappa_*\exp\{-\gamma_2(k-1)\},\\
&\tau(k;\{\mathbf{X}_t^M-\mathbb{E}\mathbf{X}_t^M\}_{t\in\mathbb{Z}},\|\cdot\|)\le C\gamma_1\kappa_1\kappa_*\exp\{-\gamma_2(k-1)\},\\
&\tau(k;\{\mathbf{Z}_t^M\}_{t\in\mathbb{Z}},\|\cdot\|)\le C'\exp\{\gamma_2\min(k,m)\}\max(\gamma_1\kappa_1\kappa_*,\kappa_*^2)\exp\{-\gamma_2(k-1)\},\\
&\tau(k;\{\mathbf{Z}_t^M-\mathbb{E}\mathbf{Z}_t^M\}_{t\in\mathbb{Z}},\|\cdot\|)\le C'\exp\{\gamma_2\min(k,m)\}\max(\gamma_1\kappa_1\kappa_*,\kappa_*^2)\exp\{-\gamma_2(k-1)\}
\end{aligned}$$

for $k\ge 1$ and some constants $C,C'>0$ only depending on ϵ.

Lemma 12 (Proof in Section 4.4.3)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a sequence of random vectors under Assumptions **(A1)-(A3)**. Take $M\ge C\gamma_1\kappa_1\kappa_*$ for some constant $C>0$ only depending on ϵ. Then we obtain*

$$\begin{aligned}
\nu_{\mathbf{X}^M}^2&\le C'\frac{\kappa_1^2\{\kappa_1^2+\kappa_1\kappa_*\gamma_1+\kappa_*^2(\gamma_3+2)\}}{1-\exp\{-\min(\frac{5+\epsilon}{6\epsilon+10}\gamma_2,\gamma_4)\}},\\
\nu_{\mathbf{Z}^M}^2&\le C''\frac{\kappa_1^2\{(2m+1)\kappa_1^2+\kappa_1\kappa_*\gamma_1+\kappa_*^2(\gamma_3+2m+2)\}}{1-\exp\{-\min(\frac{5+\epsilon}{6\epsilon+10}\gamma_2,\gamma_4)\}}
\end{aligned}$$

for some constants $C',C''>0$ only depending on ϵ.

Therefore, by applying Theorems 10, Lemma 11, and Lemma 12 with the chosen M_δ, we obtain for any $x>0$,

$$\mathbb{P}\left[\lambda_{\max}\left\{\frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i^{M_\delta}-\mathbb{E}\mathbf{X}_i^{M_\delta})\right\}\ge x+\sqrt{\delta/n}\right]\le p\exp\left(-\frac{n^2x^2}{A_1n+A_2M_\delta^2+A_3nxM_\delta}\right),\tag{4.5}$$

where

$$A_1 := \frac{C\{\kappa_*\gamma_1 + \kappa_*^2(\gamma_3+2)+1\}}{1-\exp\{-\min(\frac{5+\epsilon}{6\epsilon+10}\gamma_2, \gamma_4)\}}, \quad A_2 := \frac{453^2}{\gamma_2}, \text{ and } A_3 := \frac{2\log n}{\log 2}\max\left\{1, \frac{48\log(np)}{\gamma_2}\right\}$$

for some constant $C > 0$ only depending on ϵ.

Similarly, notice that $\lambda_{\min}(\sum_{j=1}^n \mathbf{X}_j^{M_\delta}) = \lambda_{\max}(-\sum_{j=1}^n \mathbf{X}_j^{M_\delta})$. Hence the same argument renders the same upper bound

$$\mathbb{P}\left[\lambda_{\min}\left\{\frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i^{M_\delta} - \mathbb{E}\mathbf{X}_i^{M_\delta})\right\} \le -(x+\sqrt{\delta/n})\right] \le p\exp\left(-\frac{n^2x^2}{A_1 n + A_2 M_\delta^2 + A_3 n x M_\delta}\right) \tag{4.6}$$

with the same constants as above.

For the last term of (4.4), with the choice of M_δ and Lemma 9, we obtain

$$\sum_{i=1}^n \mathbb{P}(\mathbf{X}_i \neq \mathbf{X}_i^{M_\delta}) = \sum_{i=1}^n \mathbb{P}(\|\mathbf{X}_i\| > M_\delta) \le \delta. \tag{4.7}$$

Combining (4.5), (4.6), and (4.7), we obtain

$$\mathbb{P}(\|\hat{\mathbf{\Sigma}}_0 - \mathbb{E}\hat{\mathbf{\Sigma}}_0\| \ge x + \sqrt{\delta/n}) \le 2p\exp\left(-\frac{n^2x^2}{A_1 n + A_2 M_\delta^2 + A_3 n x M_\delta}\right) + \delta$$

with the constants A_1, A_2, A_3 defined above.

Case II: Now we consider the case when $0 < m \le n-2$. Since $\mathbf{Z}_t := \boldsymbol{Y}_t\boldsymbol{Y}_{t+m}^\mathsf{T}$ is not symmetric for all $t \in \mathbb{Z}$, by applying matrix dilation (See Tropp (2015), Section 2.1.16 for more details), we define the symmetric version of $\mathbf{Z}_t^M$ as

$$\overline{\mathbf{Z}}_t^M := \begin{bmatrix} \mathbf{0} & \mathbf{Z}_t^M \\ (\mathbf{Z}_t^M)^\mathsf{T} & \mathbf{0} \end{bmatrix}.$$

Observe that $\lambda_{\max}(\overline{\mathbf{Z}}_t^M) = \|\overline{\mathbf{Z}}_t^M\| = \|\mathbf{Z}_t^M\|$. By Lemma 11, $\{\overline{\mathbf{Z}}_t^M\}_{t\in\mathbb{Z}}$ and $\{\overline{\mathbf{Z}}_t^M - \mathbb{E}\overline{\mathbf{Z}}_t^M\}_{t\in\mathbb{Z}}$ are also sequences of τ-mixing random matrices. Define

$$\nu^2_{\overline{\mathbf{Z}}^M} := \sup_{K\subseteq\{1,\ldots,n-m\}} \frac{1}{\text{card}\,(K)}\lambda_{\max}\left\{\mathbb{E}\left(\sum_{i\in K}\overline{\mathbf{Z}}_i^M - \mathbb{E}\overline{\mathbf{Z}}_i^M\right)^2\right\}.$$

Notice that $\nu^2_{\overline{\mathbf{Z}}^M}$ and $\nu^2_{\mathbf{Z}^M}$ have the same upper bound since spectral norm of block diagonal matrix is less than or equal to the spectral norm of each block.

Now we apply similar arguments in Case I to $\{\overline{\mathbf{Z}}_t\}_{t\in\mathbb{Z}}$ and $\{\overline{\mathbf{Z}}_t^M\}_{t\in\mathbb{Z}}$.

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n-m}\left\|\sum_{i=1}^{n-m}(\mathbf{Z}_i-\mathbb{E}\mathbf{Z}_i)\right\|\geq x\right\}\\
\leq&\mathbb{P}\left[\lambda_{\max}\left\{\sum_{i=1}^{n-m}(\overline{\mathbf{Z}}_i^M-\mathbb{E}\overline{\mathbf{Z}}_i^M)\right\}\geq (n-m)x-\sum_{i=1}^{n-m}\|\mathbb{E}\overline{\mathbf{Z}}_i-\mathbb{E}\overline{\mathbf{Z}}_i^M\|\right]+\sum_{i=1}^{n-m}\mathbb{P}\Big(\mathbf{Z}_i\neq\mathbf{Z}_i^M\Big).
\end{aligned}$$

The rest is straightforward by using Theorem 10, Lemma 9, Lemma 11, and Lemma 12, and we thus finish the rest of the proof.

Lastly, we consider $\kappa_1\neq 1$. Notice that for any sequence $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ satisfying Assumptions **(A1)**-**(A3)**, the sequence $\{\boldsymbol{Y}_t/\kappa_1\}_{t\in\mathbb{Z}}$ will satisfy Assumptions **(A1)** automatically and Assumptions **(A2)**-**(A3)** with $\kappa_1=1$. Hence, applying the above to $\{\boldsymbol{Y}_t/\kappa_1\}_{t\in\mathbb{Z}}$ renders the results. This completes the proof of Proposition 8. □

4.4.2 Proof of Theorem 4

Proof. The proof of Theorem 4 consists of two cases.

Case I. When $m=0$, we first state a more general result of Gaussian process. Proposition 13 considers a general Gaussian process without further assumptions on the covariance and autocovariance matrices. The proof modifies that of Theorem 5.1 in van Handel (2017) with dependence among observations taken into account.

Proposition 13 (Proof in Section 4.4.2)**.** *Let $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ be a stationary sequence of mean-zero Gaussian random vectors with autocovariance matrices $\boldsymbol{\Sigma}_m$ for $0\leq m\leq n-1$. Then*

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_0-\boldsymbol{\Sigma}_0\|\leq&\frac{2}{n}\Bigg\{2\Big(\|\boldsymbol{\Sigma}_0\|_*+2\sum_{m=1}^{n-1}\|\boldsymbol{\Sigma}_m\|_*\Big)+\sqrt{2n\|\boldsymbol{\Sigma}_0\|\Big(\|\boldsymbol{\Sigma}_0\|_*+2\sum_{m=1}^{n-1}\|\boldsymbol{\Sigma}_m\|_*\Big)}\\
&+\sqrt{2n\Big(\|\boldsymbol{\Sigma}_0\|+2\sum_{m=1}^{n-1}\|\boldsymbol{\Sigma}_m\|\Big)\operatorname{tr}(\boldsymbol{\Sigma}_0)}\Bigg\},
\end{aligned}$$

where $\|\cdot\|_*$ *is the matrix nuclear norm.*

The rest of the proof is to show the geometric decay of spectral norm and nuclear norm of autocovariance matrices under Assumptions **(A2)**-**(A3)** in order to apply Proposition 13. It is obvious that $\kappa_1^2 \asymp \|\boldsymbol{\Sigma}_0\|$ and $\kappa_*^2 \asymp \mathrm{tr}(\boldsymbol{\Sigma}_0)$ when the process is a centered stationary Gaussian process. We first prove the geometric decay of spectral norm of autocovariance matrices. For any $0 \le m \le n-1$ and any integer j, by Assumption **(A3)**, there exists $\tilde{\boldsymbol{Y}}_{1+m}$ that is identically distributed as $\boldsymbol{Y}_{1+m}$, independent of $\boldsymbol{Y}_1$, and

$$\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|(\boldsymbol{Y}_{1+m} - \tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\boldsymbol{u}\|_{L(1+\epsilon)} \le \gamma_3\sqrt{\|\boldsymbol{\Sigma}_0\|}\exp\{-\gamma_4(m-1)\}.$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{\Sigma}_m\| =& \|\mathbb{E}\boldsymbol{Y}_1\boldsymbol{Y}_{1+m}^\mathsf{T}\| \\
=& \|\mathbb{E}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m} - \tilde{\boldsymbol{Y}}_{1+m} + \tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\| \\
=& \|\mathbb{E}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m} - \tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\| \\
\le& \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} |\mathbb{E}\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m} - \tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\boldsymbol{v}| \\
\le& C\|\boldsymbol{\Sigma}_0\|\exp\{-\gamma_4(m-1)\},
\end{aligned}$$

where the last inequality is followed by Assumption **(A3)** and $\gamma_3 = O(1)$ for some constant $C > 0$ only depending on ϵ, γ_3.

Similarly, by Assumption **(A2)**, there exists $\tilde{\boldsymbol{Y}}_{1+m}$ that is identically distributed as $\boldsymbol{Y}_{1+m}$, independent of $\boldsymbol{Y}_1$, and

$$\|\|\boldsymbol{Y}_{1+m} - \tilde{\boldsymbol{Y}}_{1+m}\|_2\|_{L(1+\epsilon)} \le \gamma_1\sqrt{\|\boldsymbol{\Sigma}_0\|}\exp\{-\gamma_2(m-1)\}.$$

Then,

$$\begin{aligned}
\|\boldsymbol{\Sigma}_m\|_* &= \sqrt{\operatorname{tr}(\boldsymbol{\Sigma}_m^\mathsf{T}\boldsymbol{\Sigma}_m)} \\
&= \sqrt{\operatorname{tr}\{\mathbb{E}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\boldsymbol{Y}_1^\mathsf{T}\mathbb{E}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\}} \\
&\le \sqrt{\operatorname{tr}\{\mathbb{E}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\boldsymbol{Y}_1^\mathsf{T}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\}} \\
&= \sqrt{\operatorname{tr}\{\mathbb{E}\boldsymbol{Y}_1^\mathsf{T}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\}} \\
&= \sqrt{\mathbb{E}\|\boldsymbol{Y}_1\|_2^2\|\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m}\|_2^2} \\
&\le \|\|\boldsymbol{Y}_1\|_2\|_{L(\frac{1+\epsilon}{\epsilon})}\|\|\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m}\|_2\|_{L(1+\epsilon)} \\
&\le C\operatorname{tr}(\boldsymbol{\Sigma}_0)\exp\{-\gamma_2(m-1)\},
\end{aligned}$$

where the third line is followed by the fact that $\mathbb{E}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\boldsymbol{Y}_1^\mathsf{T}\mathbb{E}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T} \preceq \mathbb{E}\boldsymbol{Y}_1^\mathsf{T}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}$ ("$\preceq$" is the Loewner partial order of Hermitian matrices), and both matrices are positive semi-definite, and the last line by Assumption **(A2)** and $\gamma_1 = O(\sqrt{r(\boldsymbol{\Sigma}_0)})$. Indeed, for any $\boldsymbol{u} \in \mathbb{R}^p$, $\mathbb{E}\{\boldsymbol{u}^\mathsf{T}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\}^2(\boldsymbol{Y}_1^\mathsf{T}\boldsymbol{Y}_1) = \sum_{j=1}^p \mathbb{E}\{\boldsymbol{u}^\mathsf{T}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\}^2\boldsymbol{Y}_{1,j}^2$ and $\mathbb{E}\{\boldsymbol{u}^\mathsf{T}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\}\boldsymbol{Y}_1^\mathsf{T}\mathbb{E}\boldsymbol{Y}_1(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})^\mathsf{T}\boldsymbol{u}$ $= \sum_{j=1}^p[\mathbb{E}\{\boldsymbol{u}^\mathsf{T}(\boldsymbol{Y}_{1+m}-\tilde{\boldsymbol{Y}}_{1+m})\boldsymbol{Y}_{1,j}\}]^2$. The result follows.

Case II. When $m > 0$, we denote $\bar{\boldsymbol{Y}}_i := (\boldsymbol{Y}_i^\mathsf{T}\ \boldsymbol{Y}_{i+m}^\mathsf{T})^\mathsf{T}$ for $1 \le i \le n-m$. It is obvious that $\{\bar{\boldsymbol{Y}}_i\}$ is a centered stationary Gaussian process satisfying Assumptions **(A2)**-**(A3)**. Denote $\bar{\boldsymbol{\Sigma}}_0 := \mathbb{E}\bar{\boldsymbol{Y}}_i\bar{\boldsymbol{Y}}_i^\mathsf{T}$ and notice that $\boldsymbol{\Sigma}_m$ is the off-diagnal block submatrix of $\bar{\boldsymbol{\Sigma}}_0$. By Case I and the fact that spectral norm of submatrix is bounded above by that of the full matrix, we obtain

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_m - \boldsymbol{\Sigma}_m\| \le C\|\bar{\boldsymbol{\Sigma}}_0\|\Big(\sqrt{\frac{r(\boldsymbol{\Sigma}_0)}{n-m}} + \frac{r(\boldsymbol{\Sigma}_0)}{n-m}\Big).$$

Notice that $\|\boldsymbol{\Sigma}_0\| \le \|\bar{\boldsymbol{\Sigma}}_0\| \le \|\boldsymbol{\Sigma}_0\| + \|\boldsymbol{\Sigma}_m\| \le 2\|\boldsymbol{\Sigma}_0\|$ since $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_m$ is positive semi-definite. This completes the proof. □

Proof of Proposition 13. The proof heavily depends on the following observation. Denote $\mathbf{Y} := (\boldsymbol{Y}_1 \dots \boldsymbol{Y}_n)$ and let $\tilde{\mathbf{Y}}$ be an independent copy of $\mathbf{Y}$. Then

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\| \le \frac{2}{n}\mathbb{E}\|\mathbf{Y}\tilde{\mathbf{Y}}^\mathsf{T}\|.$$

This is same as Lemma 5.2 in van Handel (2017) by noticing that the result holds without independence assumption.

Now we state the following two core lemmas used to complete the proof.

Lemma 14 (Proof in Section 4.4.3)**.** *We have*

$$\mathbb{E}\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\| \le \frac{2\sqrt{2}}{n}\Big\{\mathbb{E}\|\mathbf{Y}\| \cdot \sqrt{\operatorname{tr}\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)} + \sqrt{2\Big(\|\mathbf{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\mathbf{\Sigma}_d\|\Big)} \cdot \sqrt{n\operatorname{tr}(\mathbf{\Sigma}_0)}\Big\},$$

where $\tilde{\mathbf{\Sigma}}_d := (\mathbf{U}_d\mathbf{\Lambda}_d\mathbf{U}_d^\top + \mathbf{V}_d\mathbf{\Lambda}_d\mathbf{V}_d^\top)/2$. *Here* $\mathbf{U}_d, \mathbf{V}_d, \mathbf{\Lambda}_d$ *are left singular vectors, right singular vectors, and singular values of* $\mathbf{\Sigma}_d$ *for all* $1 \le d \le n-1$ *respectively.*

Lemma 15 (Proof in Section 4.4.3)**.** *We have*

$$\mathbb{E}\|\mathbf{Y}\| \le \sqrt{2\operatorname{tr}\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)} + \sqrt{2n\|\mathbf{\Sigma}_0\|},$$

where $\tilde{\mathbf{\Sigma}}_d$ *for all* $1 \le d \le n-1$ *are defined in Lemma 14.*

The proof of Proposition 13 completes by combining Lemma 14 and Lemma 15. □

4.4.3 Proofs of auxiliary lemmas

Proof of Lemma 9. By Lemma A.2 in Bunea and Xiao (2015), we have $\mathbb{E}\|\boldsymbol{Y}_t\|_2^{2k} \le (2k)^k\kappa_*^{2k}$ for $t \in \mathbb{Z}$. Hence

$$\|\|\boldsymbol{Y}_t\|_2^2 - \mathbb{E}\|\boldsymbol{Y}_t\|_2^2\|_{\psi_1} \le 2\|\|\boldsymbol{Y}_t\|_2^2\|_{\psi_1} \le 4\|\|\boldsymbol{Y}_t\|_2\|_{\psi_2}^2 \le 8\kappa_*^2.$$

Thus by property of sub-exponential random variable and Chernoff inequality, we have for any $x \ge 0$,

$$\mathbb{P}(\|\boldsymbol{Y}_t\|_2^2 - \mathbb{E}\|\boldsymbol{Y}_t\|_2^2 \ge x) \le \exp\Big\{-C\min\Big(\frac{x^2}{64\kappa_*^4}, \frac{x}{8\kappa_*^2}\Big)\Big\},$$

for some arbitary constant $C > 0$. Obviously, we have for all $x \ge 0$,

$$\mathbb{P}\{\|\boldsymbol{Y}_t\|_2^2 \ge 2\kappa_*^2 + 8\kappa_*^2(x + \sqrt{x})\} \le \exp(-Cx)$$

for some arbitary constant $C > 0$. This completes the proof. □

Proof of Lemma 11. We first show that $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ is a sequence of τ-mixing random vectors with geometric decay. Under Assumption **(A2)** (without loss of generality, take $j = 0$), there exists a sequence of random vectors $\{\tilde{\boldsymbol{Y}}_t\}_{t>0}$ which is independent of $\sigma(\{\boldsymbol{Y}_t\}_{t\le 0})$, identically distributed as $\{\boldsymbol{Y}_t\}_{t>0}$, and for any integer $t \ge 1$,

$$\|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)} \le \gamma_1\kappa_1 \exp\{-\gamma_2(t-1)\}$$

for some constant $\epsilon > 0$. Then for any $m \ge 0$,

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{Y}_t\boldsymbol{Y}_{t+m}^\mathsf{T} - \tilde{\boldsymbol{Y}}_t\tilde{\boldsymbol{Y}}_{t+m}^\mathsf{T}\| \\
=&\mathbb{E}\|\boldsymbol{Y}_t\boldsymbol{Y}_{t+m}^\mathsf{T} - \boldsymbol{Y}_t\tilde{\boldsymbol{Y}}_{t+m}^\mathsf{T} + \boldsymbol{Y}_t\tilde{\boldsymbol{Y}}_{t+m}^\mathsf{T} - \tilde{\boldsymbol{Y}}_t\tilde{\boldsymbol{Y}}_{t+m}^\mathsf{T}\| \\
\le&\mathbb{E}\|\boldsymbol{Y}_t(\boldsymbol{Y}_{t+m} - \tilde{\boldsymbol{Y}}_{t+m})^\mathsf{T}\| + \mathbb{E}\|(\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t)\tilde{\boldsymbol{Y}}_{t+m}^\mathsf{T}\| \\
\le&\|\|\boldsymbol{Y}_t\|_2\|_{L(\frac{1+\epsilon}{\epsilon})}\|\|\boldsymbol{Y}_{t+m} - \tilde{\boldsymbol{Y}}_{t+m}\|_2\|_{L(1+\epsilon)} + \|\|\boldsymbol{Y}_{t+m}\|_2\|_{L(\frac{1+\epsilon}{\epsilon})}\|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)} \\
\le&C\gamma_1\kappa_1\kappa_* \exp\{-\gamma_2(t-1)\},
\end{aligned}$$

where the fourth line is followed by Hölder's inequality and the fact that

$$\sup_{t\in\mathbb{Z}} \|\|\boldsymbol{Y}_t\|_2\|_{L(\alpha)} \le \sup_{t\in\mathbb{Z}}\sup_{\boldsymbol{u}\in\bar{\mathbb{S}}^{p-1}} \|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\|_{L(\alpha)} \le \sup_{t\in\mathbb{Z}}\sup_{\boldsymbol{u}\in\bar{\mathbb{S}}^{p-1}} \sqrt{\alpha}\|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\|_{\psi_2} \le \sqrt{\alpha}\kappa_*$$

for any $\alpha \ge 1$. Here $C > 0$ is some constant only depending on ϵ.

Now define $\tilde{\mathbf{X}}_t := \tilde{\boldsymbol{Y}}_t\tilde{\boldsymbol{Y}}_t^\mathsf{T}$ for any integer $t > 0$. It is obvious that $\{\tilde{\mathbf{X}}_t\}_{t>0}$ is independent of $\{\mathbf{X}_t\}_{t\le 0}$ and identically distributed as $\{\mathbf{X}_t\}_{t>0}$. By applying Lemma 18, for any indices $0 < k \le t_1 < \cdots < t_\ell$, we obtain

$$\tau\{\sigma(\{\mathbf{X}_t\}_{t\le 0}), (\mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_\ell}); \|\cdot\|\} \le \sum_{i=1}^{\ell} \mathbb{E}\|\mathbf{X}_{t_i} - \tilde{\mathbf{X}}_{t_i}\| \le C\gamma_1\kappa_1\kappa_*\ell \exp\{-\gamma_2(k-1)\}.$$

By definition of τ-mixing coefficient, this yields

$$\tau(k; \{\mathbf{X}_t\}_{t\in\mathbb{Z}}, \|\cdot\|) \le C\gamma_1\kappa_1\kappa_* \exp\{-\gamma_2(k-1)\}$$

for some constant $C > 0$ only depending on ϵ.

Now we proceed to prove τ-mixing properties for the "truncated version". The following lemma is needed.

Lemma 16 (Proof in Section 4.4.3)**.** *Let* $\boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \boldsymbol{v}_1,\ \boldsymbol{v}_2 \in \mathbb{R}^p$ *for* $p \geq 1$ *with unit length under* ℓ_2*-norm and* $\sigma_u \geq 0$*. Then the function*

$$f(\sigma_v) = \|\sigma_v \boldsymbol{v}_1 \boldsymbol{v}_2^\mathsf{T} - \sigma_u \boldsymbol{u}_1 \boldsymbol{u}_2^\mathsf{T}\|$$

is non-decreasing in the range $\sigma_v \in [\sigma_u, \infty]$*. In particular, for any* $M \geq 0$ *such that* $M \leq \sigma_u,\ M \leq \sigma_v$*, we have*

$$\|M\boldsymbol{v}_1 \boldsymbol{v}_2^\mathsf{T} - M\boldsymbol{u}_1 \boldsymbol{u}_2^\mathsf{T}\| \leq \|\sigma_v \boldsymbol{v}_1 \boldsymbol{v}_2^\mathsf{T} - \sigma_u \boldsymbol{u}_1 \boldsymbol{u}_2^\mathsf{T}\|.$$

Now consider three cases.

(1) When $\|\mathbf{X}_t\| \leq M$ and $\|\tilde{\mathbf{X}}_t\| \leq M$, $\|\mathbf{X}_t^M - \tilde{\mathbf{X}}_t^M\| = \|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$.

(2) When $\|\mathbf{X}_t\| \leq M$ and $\|\tilde{\mathbf{X}}_t\| > M$, we have

$$\mathbf{X}_t^M = \mathbf{X}_t = \|\boldsymbol{Y}_t\|_2^2 \frac{\boldsymbol{Y}_t}{\|\boldsymbol{Y}_t\|_2} \frac{\boldsymbol{Y}_t^\mathsf{T}}{\|\boldsymbol{Y}_t\|_2} \quad \text{and} \quad \tilde{\mathbf{X}}_t^M = M \frac{\tilde{\boldsymbol{Y}}_t}{\|\tilde{\boldsymbol{Y}}_t\|_2} \frac{\tilde{\boldsymbol{Y}}_t^\mathsf{T}}{\|\tilde{\boldsymbol{Y}}_t\|_2}.$$

Since $\frac{\boldsymbol{Y}_t}{\|\boldsymbol{Y}_t\|_2}$, $\frac{\tilde{\boldsymbol{Y}}_t}{\|\tilde{\boldsymbol{Y}}_t\|_2}$ have unit length and $\|\boldsymbol{Y}_t\|_2^2 \leq M < \|\tilde{\boldsymbol{Y}}_t\|_2^2$, we have $\|\mathbf{X}_t^M - \tilde{\mathbf{X}}_t^M\| \leq \|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ by Lemma 16. By symmetry, the same argument also applies to the case where $\|\mathbf{X}_t\| > M$ and $\|\tilde{\mathbf{X}}_t\| \leq M$.

(3) When $\|\mathbf{X}_t\| > M$ and $\|\tilde{\mathbf{X}}_t\| > M$, we have $\mathbf{X}_t^M = M \frac{\boldsymbol{Y}_t}{\|\boldsymbol{Y}_t\|_2} \frac{\boldsymbol{Y}_t^\mathsf{T}}{\|\boldsymbol{Y}_t\|_2}$ and $\tilde{\mathbf{X}}_t^M = M \frac{\tilde{\boldsymbol{Y}}_t}{\|\tilde{\boldsymbol{Y}}_t\|_2} \frac{\tilde{\boldsymbol{Y}}_t^\mathsf{T}}{\|\tilde{\boldsymbol{Y}}_t\|_2}$. Again by Lemma 16, we have $\|\mathbf{X}_t^M - \tilde{\mathbf{X}}_t^M\| \leq \|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$.

By combining three cases, $\|\mathbf{X}_t^M - \tilde{\mathbf{X}}_t^M\| \leq \|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ always holds, and hence $\mathbb{E}\|\mathbf{X}_t^M - \tilde{\mathbf{X}}_t^M\| \leq \mathbb{E}\|\mathbf{X}_t - \tilde{\mathbf{X}}_t\|$ for any $t \geq 1$. Hence for any indices $0 < k \leq t_1 < \cdots < t_\ell$, by Lemma 18, we have

$$\tau\{\sigma(\{\mathbf{X}_t^M\}_{t \leq 0}), (\mathbf{X}_{t_1}^M, \ldots, \mathbf{X}_{t_\ell}^M); \|\cdot\|\} \leq C\gamma_1\kappa_1\kappa_*\ell \exp\{-\gamma_2(k-1)\}$$

for some constant $C > 0$ only depending on ϵ. By definition of τ-mixing coefficient, this yields

$$\tau(k; \{\mathbf{X}_t^M\}_{t \in \mathbb{Z}}, \|\cdot\|) \leq C\gamma_1\kappa_1\kappa_* \exp\{-\gamma_2(k-1)\}$$

for some constant $C > 0$ only depending on ϵ. Notice that $\mathbb{E}\|\mathbf{X}_t^M - \mathbb{E}\mathbf{X}_t^M - (\tilde{\mathbf{X}}_t^M - \mathbb{E}\tilde{\mathbf{X}}_t^M)\| = \mathbb{E}\|\mathbf{X}_t^M - \tilde{\mathbf{X}}_t^M\|$ since $\mathbb{E}\tilde{\mathbf{X}}_t^M = \mathbb{E}\mathbf{X}_t^M$ for any $t \geq 1$. The τ-mixing property stated above applies to $\{\mathbf{X}_t^M - \mathbb{E}\mathbf{X}_t^M\}$ directly.

Similar arguments apply to $\{\mathbf{Z}_t^M\}_{t\in\mathbb{Z}}$ and $\{\mathbf{Z}_t^M - \mathbb{E}\mathbf{Z}_t^M\}_{t\in\mathbb{Z}}$ so we omit the details. This completes the proof. □

Proof of Lemma 12. The proof consists of two steps.

Step I. We first provide an upper bound for $\nu_{\mathbf{X}}^2$. Without loss of generality, we only consider $\|\mathbb{E}(\mathbf{X}_0 - \mathbb{E}\mathbf{X}_0)(\mathbf{X}_k - \mathbb{E}\mathbf{X}_k)\|$ for $k \geq 0$. Under Assumptions **(A2)**-**(A3)**, there exists $\tilde{\boldsymbol{Y}}_k$ where $\tilde{\boldsymbol{Y}}_k$ is independent of $\sigma(\{\boldsymbol{Y}_t\}_{t\leq 0})$, identically distributed as $\boldsymbol{Y}_k$, and

$$\begin{aligned}
\|\|\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k\|_2\|_{L(1+\epsilon)} &\leq \gamma_1\kappa_1 \exp\{-\gamma_2(k-1)\},\\
\|(\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k)^\mathsf{T}\boldsymbol{u}\|_{L(1+\epsilon)} &\leq \gamma_3\kappa_1 \exp\{-\gamma_4(k-1)\}
\end{aligned}$$

for constants $\gamma_1, \gamma_2, \gamma_3, \gamma_4 > 0$ in Assumptions **(A2)**-**(A3)**.

For $k = 0$, we have

$$\|\mathbb{E}\mathbf{X}_0\mathbf{X}_0 - \mathbb{E}\mathbf{X}_0\mathbb{E}\mathbf{X}_0\| \leq C(\kappa_1^4 + \kappa_1^2\kappa_*^2)$$

by Assumption **(A1)** for some universal constant $C > 0$. For $k > 0$, we obtain

$$\begin{aligned}
\|\mathbb{E}\mathbf{X}_0\mathbf{X}_k - \mathbb{E}\mathbf{X}_0\mathbb{E}\mathbf{X}_k\| =& \|\mathbb{E}\mathbf{X}_0\mathbf{X}_k - \mathbb{E}\mathbf{X}_0\tilde{\mathbf{X}}_k\|\\
=& \|\mathbb{E}\mathbf{X}_0(\mathbf{X}_k - \tilde{\mathbf{X}}_k)\|\\
=& \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_0\boldsymbol{Y}_0^\mathsf{T}(\boldsymbol{Y}_k\boldsymbol{Y}_k^\mathsf{T} - \tilde{\boldsymbol{Y}}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T})\boldsymbol{v}|\\
\leq& \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \mathbb{E}|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_0\boldsymbol{Y}_0^\mathsf{T}\boldsymbol{Y}_k(\boldsymbol{Y}_k^\mathsf{T} - \tilde{\boldsymbol{Y}}_k^\mathsf{T})\boldsymbol{v} + \boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_0\boldsymbol{Y}_0^\mathsf{T}(\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k)\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v}|\\
\leq& \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} \{\mathbb{E}|\boldsymbol{Y}_0^\mathsf{T}\boldsymbol{Y}_k|^{\frac{3(1+\epsilon)}{2\epsilon}}\}^{\frac{2\epsilon}{3(1+\epsilon)}}\|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_0\|_{L(\frac{3(1+\epsilon)}{\epsilon})}\|(\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k)^\mathsf{T}\boldsymbol{v}\|_{L(1+\epsilon)} +\\
& \{\mathbb{E}|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_0\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v}|^{\frac{3(1+\epsilon)}{2\epsilon}}\}^{\frac{2\epsilon}{3(1+\epsilon)}}\|\|\boldsymbol{Y}_0\|_2\|_{L(\frac{3(1+\epsilon)}{\epsilon})}\|\|\boldsymbol{Y}_k - \tilde{\boldsymbol{Y}}_k\|_2\|_{L(1+\epsilon)}\\
\leq& C\kappa_1^2\kappa_*(\kappa_*\gamma_3 + \kappa_1\gamma_1)\exp\{-\min(\gamma_2,\gamma_4)(k-1)\},
\end{aligned}$$

where the first line is followed by $\mathbb{E}\mathbf{X}_k = \mathbb{E}\tilde{\mathbf{X}}_k$, fifth line by Hölder's inequality, and sixth line by Assumptions **(A1)**-**(A3)** for some constant $C > 0$ only depending on ϵ.

Hence for any $K \subseteq \{1, \ldots, n\}$,

$$\begin{aligned}
&\frac{1}{\operatorname{card}(K)} \lambda_{\max} \left\{ \mathbb{E} \left(\sum_{i \in K} \mathbf{X}_i - \mathbb{E}\mathbf{X}_i \right)^2 \right\} \\
\leq & \frac{1}{\operatorname{card}(K)} \left\| \sum_{i,j \in K} \mathbb{E}(\mathbf{X}_i - \mathbb{E}\mathbf{X}_i)(\mathbf{X}_j - \mathbb{E}\mathbf{X}_j) \right\| \\
\leq & \frac{1}{\operatorname{card}(K)} \sum_{i,j \in K} \|\mathbb{E}(\mathbf{X}_i - \mathbb{E}\mathbf{X}_i)(\mathbf{X}_j - \mathbb{E}\mathbf{X}_j)\| \\
\leq & C \left[\kappa_1^4 + \kappa_1^2 \kappa_*^2 + \frac{\kappa_1^2 \kappa_* (\kappa_* \gamma_3 + \kappa_1 \gamma_1)}{\operatorname{card}(K)} \sum_{i,j \in K, i \neq j} \exp\{-\min(\gamma_2, \gamma_4)(|i-j|-1)\} \right] \\
\leq & C \left[\frac{\kappa_1^2 \{\kappa_1^2 + \kappa_1 \kappa_* \gamma_1 + \kappa_*^2 (\gamma_3 + 1)\}}{1 - \exp(-\min\{\gamma_2, \gamma_4\})} \right].
\end{aligned}$$

Step II. We first bound $\nu^2_{\mathbf{X}^M}$. By definition, we have

$$\left\| \mathbb{E} \left(\sum_{i \in K} \mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M \right)^2 \right\| = \left\| \sum_{i,j \in K} \mathbb{E}(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M)(\mathbf{X}_j^M - \mathbb{E}\mathbf{X}_j^M) \right\| = \left\| \sum_{i,j \in K} (\mathbb{E}\mathbf{X}_i^M \mathbf{X}_j^M - \mathbb{E}\mathbf{X}_i^M \mathbb{E}\mathbf{X}_j^M) \right\|.$$

Without loss of generality, we consider $\|\mathbb{E}\mathbf{X}_0^M \mathbf{X}_k^M - \mathbb{E}\mathbf{X}_0^M \mathbb{E}\mathbf{X}_k^M\|$ for $k \geq 0$. Let $\tilde{\mathbf{X}}_k^M$ be defined as in the proof of Lemma 11. Then $\tilde{\mathbf{X}}_k^M$ is independent of $\tilde{\mathbf{X}}_0^M$ and distributed as $\mathbf{X}_k^M$. Hence

$$\|\mathbb{E}\mathbf{X}_0^M \mathbf{X}_k^M - \mathbb{E}\mathbf{X}_0^M \mathbb{E}\mathbf{X}_k^M\| = \|\mathbb{E}\mathbf{X}_0^M \mathbf{X}_k^M - \mathbb{E}\mathbf{X}_0^M \mathbb{E}\tilde{\mathbf{X}}_k^M\|.$$

Then we could rewrite

$$\begin{aligned}
\|\mathbb{E}\mathbf{X}_0^M \mathbf{X}_k^M - \mathbb{E}\mathbf{X}_0^M \mathbb{E}\tilde{\mathbf{X}}_k^M\| =& \|\mathbb{E}\mathbf{X}_0 \mathbf{X}_k \zeta_0 \zeta_k - \mathbb{E}\mathbf{X}_0 \tilde{\mathbf{X}}_k \zeta_0 \tilde{\zeta}_k\| \\
=& \|\mathbb{E}\mathbf{X}_0 (\mathbf{X}_k - \tilde{\mathbf{X}}_k) \zeta_0 \zeta_k + \mathbb{E}\mathbf{X}_0 \tilde{\mathbf{X}}_k \zeta_0 (\zeta_k - \tilde{\zeta}_k)\|,
\end{aligned}$$

where $\zeta_i = \frac{M \wedge \|\mathbf{X}_i\|}{\|\mathbf{X}_i\|}$, $\tilde{\zeta}_i = \frac{M \wedge \|\tilde{\mathbf{X}}_i\|}{\|\tilde{\mathbf{X}}_i\|}$. Since ζ_0, ζ_k are bounded by 1, we have

$$\begin{aligned}
\|\mathbb{E}\mathbf{X}_0 (\mathbf{X}_k - \tilde{\mathbf{X}}_k) \zeta_0 \zeta_k\| &= \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{p-1}} \mathbb{E} |\boldsymbol{u}^\mathsf{T} \mathbf{X}_0 (\mathbf{X}_k - \tilde{\mathbf{X}}_k) \zeta_0 \zeta_k \boldsymbol{v}| \\
&\leq \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{p-1}} \mathbb{E} |\boldsymbol{u}^\mathsf{T} \mathbf{X}_0 (\mathbf{X}_k - \tilde{\mathbf{X}}_k) \boldsymbol{v}| \\
&= \|\mathbb{E}\mathbf{X}_0 (\mathbf{X}_k - \tilde{\mathbf{X}}_k)\| \\
&\leq C \kappa_1^2 (\kappa_1 \kappa_* \gamma_1 + \kappa_*^2 \gamma_3) \exp\{-\min(\gamma_2, \gamma_4)(k-1)\},
\end{aligned}$$

where the last inequality is from result in Step I for some constant $C > 0$ only depending on ϵ.

On the other hand, by applying Hölder's inequality, we have

$$\begin{aligned}\|\mathbb{E}\mathbf{X}_0\tilde{\mathbf{X}}_k\zeta_0(\zeta_k-\tilde{\zeta}_k)\| &= \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\mathbb{E}|\boldsymbol{u}^\top\mathbf{X}_0\tilde{\mathbf{X}}_k\boldsymbol{v}||\zeta_k-\tilde{\zeta}_k| \\ &\le \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\{\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{Y}_0\boldsymbol{Y}_0^\top\tilde{\boldsymbol{Y}}_k\tilde{\boldsymbol{Y}}_k^\top\boldsymbol{v}|^{\frac{5(1+\epsilon)}{4\epsilon}}\}^{\frac{4\epsilon}{5(1+\epsilon)}}\{\mathbb{E}|\zeta_k-\tilde{\zeta}_k|^{\frac{5(1+\epsilon)}{5+\epsilon}}\}^{\frac{5+\epsilon}{5(1+\epsilon)}}.\end{aligned}$$

Hence, for any $\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}$,

$$\begin{aligned}\{\mathbb{E}|\boldsymbol{u}^\top\boldsymbol{Y}_0\boldsymbol{Y}_0^\top\tilde{\boldsymbol{Y}}_k\tilde{\boldsymbol{Y}}_k^\top\boldsymbol{v}|^{\frac{5(1+\epsilon)}{4\epsilon}}\}^{\frac{4\epsilon}{5(1+\epsilon)}} &\le \|\boldsymbol{u}^\top\boldsymbol{Y}_0\|_{L(\frac{5(1+\epsilon)}{\epsilon})}\|\boldsymbol{u}^\top\tilde{\boldsymbol{Y}}_k\|_{L(\frac{5(1+\epsilon)}{\epsilon})}\|\|\boldsymbol{Y}_0\|_2\|_{L(\frac{5(1+\epsilon)}{\epsilon})}\|\|\tilde{\boldsymbol{Y}}_k\|_2\|_{L(\frac{5(1+\epsilon)}{\epsilon})} \\ &\le C\kappa_1^2\kappa_*^2,\end{aligned}$$

where the first line follows by Hölder's inequality and the last line by Assumption **(A1)** for some constant $C > 0$ only depending on ϵ.

Next, we need to bound $\|\zeta_k-\tilde{\zeta}_k\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})}$. For the sake of presentation clearness, we denote $a_k := \|\mathbf{X}_k\|$ and $\tilde{a}_k := \|\tilde{\mathbf{X}}_k\|$, and rewrite

$$\begin{aligned}&\|\zeta_k-\tilde{\zeta}_k\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \\ =&\left\|M\left|\frac{1}{a_k}-\frac{1}{\tilde{a}_k}\right|\mathbf{1}_{\{a_k>M,\tilde{a}_k>M\}}+\left(1-\frac{M}{a_k}\right)\mathbf{1}_{\{a_k>M,\tilde{a}_k\le M\}}+\left(1-\frac{M}{\tilde{a}_k}\right)\mathbf{1}_{\{a_k\le M,\tilde{a}_k>M\}}\right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \\ \le&\left\|M\left|\frac{1}{a_k}-\frac{1}{\tilde{a}_k}\right|\mathbf{1}_{\{a_k>M,\tilde{a}_k>M\}}\right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})}+\left\|\left(1-\frac{M}{a_k}\right)\mathbf{1}_{\{a_k>M,\tilde{a}_k\le M\}}\right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \\ &+\left\|\left(1-\frac{M}{\tilde{a}_k}\right)\mathbf{1}_{\{a_k\le M,\tilde{a}_k>M\}}\right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})},\end{aligned} \tag{4.8}$$

where the last inequality follows by the fact that $\|\cdot\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})}$ is a norm for $\epsilon > 0$.

For the first term, we have

$$\begin{aligned}\left\|M\left|\frac{1}{a_k}-\frac{1}{\tilde{a}_k}\right|\mathbf{1}_{\{a_k>M,\tilde{a}_k>M\}}\right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} &= \left\|M\left|\frac{\tilde{a}_k-a_k}{a_k\tilde{a}_k}\right|\mathbf{1}_{\{a_k>M,\tilde{a}_k>M\}}\right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \\ &\le \frac{1}{M}\{\mathbb{E}|\tilde{a}_k-a_k|^{\frac{5(1+\epsilon)}{5+\epsilon}}\}^{\frac{5+\epsilon}{5(1+\epsilon)}} \\ &\le \frac{1}{M}\{\mathbb{E}\|\mathbf{X}_k-\tilde{\mathbf{X}}_k\|^{\frac{5(1+\epsilon)}{5+\epsilon}}\}^{\frac{5+\epsilon}{5(1+\epsilon)}} \\ &\le C\gamma_1\kappa_1\kappa_*\exp\{-\gamma_2(k-1)\}/M,\end{aligned}$$

where the last inequality is followed by Lemma 11 for some constant $C > 0$ only depending on ϵ. With the chosen $M \geq C\gamma_1\kappa_1\kappa_*$, we have

$$\left\| M\left|\frac{1}{a_k} - \frac{1}{\tilde{a}_k}\right| \mathbf{1}_{\{a_k > M, \tilde{a}_k > M\}} \right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \leq \exp\{-\gamma_2(k-1)\}.$$

For the second term, taking any $\epsilon_k > 0$, we have

$$\begin{aligned}
&\left\| \left(1 - \frac{M}{a_k}\right) \mathbf{1}_{\{a_k > M, \tilde{a}_k \leq M\}} \right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \\
=&\left\| \left(1 - \frac{M}{M+\epsilon_k}\right) \mathbf{1}_{\{M < a_k \leq M+\epsilon_k, \tilde{a}_k \leq M\}} \right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} + \left\| \left(1 - \frac{M}{a_k}\right) \mathbf{1}_{\{a_k > M+\epsilon_k, \tilde{a}_k \leq M\}} \right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \\
\leq&\frac{\epsilon_k}{M} + \left\| \mathbf{1}_{\{a_k > M+\epsilon_k, \tilde{a}_k \leq M\}} \right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \leq \frac{\epsilon_k}{M} + \{\mathbb{P}(|a_k - \tilde{a}_k| > \epsilon_k)\}^{\frac{5+\epsilon}{5(1+\epsilon)}}.
\end{aligned}$$

By Markov inequality and Lemma 11, we have

$$\mathbb{P}(|a_k - \tilde{a}_k| > \epsilon_k) \leq \frac{\mathbb{E}\|\mathbf{X}_k - \tilde{\mathbf{X}}_k\|}{\epsilon_k} \leq \frac{C\gamma_1\kappa_1\kappa_* \exp\{-\gamma_2(k-1)\}}{\epsilon_k}$$

for some constant $C > 0$ only depending on ϵ. Taking $\epsilon_k = C\gamma_1\kappa_1\kappa_* \exp\{-\frac{5+\epsilon}{6\epsilon+10}\gamma_2(k-1)\}$, we obtain

$$\left\| \left(1 - \frac{M}{a_k}\right) \mathbf{1}_{\{a_k > M, \tilde{a}_k \leq M\}} \right\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \leq 2\exp\left\{ -\frac{5+\epsilon}{6\epsilon+10}\gamma_2(k-1) \right\}.$$

The third term follows by symmetry. Putting together, we have for $k > 0$,

$$\begin{aligned}
&\|\zeta_k - \tilde{\zeta}_k\|_{L(\frac{5(1+\epsilon)}{5+\epsilon})} \leq C\exp\left\{ -\frac{5+\epsilon}{6\epsilon+10}\gamma_2(k-1) \right\}, \\
&\|\mathbb{E}\mathbf{X}_0\tilde{\mathbf{X}}_k\zeta_0(\zeta_k - \tilde{\zeta}_k)\| \leq C\kappa_1^2\kappa_*^2\exp\left\{ -\frac{5+\epsilon}{6\epsilon+10}\gamma_2(k-1) \right\}, \\
&\|\mathbb{E}\mathbf{X}_0^M\mathbf{X}_k^M - \mathbb{E}\mathbf{X}_0^M\mathbb{E}\tilde{\mathbf{X}}_k^M\| \leq C\kappa_1^2\{\kappa_1\kappa_*\gamma_1 + \kappa_*^2(\gamma_3+1)\}\exp\left\{ -\min\left(\frac{5+\epsilon}{6\epsilon+10}\gamma_2, \gamma_4\right)(k-1) \right\}
\end{aligned}$$

for some constant $C > 0$ only depending on ϵ. Hence for any $K \subseteq \{1, \ldots, n\}$,

$$
\begin{aligned}
&\frac{1}{\mathrm{card}\,(K)}\lambda_{\max}\left\{\mathbb{E}\left(\sum_{i\in K}\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M\right)^2\right\} \\
\leq&\frac{1}{\mathrm{card}\,(K)}\left\|\sum_{i,j\in K}\mathbb{E}(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M)(\mathbf{X}_j^M - \mathbb{E}\mathbf{X}_j^M)\right\| \\
\leq&\frac{1}{\mathrm{card}\,(K)}\sum_{i,j\in K}\|\mathbb{E}(\mathbf{X}_i^M - \mathbb{E}\mathbf{X}_i^M)(\mathbf{X}_j^M - \mathbb{E}\mathbf{X}_j^M)\| \\
\leq&C\left[\kappa_1^4 + \kappa_*^2\kappa_1^2 + \frac{\kappa_1^2\{\kappa_1\kappa_*\gamma_1 + \kappa_*^2(\gamma_3+1)\}}{\mathrm{card}(K)}\sum_{i,j\in K, i\neq j}\exp\left\{-\min\left(\frac{5+\epsilon}{6\epsilon+10}\gamma_2, \gamma_4\right)(|i-j|-1)\right\}\right] \\
\leq&C\frac{\kappa_1^2\{\kappa_1^2 + \kappa_1\kappa_*\gamma_1 + \kappa_*^2(\gamma_3+2)\}}{1-\exp\{-\min(\frac{5+\epsilon}{6\epsilon+10}\gamma_2, \gamma_4)\}}
\end{aligned}
$$

for some constant $C > 0$ only depending on ϵ.

Similar arguments apply to $\nu^2_{\mathbf{Z}^M}$ so we omit the details. This completes the proof. □

Proof of Lemma 16. Fix $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^p$ with unit length and $\sigma_u \geq 0$. For any $\sigma_v \geq \sigma_u$, we perform singular value decomposition for matrix $\mathbf{X}(\sigma_v) := \sigma_u\boldsymbol{u}_1\boldsymbol{u}_2^\mathsf{T} - \sigma_v\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T}$. According to Equation (8) in Brand (2006), the non-zero singular values of $\mathbf{X}(\sigma_v)$ are identical to those of

$$
\mathbf{S}(\sigma_v) = \begin{bmatrix} \sigma_u - \sigma_v\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T}\boldsymbol{u}_2 & -\sigma_v\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\|\boldsymbol{v}_2 - \boldsymbol{u}_2\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2\|_2 \\ \sigma_v\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2\|\boldsymbol{v}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\|_2 & \sigma_v^2\|\boldsymbol{v}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\|_2\|\boldsymbol{v}_2 - \boldsymbol{u}_2\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2\|_2 \end{bmatrix}.
$$

For simplicity, denote $w = \boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T}\boldsymbol{u}_2$, $\tilde{\boldsymbol{v}}_1 = \boldsymbol{v}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1$, $\tilde{\boldsymbol{u}}_1 = \boldsymbol{v}_2 - \boldsymbol{u}_2\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2$. Hence $\mathbf{S}(\sigma_v)$ could be rewritten as

$$
\mathbf{S}(\sigma_v) = \begin{bmatrix} \sigma_u - \sigma_v w & -\sigma_v\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\|\tilde{\boldsymbol{v}}_2\|_2 \\ \sigma_v\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2\|\tilde{\boldsymbol{v}}_1\|_2 & \sigma_v^2\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2 \end{bmatrix}.
$$

Using the calculation on Page 86 in Blinn (1996), $\|\mathbf{S}(\sigma_v)\| = Q(\sigma_v) + R(\sigma_v)$, where

$$
\begin{aligned}
Q(\sigma_v) &:= \sqrt{(\sigma_u - \sigma_v w + \sigma_v\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2)^2 + \sigma_v^2(\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\|\tilde{\boldsymbol{v}}_2\|_2 + \boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2\|\tilde{\boldsymbol{v}}_1\|_2)^2}/2, \\
R(\sigma_v) &:= \sqrt{(\sigma_u - \sigma_v w - \sigma_v\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2)^2 + \sigma_v^2(\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1\|\tilde{\boldsymbol{v}}_2\|_2 - \boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2\|\tilde{\boldsymbol{v}}_1\|_2)^2}/2.
\end{aligned}
$$

We are left to show that both Q and R are non-deceasing function of $\sigma_v \in [\sigma_u, \infty]$. By differentiating Q, R with respect to σ_v, we obtain

$$\frac{dQ}{d\sigma_v} = c_Q(\sigma_v)[\sigma_u(\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2 - w) + \sigma_v\{w^2 + \|\tilde{\boldsymbol{v}}_1\|_2^2\|\tilde{\boldsymbol{v}}_2\|_2^2 + (\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1)^2\|\tilde{\boldsymbol{v}}_2\|_2^2 + (\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2)^2\|\tilde{\boldsymbol{v}}_1\|_2^2\}],$$
$$\frac{dR}{d\sigma_v} = c_Q(\sigma_v)[-\sigma_u(\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2 + w) + \sigma_v\{w^2 + \|\tilde{\boldsymbol{v}}_1\|_2^2\|\tilde{\boldsymbol{v}}_2\|_2^2 + (\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1)^2\|\tilde{\boldsymbol{v}}_2\|_2^2 + (\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2)^2\|\tilde{\boldsymbol{v}}_1\|_2^2\}]$$

for some nonnegative constants $c_Q(\sigma_v),\ c_R(\sigma_v)$.

By simple algebra, we have $w^2 + \|\tilde{\boldsymbol{v}}_1\|_2^2\|\tilde{\boldsymbol{v}}_2\|_2^2 + (\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1)^2\|\tilde{\boldsymbol{v}}_2\|_2^2 + (\boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2)^2\|\tilde{\boldsymbol{v}}_1\|_2^2 = 1$ so that

$$\frac{dQ}{d\sigma_v} = c_Q(\sigma_v)[\sigma_u(\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2 - w) + \sigma_v].$$

Moreover, since $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^p$ are all length 1, we have $|w| \leq 1$ by Cauchy-Schwartz. Hence by the fact that $\sigma_v \geq \sigma_u \geq 0$, we have $\frac{dQ}{d\sigma_v} \geq 0$. On the other hand, denote $a := \boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1$ and $b := \boldsymbol{u}_2^\mathsf{T}\boldsymbol{v}_2$ and again by Cauchy-Schwartz we have $|a| \leq 1,\ |b| \leq 1$. In addition, we have

$$\begin{aligned}
\|\tilde{\boldsymbol{v}}_1\|_2 &= \sqrt{(\boldsymbol{v}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1)^\mathsf{T}(\boldsymbol{v}_1 - \boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1)} \\
&= \sqrt{\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_1 - \boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1 - \boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1 + \boldsymbol{v}_1^\mathsf{T}\boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{u}_1\boldsymbol{u}_1^\mathsf{T}\boldsymbol{v}_1} \\
&= \sqrt{1 - a^2}.
\end{aligned}$$

Similarly, we have $\|\tilde{\boldsymbol{v}}_2\|_2 = \sqrt{1 - b^2}$. Then

$$\begin{aligned}
\frac{dR}{d\sigma_v} &= c_Q(\sigma_v)\{\sigma_v - \sigma_u(\|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2 + w)\} \\
&\geq c_Q(\sigma_v)\sigma_u(1 - \|\tilde{\boldsymbol{v}}_1\|_2\|\tilde{\boldsymbol{v}}_2\|_2 - w) \\
&\geq c_Q(\sigma_v)\sigma_u(1 - \sqrt{(1 - a^2)(1 - b^2)} - ab).
\end{aligned}$$

Since $(1 - ab)^2 \geq (1 - a^2)(1 - b^2)$ and $|ab| \leq 1$, we obtain $\frac{dR}{d\sigma_v} \geq 0$. Therefore we have shown that $\|\mathbf{S}(\sigma_v)\| = Q(\sigma_v) + R(\sigma_v)$ is a non-decreasing function with respect to σ_v.

Obviously $\|M\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T} - M\boldsymbol{u}_1\boldsymbol{u}_2^\mathsf{T}\| \leq \|\sigma_u\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T} - \sigma_u\boldsymbol{u}_1\boldsymbol{u}_2^\mathsf{T}\|$ since $0 < M \leq \sigma_u$. Applying the monotonicity property proved above, we have $\|\sigma_u\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T} - \sigma_u\boldsymbol{u}_1\boldsymbol{u}_2^\mathsf{T}\| \leq \|\sigma_u\boldsymbol{v}_1\boldsymbol{v}_2^\mathsf{T} - \sigma_v\boldsymbol{u}_1\boldsymbol{u}_2^\mathsf{T}\|$. This completes the proof. □

Proof of Lemma 14. By the observation in the proof of Proposition 13, we have

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\| \le \frac{2}{n}\mathbb{E}\|\mathbf{Y}\tilde{\mathbf{Y}}^\mathsf{T}\| = \frac{2}{n}\mathbb{E}\Big(\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}}\sum_{k=1}^n \boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v}\Big) := \frac{2}{n}\mathbb{E}\Big(\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} W_{\boldsymbol{u},\boldsymbol{v}}\Big).$$

Now consider

$$\begin{aligned}
(W_{\boldsymbol{u},\boldsymbol{v}} - W_{\boldsymbol{u}',\boldsymbol{v}'})^2 =& \Big(\sum_{k=1}^n \boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v} - \sum_{k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v}'\Big)^2 \\
=& \Big(\sum_{k=1}^n \boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v} - \sum_{k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v} + \sum_{k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v} - \sum_{k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v}'\Big)^2 \\
=& \Big(\sum_{k=1}^n (\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v} + \sum_{k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}(\boldsymbol{v}-\boldsymbol{v}')\Big)^2 \\
\le& 2\Big(\sum_{k=1}^n (\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}\boldsymbol{v}\Big)^2 + 2\Big(\sum_{k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\tilde{\boldsymbol{Y}}_k^\mathsf{T}(\boldsymbol{v}-\boldsymbol{v}')\Big)^2 \\
=& 2\sum_{d=0}^{n-1}\sum_{|j-k|=d} (\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\boldsymbol{Y}_j\cdot(\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\boldsymbol{Y}_k\cdot\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{Y}}_j\cdot\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{Y}}_k \\
&+ 2\sum_{d=0}^{n-1}\sum_{|j-k|=d} \boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_j\cdot\boldsymbol{u}'^\mathsf{T}\boldsymbol{Y}_k\cdot(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\tilde{\boldsymbol{Y}}_j\cdot(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\tilde{\boldsymbol{Y}}_k.
\end{aligned}$$

Now denote the conditional expectation $\mathbb{E}_{\tilde{\mathbf{Y}}} := \mathbb{E}(\cdot|\tilde{\mathbf{Y}})$. Then,

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathbf{Y}}}(W_{\boldsymbol{u},\boldsymbol{v}} - W_{\boldsymbol{u}',\boldsymbol{v}'})^2 \\
\le& 2(\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\boldsymbol{\Sigma}_0(\boldsymbol{u}-\boldsymbol{u}')\sum_{j=1}^n \boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{Y}}_j\tilde{\boldsymbol{Y}}_j^\mathsf{T}\boldsymbol{v} + 2\sum_{d=1}^{n-1}(\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}(\boldsymbol{\Sigma}_d+\boldsymbol{\Sigma}_d^\mathsf{T})(\boldsymbol{u}-\boldsymbol{u}')\sum_{(j-k)=d}\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{Y}}_j\cdot\boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{Y}}_k \\
&+ 2\sum_{j,k=1}^n \boldsymbol{u}'^\mathsf{T}\boldsymbol{\Sigma}_{|j-k|}\boldsymbol{u}'\cdot(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\tilde{\boldsymbol{Y}}_j\cdot(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\tilde{\boldsymbol{Y}}_k \\
\le& 2(\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\Big(\boldsymbol{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)(\boldsymbol{u}-\boldsymbol{u}')\sum_{j=1}^n \boldsymbol{v}^\mathsf{T}\tilde{\boldsymbol{Y}}_j\tilde{\boldsymbol{Y}}_j^\mathsf{T}\boldsymbol{v} + 2\Big(\|\boldsymbol{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\boldsymbol{\Sigma}_d\|\Big)\sum_{j=1}^n(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\tilde{\boldsymbol{Y}}_j\tilde{\boldsymbol{Y}}_j^\mathsf{T}(\boldsymbol{v}-\boldsymbol{v}') \\
\le& 2\|\Big(\boldsymbol{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)^{\frac{1}{2}}(\boldsymbol{u}-\boldsymbol{u}')\|^2\|\tilde{\mathbf{Y}}\|^2 + 2\Big(\|\boldsymbol{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\boldsymbol{\Sigma}_d\|\Big)\|(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\tilde{\mathbf{Y}}\|^2,
\end{aligned}$$

where the second inequality is followed by defining $\tilde{\boldsymbol{\Sigma}}_d := (\mathbf{U}_d\boldsymbol{\Lambda}_d\mathbf{U}_d^\mathsf{T} + \mathbf{V}_d\boldsymbol{\Lambda}_d\mathbf{V}_d^\mathsf{T})/2$. Here $\mathbf{U}_d, \mathbf{V}_d, \boldsymbol{\Lambda}_d$ are left singular vectors, right singular vectors and singular values of $\boldsymbol{\Sigma}_d$ for all

$1 \le d \le n-1$. Note that $\tilde{\mathbf{\Sigma}}_d$ are symmetric and positive semidefinite for all d, and hence so is $\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1} \tilde{\mathbf{\Sigma}}_d$.

Define the following Gaussian process:

$$Y_{\boldsymbol{u},\boldsymbol{v}} := \sqrt{2}\|\tilde{\mathbf{Y}}\|\boldsymbol{u}^{\mathsf{T}}\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)^{\frac{1}{2}}\boldsymbol{g} + \sqrt{2}\Big(\|\mathbf{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\mathbf{\Sigma}_d\|\Big)^{\frac{1}{2}}\boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{Y}}\boldsymbol{g}',$$

where $\boldsymbol{g}, \boldsymbol{g}'$ are independent standard Gaussian random vectors in $\mathbb{R}^p$ and $\mathbb{R}^n$ respectively. Thus by previous inequality, we have

$$\mathbb{E}_{\tilde{\mathbf{Y}}}(W_{\boldsymbol{u},\boldsymbol{v}} - W_{\boldsymbol{u}',\boldsymbol{v}'})^2 \le \mathbb{E}_{\tilde{\mathbf{Y}}}(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\boldsymbol{u}',\boldsymbol{v}'})^2.$$

Hence by Slepian-Fernique inequality (Slepian, 1962), we have

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathbf{Y}}} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} W_{\boldsymbol{u},\boldsymbol{v}} \\
\le&\mathbb{E}_{\tilde{\mathbf{Y}}} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} Y_{\boldsymbol{u},\boldsymbol{v}} \\
=&\sqrt{2}\|\tilde{\mathbf{Y}}\| \cdot \mathbb{E} \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \boldsymbol{u}^{\mathsf{T}}\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)^{\frac{1}{2}}\boldsymbol{g} + \sqrt{2}\Big(\|\mathbf{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\mathbf{\Sigma}_d\|\Big)^{\frac{1}{2}} \cdot \mathbb{E}_{\tilde{\mathbf{Y}}} \sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}} \boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{Y}}\boldsymbol{g}' \\
\le&\sqrt{2}\|\tilde{\mathbf{Y}}\| \cdot \mathbb{E}\|\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)^{\frac{1}{2}}\boldsymbol{g}\| + \sqrt{2}\Big(\|\mathbf{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\mathbf{\Sigma}_d\|\Big)^{\frac{1}{2}} \cdot \mathbb{E}_{\tilde{\mathbf{Y}}}\|\tilde{\mathbf{Y}}\boldsymbol{g}'\| \\
\le&\sqrt{2}\|\tilde{\mathbf{Y}}\| \cdot \sqrt{\operatorname{tr}\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)} + \sqrt{2}\Big(\|\mathbf{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\mathbf{\Sigma}_d\|\Big)^{\frac{1}{2}} \cdot \sqrt{\operatorname{tr}(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathsf{T}})}.
\end{aligned}$$

Taking expectation with respect to $\tilde{\mathbf{Y}}$ and using the fact that $\tilde{\mathbf{Y}}$ is an independent copy of $\mathbf{Y}$, we obtain

$$\mathbb{E} \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} W_{\boldsymbol{u},\boldsymbol{v}} \le \sqrt{2}\mathbb{E}\|\mathbf{Y}\| \cdot \sqrt{\operatorname{tr}\Big(\mathbf{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\mathbf{\Sigma}}_d\Big)} + \sqrt{2}\sqrt{\|\mathbf{\Sigma}_0\| + 2\sum_{d=1}^{n-1}\|\mathbf{\Sigma}_d\|} \cdot \sqrt{n\operatorname{tr}(\mathbf{\Sigma}_0)}.$$

This completes the proof of Lemma 14. □

Proof of Lemma 15. Define $W_{\boldsymbol{u},\boldsymbol{v}} := \boldsymbol{u}^\mathsf{T}\mathbf{Y}\boldsymbol{v}$. Then,

$$\begin{aligned}\mathbb{E}(W_{\boldsymbol{u},\boldsymbol{v}} - W_{\boldsymbol{u}',\boldsymbol{v}'})^2 =& \mathbb{E}(\boldsymbol{u}^\mathsf{T}\mathbf{Y}\boldsymbol{v} - \boldsymbol{u}'^\mathsf{T}\mathbf{Y}\boldsymbol{v}')^2 \\ \leq& 2\mathbb{E}((\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\mathbf{Y}\boldsymbol{v})^2 + 2\mathbb{E}(\boldsymbol{u}'^\mathsf{T}\mathbf{Y}(\boldsymbol{v}-\boldsymbol{v}'))^2 \\ =& 2\sum_{i,j}(\boldsymbol{u}-\boldsymbol{u}')^\mathsf{T}\boldsymbol{\Sigma}_{|i-j|}(\boldsymbol{u}-\boldsymbol{u}')\boldsymbol{v}_i\boldsymbol{v}_j + 2\sum_{i,j}\boldsymbol{u}'^\mathsf{T}\boldsymbol{\Sigma}_{|i-j|}\boldsymbol{u}'(\boldsymbol{v}_i - \boldsymbol{v}_i')(\boldsymbol{v}_j - \boldsymbol{v}_j').\end{aligned}$$

In addition, define

$$\boldsymbol{\Sigma}_L := \begin{bmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_1 & \cdots & \boldsymbol{\Sigma}_{n-1} \\ \boldsymbol{\Sigma}_1^\mathsf{T} & \boldsymbol{\Sigma}_0 & \cdots & \boldsymbol{\Sigma}_{n-2} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{\Sigma}_{n-1}^\mathsf{T} & \boldsymbol{\Sigma}_{n-2}^\mathsf{T} & \cdots & \boldsymbol{\Sigma}_0 \end{bmatrix}, \boldsymbol{\Sigma}_{L,\boldsymbol{u}} := \begin{bmatrix} \boldsymbol{u}^\mathsf{T} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{u}^\mathsf{T} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{u}^\mathsf{T} \end{bmatrix} \boldsymbol{\Sigma}_L \begin{bmatrix} \boldsymbol{u} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{u} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{u} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{\boldsymbol{u}}^\circ := (\boldsymbol{u}^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{u})\mathbf{1}_n\mathbf{1}_n^\mathsf{T}, \qquad \boldsymbol{\Sigma}^\diamond := \|\boldsymbol{\Sigma}_0\|\mathbf{1}_n\mathbf{1}_n^\mathsf{T}.$$

Since $\boldsymbol{\Sigma}_L$ is a positive semi-definite matrix, we have

$$\boldsymbol{\Sigma}_{L,\boldsymbol{u}} \preceq \boldsymbol{\Sigma}_{\boldsymbol{u}}^\circ \preceq \boldsymbol{\Sigma}^\diamond$$

for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$, where "$\preceq$" is the Loewner partial order of Hermitian matrices. Hence,

$$\mathbb{E}(W_{\boldsymbol{u},\boldsymbol{v}} - W_{\boldsymbol{u}',\boldsymbol{v}'})^2 \leq 2\|\Big(\boldsymbol{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)^{\frac{1}{2}}(\boldsymbol{u}-\boldsymbol{u}')\|^2 + 2\|\boldsymbol{\Sigma}_0\|(\boldsymbol{v}-\boldsymbol{v}')^\mathsf{T}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}(\boldsymbol{v}-\boldsymbol{v}').$$

Then define the following Gaussian process:

$$Y_{\boldsymbol{u},\boldsymbol{v}} := \sqrt{2}\boldsymbol{u}^\mathsf{T}\Big(\boldsymbol{\Sigma}_0 + 2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)^{\frac{1}{2}}\boldsymbol{g} + \sqrt{2}\|\boldsymbol{\Sigma}_0\|^{\frac{1}{2}}\boldsymbol{v}^\mathsf{T}\boldsymbol{g}',$$

where $\boldsymbol{g} \in \mathbb{R}^p, \boldsymbol{g}' \in \mathbb{R}^n$ are independent Gaussian random vectors with mean $\mathbf{0}$ and covariance matrices $\mathbf{I}_p$ and $\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ respectively. Thus by previous inequality, we have

$$\mathbb{E}(W_{\boldsymbol{u},\boldsymbol{v}} - W_{\boldsymbol{u}',\boldsymbol{v}'})^2 \leq \mathbb{E}(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\boldsymbol{u}',\boldsymbol{v}'})^2.$$

Hence by Slepian-Fernique inequality, we have

$$\begin{aligned}
&\mathbb{E}\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} W_{\boldsymbol{u},\boldsymbol{v}} \le \mathbb{E}\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathbb{S}^{p-1}} Y_{\boldsymbol{u},\boldsymbol{v}} \\
=&\sqrt{2}\mathbb{E}\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \boldsymbol{u}^\top\Big(\boldsymbol{\Sigma}_0+2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)^{\frac{1}{2}}\boldsymbol{g}+\sqrt{2}\|\boldsymbol{\Sigma}_0\|^{\frac{1}{2}}\cdot\mathbb{E}\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\boldsymbol{v}^\top\boldsymbol{g}' \\
\le&\sqrt{2}\mathbb{E}\|\Big(\boldsymbol{\Sigma}_0+2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)^{\frac{1}{2}}\boldsymbol{g}\|+\sqrt{2}\|\boldsymbol{\Sigma}_0\|^{\frac{1}{2}}\cdot\mathbb{E}\|\boldsymbol{g}'\| \\
\le&\sqrt{2}\sqrt{\operatorname{tr}\Big(\boldsymbol{\Sigma}_0+2\sum_{d=1}^{n-1}\tilde{\boldsymbol{\Sigma}}_d\Big)}+\sqrt{2}\|\boldsymbol{\Sigma}_0\|^{\frac{1}{2}}\cdot\sqrt{n}.
\end{aligned}$$

This completes the proof of Lemma 15. □

4.4.4 Proof of results in Section 4.3

Proof of Theorem 5. We first examine Assumptions **(A1)** and **(A4)**. First of all, we will study VAR(1) model, i.e., $\boldsymbol{Y}_t=\mathbf{A}\boldsymbol{Y}_{t-1}+\boldsymbol{E}_t$. Notice that for VAR(1), we could rewrite the original sequence as a moving-average model, i.e., $\boldsymbol{Y}_t=\sum_{j=0}^{\infty}\mathbf{A}^j\boldsymbol{E}_{t-j}$. For any $\boldsymbol{u}\in\mathbb{R}^p$, we have

$$\begin{aligned}
\|\boldsymbol{u}^\top\boldsymbol{Y}_t\|_{\psi_2}&=\Big\|\sum_{j=0}^{\infty}\boldsymbol{u}^\top\mathbf{A}^j\boldsymbol{E}_{t-j}\Big\|_{\psi_2} \\
&\le C\Big(\sum_{j=0}^{\infty}\|\boldsymbol{u}^\top\mathbf{A}^j\boldsymbol{E}_{t-j}\|_{\psi_2}^2\Big)^{\frac{1}{2}} \\
&\le Cc'\Big(\sum_{j=0}^{\infty}\|\boldsymbol{u}^\top\mathbf{A}^j\boldsymbol{E}_{t-j}\|_{L(2)}^2\Big)^{\frac{1}{2}}=Cc'\|\boldsymbol{u}^\top\boldsymbol{Y}_t\|_{L(2)}
\end{aligned}$$

for some universal constant $C>0$. Here the second line and last equality are followed by the fact that $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}}$ is a sequence of independent random vector, and the third line by the moment assumption on $\{\boldsymbol{E}_t\}_{t\in\mathbb{Z}}$. Since $\boldsymbol{Y}_{t-1}$ is a stable process when $\|\mathbf{A}\|<1$, $\|\boldsymbol{u}^\top\boldsymbol{Y}_t\|_{\psi_2}\le Cc'\|\boldsymbol{u}^\top\boldsymbol{Y}_t\|_{L(2)}<\infty$ for all $\boldsymbol{u}\in\mathbb{R}^p$.

Denote $\bar{\boldsymbol{Y}}_t:=(\boldsymbol{Y}_t^\top\dots\boldsymbol{Y}_{t-d}^\top)^\top$ and $\bar{\boldsymbol{E}}_t:=(\boldsymbol{E}_t^\top\ \mathbf{0}^\top\dots\mathbf{0}^\top)^\top$. For $\{\boldsymbol{Y}_t\}_{t\in\mathbb{Z}}$ generated from a VAR(d) model, $\{\bar{\boldsymbol{Y}}_t\}_{t\in\mathbb{Z}}$ is a VAR(1) process, i.e., $\bar{\boldsymbol{Y}}_t=\bar{\mathbf{A}}\cdot\bar{\boldsymbol{Y}}_{t-1}+\bar{\boldsymbol{E}}_t$. Thus by previous

argument, taking any $\boldsymbol{v} \in \mathbb{R}^{p(d+1)}$ where only the first p digits are non-zero and denoting $\boldsymbol{v}' \in \mathbb{R}^p$ to be first-p part of $\boldsymbol{v}$, we have $\|\boldsymbol{v}'^{\mathsf{T}}\boldsymbol{Y}_t\|_{\psi_2} = \|\boldsymbol{v}^{\mathsf{T}}\bar{\boldsymbol{Y}}_t\|_{\psi_2} \leq C\|\boldsymbol{v}^{\mathsf{T}}\bar{\boldsymbol{Y}}_t\|_{L(2)} = C\|\boldsymbol{v}'^{\mathsf{T}}\bar{\boldsymbol{Y}}_t\|_{L(2)} < \infty$ for some constant $C > 0$ only depending on c' where the last inequality is followed by the fact that $\{\boldsymbol{Y}_t\}$ is a stable process (see Lemma 17). Assumptions **(A1)** and **(A4)** are verified.

Then we examine Assumption **(A2)**. Without loss of generality, take $j = 0$ in Assumption **(A2)**. Let $\{\tilde{\boldsymbol{Y}}_t\}_{t=1-d}^0$ be a sequence of random vectors independent of $\{\boldsymbol{Y}_t\}_{t\leq 0}$ and identically distributed as $\{\boldsymbol{Y}_t\}_{t=1-d}^0$. Define $\tilde{\boldsymbol{Y}}_t = \mathbf{A}_1\tilde{\boldsymbol{Y}}_{t-1} + \cdots + \mathbf{A}_d\tilde{\boldsymbol{Y}}_{t-d} + \boldsymbol{E}_t$ for every $t > 0$. It is obvious that $\{\tilde{\boldsymbol{Y}}_t\}_{t>0}$ is independent of $\{\tilde{\boldsymbol{Y}}_t\}_{t\leq 0}$ and identically distributed as $\{\boldsymbol{Y}_t\}_{t>0}$. Moreover, for any $t \geq 1$, we have

$$\begin{aligned}
\|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)} &= \{\mathbb{E}\|\mathbf{A}_1\boldsymbol{Y}_{t-1} + \cdots + \mathbf{A}_d\boldsymbol{Y}_{t-d} + \boldsymbol{E}_t - (\mathbf{A}_1\tilde{\boldsymbol{Y}}_{t-1} + \cdots + \mathbf{A}_d\tilde{\boldsymbol{Y}}_{t-d} + \boldsymbol{E}_t)\|_2^{1+\epsilon}\}^{\frac{1}{1+\epsilon}} \\
&\leq \{\mathbb{E}\|\mathbf{A}_1(\boldsymbol{Y}_{t-1} - \tilde{\boldsymbol{Y}}_{t-1}) + \cdots + \mathbf{A}_d(\boldsymbol{Y}_{t-d} - \tilde{\boldsymbol{Y}}_{t-d})\|_2^{1+\epsilon}\}^{\frac{1}{1+\epsilon}} \\
&\leq \sum_{k=1}^{d} a_k\{\mathbb{E}\|\boldsymbol{Y}_{t-k} - \tilde{\boldsymbol{Y}}_{t-k}\|_2^{1+\epsilon}\}^{\frac{1}{1+\epsilon}},
\end{aligned}$$

where the third line follows by $\|\cdot\|_{L(1+\epsilon)}$ is a norm for $\epsilon > 0$. Denoting $\phi_t = \|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)}$, we have $\phi_t \leq \sum_{k=1}^d a_k\phi_{t-k}$. Let $\boldsymbol{v}$ be the unit vector with 1 at first position and 0 elsewhere. Then by iteration, we have

$$\boldsymbol{v}^{\mathsf{T}}(\phi_t, \ldots, \phi_{t-d+1})^{\mathsf{T}} \leq \boldsymbol{v}^{\mathsf{T}}\bar{\mathbf{A}}^t(\phi_0, \ldots, \phi_{1-d})^{\mathsf{T}} \leq \|\bar{\mathbf{A}}^t\|\|(\phi_0, \ldots, \phi_{1-d})^{\mathsf{T}}\|_2.$$

Note that $\phi_t = C\kappa_*$ for $t \leq 0$ by Assumption **(A1)** for some constant $C > 0$ only depending on ϵ. By the following lemma which provides sufficient and necessary conditions for matrix $\bar{\mathbf{A}}$ to have spectral radius strictly less than 1, we could choose some arbitary ρ_1 such that $\rho(\bar{\mathbf{A}}) < \rho_1 < 1$.

Lemma 17. *For $\bar{\mathbf{A}}$ defined above, $\rho(\bar{\mathbf{A}}) < 1$ if and only if $\sum_{k=1}^d a_k < 1$, where $\rho(\bar{\mathbf{A}})$ is the spectral radius of $\bar{\mathbf{A}}$.*

Proof of Lemma 17. The result is well known and here we include a proof merely for completeness. First of all, we prove the sufficient condition. A key observation is that the

characteristic equation $\det(\bar{\mathbf{A}} - \lambda \mathbf{I}_d) = 0$ for matrix $\bar{\mathbf{A}}$ is

$$f(\lambda) = \lambda^d - a_1\lambda^{d-1} - \cdots - a_{d-1}\lambda^1 - a_d = 0.$$

Assume $\sum_{j=1}^d a_j \geq 1$. We obtain $f(1) = 1 - \sum_{j=1}^d a_j \leq 0$ and $f(\infty) = \infty$. By continuity of $f(\lambda)$, there exists at least one root whose modulus is greater than or equal to 1. This contradicts with the fact that $\rho(\bar{\mathbf{A}})$ is strictly less than 1.

Secondly, we prove the necessary condition. Suppose there exists a root $z \in \mathbb{C}$ (the set of complex numbers) of $f(\lambda)$ such that $|z| \geq 1$. Here $|z|$ is the modulus of z. Then

$$|z|^d = |a_1 z^{d-1} + \cdots + a_{d-1}z^1 + a_d| \leq a_1|z|^{d-1} + \cdots + a_{d-1}|z|^1 + a_d.$$

Since $|z| \geq 1$, we have $|z|^k \leq |z|^d$ for $0 \leq k \leq d-1$. Hence $|z|^d \leq (a_1 + \cdots + a_d)|z|^d$ implies $a_1 + \cdots + a_d \geq 1$. This contradicts the fact that $\sum_{j=1}^d a_j$ is strictly less than 1. This completes the proof. □

By Gelfand's formula, there exists a $K > 0$, such that for all $t \geq K$, $\|\bar{\mathbf{A}}^t\| < \rho_1^t$. For $t < K$, we have

$$\phi_t \leq 2d\kappa_* \left(\frac{\|\bar{\mathbf{A}}\|}{\rho_1}\right)^K \rho_1^t.$$

For $t \geq K$, we have $\phi_t \leq Cd\kappa_*\rho_1^t$ for some constant $C > 0$ only depending on ϵ. Taking $\gamma_1 = Cd(\kappa_*/\kappa_1)(\|\bar{\mathbf{A}}\|/\rho_1)^K$ for some constant $C > 0$ only depending on ϵ and $\gamma_2 = \log(\rho_1^{-1})$ verifies Assumption **(A2)**.

Lastly, we verify Assumption **(A3)**. Following the same construction as in verifying Assumption **(A2)**, we have for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$,

$$\begin{aligned}
&\|(\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t)^\mathsf{T}\boldsymbol{u}\|_{L(1+\epsilon)} \\
=&(\mathbb{E}|\{\mathbf{A}_1\boldsymbol{Y}_{t-1} + \cdots + \mathbf{A}_d\boldsymbol{Y}_{t-d} + \boldsymbol{E}_t - (\mathbf{A}_1\tilde{\boldsymbol{Y}}_{t-1} + \cdots + \mathbf{A}_d\tilde{\boldsymbol{Y}}_{t-d} + \boldsymbol{E}_t)\}^\mathsf{T}\boldsymbol{u}|^{1+\epsilon})^{\frac{1}{1+\epsilon}} \\
\leq&(\mathbb{E}|\{\mathbf{A}_1\boldsymbol{Y}_{t-1} + \cdots + \mathbf{A}_d\boldsymbol{Y}_{t-d} - (\mathbf{A}_1\tilde{\boldsymbol{Y}}_{t-1} + \cdots + \mathbf{A}_d\tilde{\boldsymbol{Y}}_{t-d})\}^\mathsf{T}\boldsymbol{u}|^{1+\epsilon})^{\frac{1}{1+\epsilon}} \\
\leq&\sum_{k=1}^d a_k\{\mathbb{E}|(\boldsymbol{Y}_{t-k} - \tilde{\boldsymbol{Y}}_{t-k})^\mathsf{T}\boldsymbol{u}_k|^{1+\epsilon}\}^{\frac{1}{1+\epsilon}},
\end{aligned}$$

for $\boldsymbol{u}_k := \mathbf{A}_k\boldsymbol{u}/\|\mathbf{A}_k\boldsymbol{u}\|_2$, $k \in \{1, \ldots, d\}$. The result follows as we follow the same arguments to verify Assumption **(A2)**. This completes the proof of Theorem 5. □

Proof of Theorem 6. First of all, we verify Assumptions **(A1)** and **(A4)**. It is trivial that Assumptions **(A1)** and **(A4)** are satisfied if $W_t = 0$ almost surely for all $t \in \mathbb{Z}$. If $W_t \neq 0$ almost surely, then for all $\boldsymbol{u} \in \mathbb{R}^p$, $\|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\|_{\psi_2} \leq \|W_t\|_{L(\infty)}\|\boldsymbol{u}^\mathsf{T}\boldsymbol{E}_t\|_{\psi_2} \leq c'\kappa_W\|\boldsymbol{u}^\mathsf{T}\boldsymbol{E}_t\|_{L(2)} \leq c'\frac{\kappa_W}{\inf_{t\in\mathbb{Z}}\|W_t\|_{L(2)}}\|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\|_{L(2)} < \infty$. This verifies Assumptions **(A1)** and **(A4)**.

For Assumption **(A2)**, without loss of generality, take $j = 0$. Since $\{W_t\}_{t\in\mathbb{Z}}$ is a sequence of uniformly bounded τ-mixing random variables, we may find $\{\tilde{W}_t\}_{t>0}$ which is independent of $\{W_t\}_{t\leq 0}$, identically distributed as $\{W_t\}_{t>0}$, and for any $t \geq 1$,

$$\mathbb{E}|\tilde{W}_t - W_t| \leq \kappa_W\gamma_5 \exp\{-\gamma_6(t-1)\}.$$

Define $\tilde{\boldsymbol{Y}}_t := \tilde{W}_t\boldsymbol{E}_t$ for all $t \geq 1$. It is obvious that $\{\tilde{\boldsymbol{Y}}_t\}_{t>0}$ is independent of $\{\boldsymbol{Y}_t\}_{t\leq 0}$ and identically distributed as $\{\boldsymbol{Y}_t\}_{t>0}$. Moreover, for any integer $t \geq 1$,

$$\begin{aligned}
\|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)} &\leq (\mathbb{E}\|W_t\boldsymbol{E}_t - \tilde{W}_t\boldsymbol{E}_t\|_2^{1+\epsilon})^{\frac{1}{1+\epsilon}} \\
&\leq (\mathbb{E}|W_t - \tilde{W}_t| \cdot |W_t - \tilde{W}_t|^{1+\epsilon})^{1+\epsilon}(\mathbb{E}\|\boldsymbol{E}_t\|_2^{1+\epsilon})^{\frac{1}{1+\epsilon}} \\
&\leq C\kappa'_*\kappa_W\gamma_5^{\frac{1}{1+\epsilon}} \exp\Big\{-\frac{1}{1+\epsilon}\gamma_6(t-1)\Big\}
\end{aligned}$$

for some constant $C > 0$ only depending on ϵ. Taking $\gamma_1 = C\kappa'_*\kappa_W\gamma_5^{\frac{1}{1+\epsilon}}/\kappa_1$ and $\gamma_2 = \frac{1}{1+\epsilon}\gamma_6$ verifies Assumption **(A2)**.

For Assumption **(A3)**, without loss of generality, take $j = 0$. Let $\{\tilde{\boldsymbol{Y}}_t\}_{t>0}$ be the same construction as above. For any integer $t \geq 1$,

$$\begin{aligned}
\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|(\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t)^\mathsf{T}\boldsymbol{u}\|_{L(1+\epsilon)} &= \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \{\mathbb{E}|(W_t\boldsymbol{E}_t - \tilde{W}_t\boldsymbol{E}_t)^\mathsf{T}\boldsymbol{u}|^{1+\epsilon}\}^{\frac{1}{1+\epsilon}} \\
&= (\mathbb{E}|W_t - \tilde{W}_t|^{1+\epsilon})^{\frac{1}{1+\epsilon}} \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} (\mathbb{E}|\boldsymbol{E}_t^\mathsf{T}\boldsymbol{u}|^{1+\epsilon})^{\frac{1}{1+\epsilon}} \\
&\leq C\kappa'_1\kappa_W\gamma_5^{\frac{1}{\epsilon}} \exp\Big\{-\frac{1}{1+\epsilon}\gamma_6(t-1)\Big\}
\end{aligned}$$

for some constant $C > 0$ only depending on ϵ. Taking $\gamma_3 = C\kappa'_1\kappa_W\gamma_5^{\frac{1}{1+\epsilon}}/\kappa_1$ and $\gamma_4 = \frac{1}{1+\epsilon}\gamma_6$ verifies Assumption **(A2)**. This completes the proof of Theorem 6. □

Proof of Theorem 7. We first verify Assumptions **(A2)** and **(A3)**. Without loss of generality, take $j = 0$ in Assumption **(A2)**. Let $\tilde{\boldsymbol{Y}}_0$ be a random vector independent of $\{\boldsymbol{Y}_t\}_{t\le 0}$ and identically distributed as $\boldsymbol{Y}_0$. Define $\tilde{\boldsymbol{Y}}_t = \mathbf{A}\tilde{\boldsymbol{Y}}_{t-1} + H(\tilde{\boldsymbol{Y}}_{t-1})\boldsymbol{E}_t$ for every $t \ge 1$. It is obvious that $\{\tilde{\boldsymbol{Y}}_t\}_{t>0}$ is independent of $\{\boldsymbol{Y}_t\}_{t\le 0}$ and identically distributed as $\{\boldsymbol{Y}_t\}_{t>0}$. We obtain for any $t \ge 1$,

$$\begin{aligned}
\|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)} &= [\mathbb{E}\|\mathbf{A}\boldsymbol{Y}_{t-1} + H(\boldsymbol{Y}_{t-1})\boldsymbol{E}_t - \{\mathbf{A}\tilde{\boldsymbol{Y}}_{t-1} + H(\tilde{\boldsymbol{Y}}_{t-1})\boldsymbol{E}_t\}\|_2^{1+\epsilon}]^{\frac{1}{1+\epsilon}} \\
&\le [\mathbb{E}\|\mathbf{A}\boldsymbol{Y}_{t-1} - \mathbf{A}\tilde{\boldsymbol{Y}}_{t-1} + \{H(\boldsymbol{Y}_{t-1}) - H(\tilde{\boldsymbol{Y}}_{t-1})\}\boldsymbol{E}_t\|_2^{1+\epsilon}]^{\frac{1}{1+\epsilon}} \\
&\le (a_1 + a_2)\|\|\boldsymbol{Y}_{t-1} - \tilde{\boldsymbol{Y}}_{t-1}\|_2\|_{L(1+\epsilon)}.
\end{aligned}$$

By iteration, we obtain

$$\|\|\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t\|_2\|_{L(1+\epsilon)} \le (a_1 + a_2)^t(\mathbb{E}\|\boldsymbol{Y}_0 - \tilde{\boldsymbol{Y}}_0\|_2^{1+\epsilon})^{\frac{1}{1+\epsilon}} \le C\kappa_*(a_1 + a_2)^t$$

for some constant $C > 0$ only depending on ϵ. Taking $\gamma_1 = C\kappa_*/\kappa_1$ and $\gamma_2 = -\log(a_1 + a_2)$ verifies Assumption **(A2)**.

For Assumption **(A3)**, following the construction above, we have for any $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and $t \ge 1$,

$$\begin{aligned}
\|(\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t)^\mathsf{T}\boldsymbol{u}\|_{L(1+\epsilon)} &= [\mathbb{E}|\{\mathbf{A}\boldsymbol{Y}_{t-1} + H(\boldsymbol{Y}_{t-1})\boldsymbol{E}_t - (\mathbf{A}\tilde{\boldsymbol{Y}}_{t-1} + H(\tilde{\boldsymbol{Y}}_{t-1})\boldsymbol{E}_t)\}^\mathsf{T}\boldsymbol{u}|^{1+\epsilon}]^{\frac{1}{1+\epsilon}} \\
&\le [\mathbb{E}|\{\mathbf{A}\boldsymbol{Y}_{t-1} - \mathbf{A}\tilde{\boldsymbol{Y}}_{t-1} + (H(\boldsymbol{Y}_{t-1}) - H(\boldsymbol{Y}_{t-1}))\boldsymbol{E}_t\}^\mathsf{T}\boldsymbol{u}|^{1+\epsilon}]^{\frac{1}{1+\epsilon}} \\
&\le a_1\|(\boldsymbol{Y}_{t-1} - \tilde{\boldsymbol{Y}}_{t-1})^\mathsf{T}\boldsymbol{v}\|_{L(1+\epsilon)} + a_2\frac{\kappa_1'}{\kappa_*'}\|\|\boldsymbol{Y}_{t-1} - \tilde{\boldsymbol{Y}}_{t-1}\|_2\|_{L(1+\epsilon)},
\end{aligned}$$

where $\boldsymbol{v} := \mathbf{A}\boldsymbol{u}/\|\mathbf{A}\boldsymbol{u}\|_2 \in \mathbb{S}^{p-1}$. By iteration, we obtain

$$\|(\boldsymbol{Y}_t - \tilde{\boldsymbol{Y}}_t)^\mathsf{T}\boldsymbol{u}\|_{L(1+\epsilon)} \le C\{\kappa_1 a_1^t + 2\kappa_*\frac{\kappa_1'}{\kappa_*'}a_2\sum_{\ell=0}^{t-1}a_1^\ell(a_1 + a_2)^{t-1-\ell}\} \le C(a_1 + a_2)^t\max(\kappa_*\frac{\kappa_1'}{\kappa_*'}, \kappa_1)$$

for some constant $C > 0$ only depending on ϵ. Taking $\gamma_3 = C\max(\frac{\kappa_*\kappa_1'}{\kappa_1\kappa_*'}, 1)$ and $\gamma_4 = -\log(a_1 + a_2)$ verifies Assumption **(A3)**.

By further assuming that $\{\boldsymbol{Y}_t\}$ is a stationary process and $H(\cdot)$ is uniformly bounded, we have that for all $t \in \mathbb{Z}$, $\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_t\|_{\psi_2} \le \|\mathbf{A}\|\sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}}\|\boldsymbol{u}^\mathsf{T}\boldsymbol{Y}_{t-1}\|_{\psi_2} + D_2\sup_{\boldsymbol{v}\in\mathbb{S}^{p-1}}\|\boldsymbol{v}^\mathsf{T}\boldsymbol{E}_t\|$.

By stationarity, this renders $\kappa_1 = \sup_{\boldsymbol{u}\in\mathbb{S}^{p-1}} \|\boldsymbol{u}^\top \boldsymbol{Y}_t\|_{\psi_2} \leq \frac{1}{1-\|\mathbf{A}\|} D_2 \kappa_1' < \infty$. Similar argument applies to κ_*. This verifies Assumption **(A1)** under additional assumptions and completes the proof of Theorem 7. □

Chapter 5

CONCLUSION

This chapter contains the main contributions of this work together with possible directions for future research.

5.1 Contributions to Research

In this work, I have developed several methods and theories in high dimensional time series estimations and forecasts. First of all, I proposed a new Bayesian hierarchical model to forecast long term all-age smoking attributable fraction (ASAF). To the best of the author's knowledge, this is the first BHM to forecast ASAF for multi-population simultaneously. Out-of-sample validation on the existing data shows that the proposed method is accurate and calibrated. The potential uses of the projected ASAF include monitoring future smoking epidemics and its impact on human mortality measures, assisting the forecasts of various aspects of human mortality measures, and providing a baseline forecast to assess the effectiveness of additional smoking-related policies.

As an application of the previously mentioned BHM for ASAF, I proposed a joint model for forecasting male-female life expectancy at birth to demonstrate the usefulness of ASAF in determining the future between-gender gap. In addition, I proposed a new modeling framework for the male life expectancy forecast by incorporating the smoking epidemic. This framework considers the ideas proposed in Bongaarts (2006) and Janssen et al. (2013) as the general guidance while it is built on two newly proposed BHMs for the age-specific smoking attributable fraction (ASSAF) and the non-smoking life expectancy at birth. These two models differ fundamentally from the ones used in the mentioned literature and are shown to produce more accurate and calibrated forecasting results than commonly used methods.

The new framework provides further evidence for the adverse impacts of smoking on human mortality measures and sheds light on modeling other health-risking lifestyles into mortality forecasts.

The other component of this work is the derivation of the optimal moment bounds for autocovariance matrices estimation under a general class of high-dimensional time series models. This new result makes a step to extend current results based on independent samples to a more general dependent data structure. The major contribution of the work is to derive a Bernstein-type inequality for a sequence of dependent random matrices, which further extends along the line of Merlevède et al. (2009), Tropp (2015), and Banna et al. (2016).

5.2 Future Research

5.2.1 Smoking- and mortality-related Forecast

One major drawback of our ASAF and life expectancy forecasts is that we can not apply this method for all countries around the world. One possible direction is to gather useful smoking-related data for more countries especially in particular developing countries and those still experiencing a maturation of the smoking epidemic such as China, India, Indonesia, and most African countries. Since the estimation of ASAF and ASSAF, using either the Peto-Lopez method (Peto et al., 1992) or PGW method (Preston et al., 2009), depends on high quality lung cancer mortality data, further collections of these data for countries of interest should be conducted outside the WHO Mortality Database. Another possible direction is to impute the missing data, which would require further investigation of country-specific covariates.

In addition, the female SAF forecast can be further refined. As more and more data becomes available for the female population, the female SAF pattern is expected to resemble that of the male one with more easy-to-estimate time-to-peak and rate of decline, which would potentially yield more accurate and sharp forecasts compared with current results. Moreover, the question of whether SAF will reach 0 in the future could be further examined. Although the use of traditional tobacco and manufactured cigarettes is decreasing continuously among

male adults in developed countries, the use of new devices to smoke such as electronic cigarettes is increasing especially among adolescents. Recent, an Illinois resident recently died of severe respiratory illness, possibly due to the use of vaping-e-cigarette. If such trend continues, a short-term resurrection of the smoking epidemic would not be impossible. However, the long-term influence of smoking is unlikely to resummit back to a historical level.

Thirdly, the out-of-sample validation of our proposed method for ASAF works well for populations with strong and clear patterns while it works less satisfactorily for non-clear-pattern data. One possible solution would be collecting more accurate data. Another one would be using country-specific covariates to assist the forecast. On the other hand, since many such countries also suffer from other risk factors of respiratory diseases such as air pollution, using the Peto-Lopez method directly might not be the most accurate, and adopting variants such as in Ezzati and Lopez (2003) would be more proper.

For projecting life expectancy at birth, we encounter similar issues as discussed above, such as limited high quality data and a large prediction variance for female forecasts. Besides smoking, other lifestyle-type health risk factors like obesity and alcohol consumption have also been investigated in literature (Trias Llimós and Janssen, 2019; Vidra et al., 2017). It is of interest to see whether those risk factors could fit in our framework to produce a probabilistic forecast.

5.2.2 Autocovariance Estimation Theory

One direct extension of our results is to remove the extra logarithm factors in the moment bound in Theorem 3. As proven in Koltchinskii and Lounici (2017a) and Theorem 4 in Chapter 4, such improvement is possible for independent subgaussian samples and samples from a stationary Gaussian process. Another direction is to extend the results under less restrictive moment assumption. Bai and Yin (1993) showed that $\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|$ converges to 0 in probability as long as the 4-th moment exists. Mendelson and Paouris (2014) proved an optimal non-asymptotic tail bound for covariance estimation under $4 + \delta$ moment as-

sumption. For the potential use of our theorem, one could apply it to show the consistency of principal component analysis (PCA) under high-dimensional time series models and to provide theoretical guarantee for the spectral clustering of mixture models with dependent data.

BIBLIOGRAPHY

Adamczak, R. (2007). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13(34):1000–1034.

Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography*, 48(3):815–839.

Andrews, D. W. (1984). Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4):930–934.

Au, J. S., Mang, O. W., Foo, W., and Law, S. C. (2004). Time trends of lung cancer incidence by histologic types and smoking prevalence in Hong Kong 1983–2000. *Lung Cancer*, 45(2):143–152.

Azizyan, M., Krishnamurthy, A., and Singh, A. (2014). Subspace learning from extremely compressed measurements. In *Asilomar Conference on Signals, Systems and Computers*, pages 311–315. IEEE.

Azose, J. J. and Raftery, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography*, 52(5):1627–1650.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294.

Banna, M., Merlevède, F., and Youssef, P. (2016). Bernstein-type inequality for a class of dependent random matrices. *Random Matrices: Theory and Applications*, 5(2):1650006.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Basu, S., Michailidis, G., et al. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Berbee, H. C. (1979). *Random Walks with Stationary Increments and Renewal Theory*, volume 112. Mathematisch Centrum.

Berkes, I. and Philipp, W. (1977). An almost sure invariance principle for the empirical distribution function of mixing random variables. *Probability Theory and Related Fields*, 41(2):115–137.

Berkes, I. and Philipp, W. (1979). Approximation theorems for independent and weakly dependent random vectors. *The Annals of Probability*, 7(1):29–54.

Bhatia, R. (1997). *Matrix Analysis*, volume 169. Springer.

Blinn, J. (1996). Consider the lowly 2×2 matrix. *IEEE Computer Graphics and Applications*, 16(2):82–88.

Bongaarts, J. (2006). How long will we live? *Population and Development Review*, 32(4):605–628.

Bongaarts, J. (2014). Trends in causes of death in low-mortality countries: implications for mortality projections. *Population and Development Review*, 40(2):189–212.

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*, 15:289–310.

Booth, H., Maindonald, J., and Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3):325–336.

Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, 3(1-2):3–43.

Bordenave, C., Caputo, P., Chafaï, D., and Tikhomirov, K. (2018). On the spectral radius of a random matrix. *The Annals of Probability*, (in press).

Bradley, R. C. (1985). Basic properties of strong mixing conditions. Technical report, NORTH CAROLINA UNIV AT CHAPEL HILL CENTER FOR STOCHASTIC PROCESSES.

Bradley, R. C. (2005a). Basic properties of strong mixing conditions. *Probability Surveys*, 2(2):107–144.

Bradley, R. C. (2005b). *Introduction to Strong Mixing Conditions*, volume 2. Kendrick Press.

Bradley, R. C. (2005c). *Introduction to Strong Mixing Conditions*, volume 1. Kendrick Press.

Brand, M. (2006). Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra and its Applications*, 415(1):20–30.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. Siam.

Britton, J. (2017). Death, disease, and tobacco. *The Lancet*, 389(10082):1861–1862.

Brualdi, R. A. and Schneider, H. (1983). Determinantal identities: Gauss, schur, cauchy, sylvester, kronecker, jacobi, binet, laplace, muir, and cayley. *Linear Algebra and its applications*, 52:769–791.

Bryc, W. (1981). *Sequences of Weakly Dependent Random Variables*. PhD thesis, Politechnika Warszawska.

Bryc, W. (1992). On large deviations for uniformly strong mixing sequences. *Stochastic Processes and Their Applications*, 41(2):191–202.

Bryc, W. and Dembo, A. (1996). Large deviations and strong mixing. *Annales de l'institut Henri Poincare*, 32(4):549–569.

Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli*, 21(2):1200–1230.

Burns, D. M., Lee, L., Shen, L. Z., Gilpin, E., Tolley, H. D., Vaughn, J., Shanks, T. G., et al. (1997). Cigarette smoking behavior in the United States. *Changes in cigarette-related disease risks and their implication for prevention and control. Smoking and Tobacco Control Monograph*, 8:13–42.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Centers for Disease Control and Prevention (2019). What are the risk factors for lung cancer? Last accessed: Oct. 19, 2019 https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm.

Chang, J., Guo, B., and Yao, Q. (2018). Principal component analysis for second-order stationary vector time series. *The Annals of Statistics*, 46(5):2094–2124.

Chatterjee, S. (2005). *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021.

Chen, Z., Peto, R., Zhou, M., Iona, A., Smith, M., Yang, L., Guo, Y., Chen, Y., Bian, Z., Lancaster, G., et al. (2015). Contrasting male and female trends in tobacco-attributed

mortality in China: evidence from successive nationwide prospective cohort studies. *The Lancet*, 386(10002):1447–1456.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical science*, pages 204–218.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

De Jong, P. and Tickle, L. (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18.

Dedecker, J., Doukhan, P., Lang, G., Leon, J., Louhichi, S., and Prieur, C. (2007). *Weak Dependence: With Examples and Applications*. Springer-Verlag New York.

Dedecker, J. and Prieur, C. (2004). Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability*, 17(4):861–885.

Dedecker, J. and Prieur, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236.

Ding, X., Qiu, Z., Chen, X., et al. (2017). Sparse transition matrix estimation for high-dimensional and locally stationary vector autoregressive models. *Electronic Journal of Statistics*, 11(2):3871–3902.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Ezzati, M. and Lopez, A. D. (2003). Estimates of global mortality attributable to smoking in 2000. *The Lancet*, 362(9387):847–852.

Ezzati, M. and Lopez, A. D. (2004). Regional, disease specific patterns of smoking-attributable mortality in 2000. *Tobacco Control*, 13(4):388–395.

Fan, J., Han, F., and Liu, H. (2014). Page: Robust pattern guided estimation of large covariance matrix. Technical report, Technical report, Technical report, Princeton University.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fenelon, A. and Preston, S. H. (2012). Estimating smoking-attributable mortality in the United States. *Demography*, 49(3):797–818.

Fokas, N. (2007). Growth functions, social diffusion, and social change. *Review of Sociology*, 13(1):5–30.

Forey, B., Hamling, J., Hamling, J., and Lee, P. (2006). International smoking statistics. a collection of worldwide historical data. *Web edition. Sutton, Surrey: PN Lee Statistics and Computing Ltd*, 2016.

Freedman, D. A. (1975). On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118.

Fung, M. C., Peters, G. W., and Shevchenko, P. V. (2017). A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting. *Annals of Actuarial Science*, 11(2):343–389.

Gajalakshmi, V., Peto, R., Kanaka, T. S., and Jha, P. (2003). Smoking and mortality from tuberculosis and other diseases in India: retrospective study of 43000 adult male deaths and 35000 controls. *The Lancet*, 362(9383):507–515.

Garg, A. and Srivastava, N. (2017). Matrix concentration for expander walks. *arXiv preprint arXiv:1704.03864*.

Gelman, A. and Rubin, D. B. (1992a). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Gelman, A. and Rubin, D. B. (1992b). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Gershgorin, S. A. (1931). Uber die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Navk SSSR, Ser. Mat.*, 7:749–754.

Giordano, R. (2019). *rstansensitivity: Tools for calculating hyperparameter sensitivity in Stan.* R package version 0.0.0.9000.

Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Godwin, J. and Raftery, A. E. (2017). Bayesian projection of life expectancy accounting for the HIV/AIDS epidemic. *Demographic Research*, 37:1549.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Grübler, A., Nakićenović, N., and Victor, D. G. (1999). Dynamics of energy technologies and global change. *Energy Policy*, 27(5):247–280.

Gu, D., Kelly, T. N., Wu, X., Chen, J., Samet, J. M., Huang, J.-f., Zhu, M., Chen, J.-c., Chen, C.-S., Duan, X., et al. (2009). Mortality attributable to smoking in China. *New England Journal of Medicine*, 360(2):150–159.

Guo, S., Wang, Y., and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, page asw046.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Han, F. (2017). An exponential inequality for U-statistics under mixing conditions. *Journal of Theoretical Probability*, (to appear).

Han, F. and Liu, H. (2018). ECA: High-dimensional elliptical component analysis in non-gaussian distributionsigh-dimensional elliptical component analysis in non-gaussian distributions. *Journal of the American Statistical Association*, 113(521):252–268.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150.

Han, F., Xu, S., and Liu, H. (2017). Rate-optimal estimation of a high-dimensional semi-parametric time series model.

Han, Y. and Tsay, R. S. (2017). High-dimensional linear regression for dependent observations with application to nowcasting. *arXiv preprint arXiv:1706.07899*.

Hanson, G., Venturelli, P., and Fleckenstein, A. (2011). *Drugs and society*. Jones & Bartlett Publishers.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Head, G. A., Lukoshkova, E. V., Mayorov, D. N., and van den Buuse, M. (2004). Non-symmetrical double-logistic analysis of 24-h blood pressure recordings in normotensive and hypertensive rats. *Journal of Hypertension*, 22(11):2075–2085.

Head, G. A. and McCarty, R. (1987). Vagal and sympathetic components of the heart rate range and gain of the baroreceptor-heart rate reflex in conscious rats. *Journal of the Autonomic Nervous System*, 21(2-3):203–213.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Houdré, C., Mason, D. M., Reynaud-Bouret, P., and Rosiński, J. (2016). *High Dimensional Probability VII: The Cargèse Volume*, volume 71. Birkhäuser.

Hyndman, R. J., Booth, H., Tickle, L., and Maindonald., J. (2019). *Demography: Forecasting Mortality, Fertility, Migration and Population Data*. R package version 1.22. https://CRAN.R-project.org/package=demography.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, 7(4):349–382.

Islami, F., Torre, L. A., and Jemal, A. (2015). Global trends of lung cancer mortality and smoking prevalence. *Translational Lung Cancer Research*, 4(4):327.

Janssen, F. (2018). Advances in mortality forecasting: introduction. *Genus*, 74(1):21.

Janssen, F. and van Poppel, F. (2015). The adoption of smoking and its effect on the mortality gender gap in Netherlands: a historical perspective. *BioMed Research International*, 2015.

Janssen, F., van Wissen, L. J., and Kunst, A. E. (2013). Including the smoking epidemic in internationally coherent mortality projections. *Demography*, 50(4):1341–1362.

King, G. and Soneji, S. (2011). The future of death in America. *Demographic Research*, 25:1–38.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Kolmogorov, A. N. and Rozanov, Y. A. (1960). On strong mixing conditions for stationary Gaussian processes. *Theory of Probability and Its Applications*, 5(2):204–208.

Koltchinskii, V. and Lounici, K. (2017a). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133.

Koltchinskii, V. and Lounici, K. (2017b). New asymptotic results in principal component analysis. *Sankhya A*, 79(2):254–297.

Koltchinskii, V. and Lounici, K. (2017c). Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157.

Koltchinskii, V., Lounici, K., Tsybakov, A. B., et al. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Kong, K. A., Jung-Choi, K.-H., Lim, D., Lee, H. A., Lee, W. K., Baik, S. J., Park, S. H., and Park, H. (2016). Comparison of prevalence-and smoking impact ratio-based methods of estimating smoking-attributable fractions of deaths. *Journal of Epidemiology*, 26(3):145–154.

Krebs, J. T. (2017). A Bernstein inequality for exponentially growing graphs. *arXiv:1701.04188*.

Kucharavy, D. and De Guio, R. (2011). Logistic substitution model and technological forecasting. *Procedia Engineering*, 9:402–416.

Lal, P. G., Wilson, N. C., and Gupta, P. C. (2012). Attributable deaths from smoking in the last 100 years in India. *Current Science*, pages 1085–1090.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Lam, C., Yao, Q., et al. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, T., Ho, S., Hedley, A., Mak, K., and Peto, R. (2001). Mortality and smoking in Hong Kong: case-control study of all adult deaths in 1998. *BMJ*, 323(7309):361.

Lariscy, J. T., Hummer, R. A., and Rogers, R. G. (2018). Cigarette smoking and all-cause and cause-specific adult mortality in the United States. *Demography*, 55(5):1855–1885.

Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4):537–549.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio Int Contra Cancrum*, 9:531–941.

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

Li, Y. and Raftery, A. E. (2019). Estimating and forecasting the smoking-attributable mortality fraction for both sexes jointly in 69 countries. *arXiv preprint arXiv:1902.07791*.

Liu, B.-Q., Peto, R., Chen, Z.-M., Boreham, J., Wu, Y.-P., Li, J.-Y., Campbell, T. C., and Chen, J.-S. (1998). Emerging tobacco hazards in China: 1. retrospective proportional mortality study of one million deaths. *BMJ*, 317(7170):1411–1422.

Liu, W., Xiao, H., and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statistica Sinica*, 23(3):1257–1272.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics*, 40(3):2726–2734.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Luo, L. (2013). Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem (with discussion). *Demography*, 50:1945–1988.

Luo, Q., Yu, X. Q., Wade, S., Caruana, M., Pesola, F., Canfell, K., and O'Connell, D. L. (2018). Lung cancer mortality in Australia: Projected outcomes to 2040. *Lung Cancer*, 125:68–76.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Ma, J., Siegel, R. L., Jacobs, E. J., and Jemal, A. (2018). Smoking-attributable mortality by state in 2014, US. *American Journal of Preventive Medicine*, 54(5):661–670.

Mackenbach, J. P., Huisman, M., Andersen, O., Bopp, M., Borgan, J.-K., Borrell, C., Costa, G., Deboosere, P., Donkin, A., Gadeyne, S., et al. (2004). Inequalities in lung cancer mortality by the educational level in 10 european populations. *European Journal of Cancer*, 40(1):126–135.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. (2014). Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945.

Marchetti, C., Meyer, P. S., and Ausubel, J. H. (1996). Human population dynamics revisited with the logistic model: how much can be modeled and predicted? *Technological Forecasting and Social Change*, 52(1):1–30.

Medeiros, M. C. and Mendes, E. F. (2016). ℓ_1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271.

Mehta, N. and Preston, S. H. (2012). Continued increases in the relative risk of death from smoking. *American Journal of Public Health*, 102(11):2181–2186.

Melnyk, I. and Banerjee, A. (2016). Estimating structured vector autoregressive models. In *International Conference on Machine Learning*, pages 830–839.

Mendelson, S. (2010). Empirical processes with a bounded ψ_1 diameter. *Geometric and Functional Analysis*, 20(4):988–1027.

Mendelson, S. and Paouris, G. (2014). On the singular values of random matrices. *Journal of the European Mathematical Society*, 16:823–834.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: the Luminy Volume*, pages 273–292. Institute of Mathematical Statistics.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3):435–474.

Meyer, P. (1994). Bi-logistic growth. *Technological Forecasting and Social Change*, 47(1):89–102.

Meyer, P. S., Yung, J. W., and Ausubel, J. H. (1999). A primer on logistic growth and

substitution: the mathematics of the Loglet Lab software. *Technological Forecasting and Social Change*, 61(3):247–271.

Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4):303–306.

Mishra, S., Joseph, R. A., Gupta, P. C., Pezzack, B., Ram, F., Sinha, D. N., Dikshit, R., Patra, J., and Jha, P. (2016). Trends in bidi and cigarette smoking in India from 1998 to 2015, by age, gender and education. *BMJ Global Health*, 1(1):e000005.

Mons, U. and Brenner, H. (2017). Demographic ageing and the evolution of smoking-attributable mortality: the example of Germany. *Tobacco Control*, 26(4):455–457.

Muszyńska, M. M., Fihel, A., and Janssen, F. (2014). Role of smoking in regional variation in mortality in p oland. *Addiction*, 109(11):1931–1941.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

Ng, M., Freeman, M. K., Fleming, T. D., Robinson, M., Dwyer-Lindgren, L., Thomson, B., Wollum, A., Sanman, E., Wulf, S., Lopez, A. D., et al. (2014). Smoking prevalence and cigarette consumption in 187 countries, 1980-2012. *Journal of the American Medical*, 311(2):183–192. Dataset last assess: Oct. 15, 2018 http://www.healthdata.org/data-tools.

Niu, S.-R., Yang, G.-H., Chen, Z.-M., Wang, J.-L., Wang, G.-H., He, X.-Z., Schoepff, H., Boreham, J., Pan, H.-C., and Peto, R. (1998). Emerging tobacco hazards in China: 2. early mortality results from a prospective study. *BMJ*, 317(7170):1423–1424.

Oeppen, J. and Vaupel, J. W. (2002). Enhanced: broken limits to life expectancy. *Science*, 296(5570):1029–1031.

Oliveira, R. (2010). Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15:203–212.

Oliveira, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv:0911.0600*.

Pampel, F. (2005). Forecasting sex differences in mortality in high income nations: The contribution of smoking. *Demographic Research*, 13(18):455.

Pampel, F. C. (2006). Global patterns and determinants of sex differences in smoking. *International Journal of Comparative Sociology*, 47(6):466–487.

Parascandola, M. and Xiao, L. (2019). Tobacco and the lung cancer epidemic in China. *Translational Lung Cancer Research*, 8(Suppl 1):S21.

Parkin, D. M., Boyd, L., and Walker, L. (2011). The fraction of cancer attributable to lifestyle and environmental factors in the uk in 2010. *British Journal of Cancer*, 105(S2):S77.

Paulin, D., Mackey, L., and Tropp, J. A. (2016). Efron–Stein inequalities for random matrices. *The Annals of Probability*, 44(5):3431–3473.

Pedroza, C. (2006). A bayesian forecasting model: predicting US male mortality. *Biostatistics*, 7(4):530–550.

Pérez-Ríos, M. and Montes, A. (2008). Methodologies used to estimate tobacco-attributable mortality: a review. *BMC public health*, 8(1):22.

Peters, F., Mackenbach, J., and Nusselder, W. (2016). Do life expectancy projections need to account for the impact of smoking. *Netspar Design Papers*, 2016(52):1–54.

Peto, R., Boreham, J., Lopez, A. D., Thun, M., and Heath, C. (1992). Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *The Lancet*, 339(8804):1268–1278.

Peto, R., Lopez, A. D., Boreham, J., and Thun, M. (2006). Mortality from smoking in developed countries. *Population*, 673290(284395):300245.

Peto, R., Lopez, A. D., Boreham, J., Thun, M., and Heath, C. (1994). Mortality from smoking in developed countries 1950-2000. Indirect estimates from national statistics.

Petz, D. (1994). A survey of certain trace inequalities. *Banach Center Publications*, 30(1):287–298.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.

Preston, S., Heuveline, P., and Guillot, M. (2000a). *Demography: Measuring and Modeling Population Processes*. Wiley-Blackwell.

Preston, S., Heuveline, P., and Guillot, M. (2000b). *Demography: measuring and modeling population processes*. Wiley-Blackwell.

Preston, S. H., Glei, D. A., and Wilmoth, J. R. (2009). A new method for estimating smoking-attributable mortality in high-income countries. *International Journal of Epidemiology*, 39(2):430–438.

Preston, S. H., Glei, D. A., and Wilmoth, J. R. (2011). Contribution of smoking to international differences in life expectancy. In Crimmins, E. M., Preston, S. H., and Cohen, B., editors, *International Differences in Mortality at Older Ages: Dimensions and Sources*, pages 105–31. The National Academies Press, Washington, DC.

Preston, S. H., Stokes, A., Mehta, N. K., and Cao, B. (2014). Projecting the effect of changes

in smoking and obesity on future life expectancy in the United States. *Demography*, 51(1):27–49.

Preston, S. H. and Wang, H. (2006). Sex mortality differences in the United States: The role of cohort smoking patterns. *Demography*, 43(4):631–646.

Raftery, A. E., Alkema, L., and Gerland, P. (2014a). Bayesian population projections for the United Nations. *Statistical Science*, 29:58–68.

Raftery, A. E. and Bao, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics*, 66(4):1162–1173.

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., Lalic, N., and Gerland, P. (2014b). Joint probabilistic projection of female and male life expectancy. *Demographic Research*, 30:795–822.

Raftery, A. E. and Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7(4):493–497.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., and Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109:13915–13921.

Rani, M., Bonu, S., Jha, P., Nguyen, S., and Jamjoum, L. (2003). Tobacco use in India: prevalence and predictors of smoking and chewing in a national cross sectional household survey. *Tobacco Control*, 12(4):e4–e4.

Reitsma, M. B., Fullman, N., Ng, M., Salama, J. S., Abajobir, A., Abate, K. H., Abbafati, C., Abera, S. F., Abraham, B., Abyu, G. Y., et al. (2017). Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the global burden of disease study 2015. *The Lancet*, 389(10082):1885–1906.

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.

Rogers, R. G., Hummer, R. A., Krueger, P. M., and Pampel, F. C. (2005). Mortality attributable to cigarette smoking in the United States. *Population and Development Review*, 31(2):259–292.

Rosen, L. (2013). An intuitive approach to understanding the attributable fraction of disease due to a risk factor: the case of smoking. *International Journal of Environmental Research and Public Health*, 10(7):2932–2943.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1):43.

Roser, M. and Ritchie, H. (2019). Smoking. *Our World in Data*. https://ourworldindata.org/smoking.

Rostron, B. L. and Wilmoth, J. R. (2011). Estimating the effect of smoking on slowdowns in mortality declines in developed countries. *Demography*, 48(2):461–479.

Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72.

Ševčíková, H., Li, N., and Gerland, P. (2019a). *MortCast: Estimation and Projection of Age-Specific Mortality Rates*. R package version 2.1-1. https://CRAN.R-project.org/package=MortCast.

Ševčíková, H., Li, N., Kantorová, V., Gerland, P., and Raftery, A. E. (2016). Age-specific mortality and fertility rates for probabilistic population projections. In *Dynamic Demographic Analysis*, pages 285–310. Springer.

Ševčíková, H., Raftery, A., and Chunn, J. (2019b). *bayesLife: Bayesian Projection of Life Expectancy*. R package version 4.0-2. https://CRAN.R-project.org/package=bayesLife.

Shabani, A., Sepaskhah, A., Kamgar-Haghighi, A., and Honar, T. (2018). Using double logistic equation to describe the growth of winter rapeseed. *The Journal of Agricultural Science*, 156(1):37–45.

Shang, H. L. (2016). Mortality and life expectancy forecasting for a group of populations in developed countries: a multilevel functional data method. *The Annals of Applied Statistics*, 10(3):1639–1672.

Shibuya, K., Inoue, M., and Lopez, A. D. (2005). Statistical modeling and projections of lung cancer mortality in 4 industrialized countries. *International Journal of Cancer*, 117(3):476–485.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer.

Simonato, L., Agudo, A., Ahrens, W., Benhamou, E., Benhamou, S., Boffetta, P., Brennan, P., Darby, S. C., Forastiere, F., Fortes, C., et al. (2001). Lung cancer and cigarette smoking in Europe: an update of risk estimates and an assessment of inter-country heterogeneity. *International Journal of Cancer*, 91(6):876–887.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501.

Smith, T. R. and Wakefield, J. (2016). A review and comparison of age–period–cohort models for cancer incidence. *Statistical Science*, 31(4):591–610.

Srivastava, N. and Vershynin, R. (2013). Covariance estimation for distributions with $2 + \epsilon$ moments. *The Annals of Probability*, 41(5):3081–3111.

Stan Development Team (2018). *RStan: the R interface to Stan*. R package version 2.18.2. http://mc-stan.org/.

Stoeldraijer, L., Bonneux, L., van Duin, C., van Wissen, L., and Janssen, F. (2015). The

future of smoking-attributable mortality: the case of England & Wales, Denmark and the Netherlands. *Addiction*, 110(2):336–345.

Stoeldraijer, L., van Duin, C., van Wissen, L., and Janssen, F. (2013). Impact of different mortality forecasting methods and explicit assumptions on projected future life expectancy: The case of the Netherlands. *Demographic Research*, 29:323–354.

Tachfouti, N., Raherison, C., Obtel, M., and Nejjari, C. (2014). Mortality attributable to tobacco: review of different methods. *Archives of Public Health*, 72(1):22.

Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer.

Teng, A., Atkinson, J., Disney, G., Wilson, N., and Blakely, T. (2017). Changing smoking-mortality association over time and across social groups: National census-mortality cohort studies from 1981 to 2011. *Scientific Reports*, 7(1):11465.

Thun, M., Peto, R., Boreham, J., and Lopez, A. D. (2012). Stages of the cigarette epidemic on entering its second century. *Tobacco control*, 21(2):96–101.

Thun, M. J., Apicella, L. F., and Henley, S. J. (2000). Smoking vs other risk factors as the cause of smoking-attributable deaths: confounding in the courtroom. *Jama*, 284(6):706–712.

Tikhomirov, K. (2017). Sample covariance matrices of heavy-tailed distributions. *International Mathematics Research Notices*, 2018(20):6254–6289.

Trias Llimós, S. and Janssen, F. (2019). Gender gaps in life expectancy and alcohol consumption in Eastern Europe. *N-IUSSP*.

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats.

United Nations (2015). *World Population Prospects 2015*. United Nations, New York, N.Y.

United Nations (2017). *World Population Prospects*. United Nations, New York, N.Y. Accessed: Oct. 15, 2018 http://population.un.org/wpp/Download/Standard/Population/.

van Handel, R. (2017). Structured random matrices. In *Convexity and Concentration*, volume 161, pages 107–156. Springer.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press.

Vidra, N., Trias-Llimós, S., and Jansse, F. (2017). Impact of obesity on trends in life expectancy among different European countries, 1975-2012. *European Journal of Public Health*, 27(suppl_3).

Volkonskii, V. and Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theory of Probability & Its Applications*, 4(2):178–197.

Wang, H. and Preston, S. H. (2009). Forecasting United States mortality using cohort smoking histories. *Proceedings of the National Academy of Sciences*, 106(2):393–398.

Wang, J.-B., Jiang, Y., Wei, W.-Q., Yang, G.-H., Qiao, Y.-L., and Boffetta, P. (2010). Estimation of cancer incidence and mortality attributable to smoking in China. *Cancer Causes & Control*, 21(6):959–965.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45.

Wen, C., Tsai, S., Chen, C., Cheng, T., Tsai, M., and Levy, D. (2005). Smoking attributable mortality for Taiwan and its projection to 2020 under different smoking scenarios. *Tobacco Control*, 14(suppl 1):i76–i80.

Whelpton, P. K. (1936). An empirical method of calculating future population. *Journal of the American Statistical Association*, 31(195):457–473.

Wilms, I., Basu, S., Bien, J., and Matteson, D. S. (2017). Sparse identification and estimation of high-dimensional vector autoregressive moving averages. *arXiv preprint arXiv:1707.09208*.

Wiśniowski, A., Smith, P. W., Bijak, J., Raymer, J., and Forster, J. J. (2015). Bayesian population forecasting: extending the Lee-Carter method. *Demography*, 52(3):1035–1059.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

World Health Organization (2017). Mortality database. Last accessed: Oct. 15, 2018 http://www.who.int/healthinfo/statistics/mortality_rawdata/en/.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14150–14154.

Wu, W. B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379.

Yang, X., Mustard, J. F., Tang, J., and Xu, H. (2012). Regional-scale phenology modeling based on meteorological records and remote sensing observations. *Journal of Geophysical Research: Biogeosciences*, 117(G3).

Zheng, Y., Ji, Y., Dong, H., and Chang, C. (2018). The prevalence of smoking, second-hand smoke exposure, and knowledge of the health hazards of smoking among internal migrants in 12 provinces in China: a cross-sectional analysis. *BMC Public Health*, 18(1):655.

Appendix A

APPENDICES TO CHAPTER 2

A.1 Full Bayesian Hierarchical Model for All-age Smoking Attributable Fraction

The details of the four-layer Bayesian Hierarchical model described in Section 2.2.4 are as follows. Here $\mathcal{N}_l^u(a,b)$ represents a normal distribution with mean a and variance b truncated at interval $[l,u]$ ($l(u)$ is omitted if it takes value $-\infty$ (∞)). Gamma(a,b) represents a Gamma distribution with shape a and rate b. Lognormal(a,b) represents a log-normal distribution with parameters a,b. InvGamma(a,b) represents a inverse-Gamma distribution with shape

a and scale b.

Level 1: $y_{c,s,t}|h_{c,s,t} \sim \mathcal{N}(h_{c,s,t}, \sigma_c^2)$;

Level 2: $h_{c,s,t_0} = g(t_0|\theta_{c,s}) + \varepsilon^h_{c,s,t_0}$,

$h_{c,s,t} = h_{c,s,t-1} + g(t|\theta_{c,s}) - g(t-1|\theta_{c,s}) + \varepsilon^h_{c,s,t}$ for $t > t_0$,

$\varepsilon^h_{c,s,t} \sim_{ind} \mathcal{N}(0, \sigma_h^2)$;

Level 3:

$a_1^{c,m} \sim \text{Gamma}(2, 2/a_1^m)$, $a_1^{c,f} \sim \text{Gamma}(2, 2/a_1^f)$,

$a_2^{c,m} \sim \mathcal{N}^{65}(a_2^m, \sigma^2_{a_2^m})$, $a_2^{c,f} = a_2^{c,m} + \triangle^c_{a_2}$,

$a_3^{c,m} \sim \text{Gamma}(2, 2/a_3^m)$, $\triangle^c_{a_2} \sim \mathcal{N}(\triangle_{a_2}, \sigma^2_{\triangle_{a_2}})$,

$a_4^{c,m} \sim \mathcal{N}_0^{100}(a_4, \sigma^2_{a_4})$, $a_3^{c,f} \sim \text{Gamma}(2, 2/a_3^f)$,

$k^{c,m} \sim \mathcal{N}_0(k^m, \sigma^2_{k^m})$, $a_4^{c,f} \sim \mathcal{N}_0^{100}(a_4, \sigma^2_{a_4})$,

$\sigma_c^2 \sim \text{Lognormal}(\nu, \rho^2)$, $k^{c,f} \sim \mathcal{N}_0(k^f, \sigma^2_{k^f})$;

Level 4:

$a_1^m \sim \text{Gamma}(\alpha_{a_1^m}, \beta_{a_1^m})$, $a_1^f \sim \text{Gamma}(\alpha_{a_1^f}, \beta_{a_1^f})$,

$a_2^m \sim \mathcal{N}(\alpha_{a_2^m}, \beta_{a_1^m})$, $\triangle_{a_2} \sim \mathcal{N}(\alpha_{\triangle_{a_2}}, \beta_{\triangle_{a_2}})$,

$a_3^m \sim \mathcal{N}(\alpha_{a_3^m}, \beta_{a_3^m})$, $a_3^f \sim \text{Gamma}(\alpha_{a_3^f}, \beta_{a_3^f})$,

$a_4 \sim \mathcal{N}(\alpha_{a_4}, \beta_{a_4})$, $k^f \sim \mathcal{N}(\alpha_{k^f}, \beta_{k^f})$,

$k^m \sim \mathcal{N}(\alpha_{k^m}, \beta_{k^m})$, $\sigma^2_{\triangle_{a_2}} \sim \text{InvGamma}(\alpha_{\sigma^2_{\triangle_{a_2}}}, \beta_{\sigma^2_{\triangle_{a_2}}})$,

$\sigma^2_{a_2^m} \sim \text{InvGamma}(\alpha_{\sigma^2_{a_2^m}}, \beta_{\sigma^2_{a_2^m}})$, $\sigma^2_{k^f} \sim \text{InvGamma}(\alpha_{\sigma^2_{k^f}}, \beta_{\sigma^2_{k^f}})$,

$\sigma^2_{a_4} \sim \text{InvGamma}(\alpha_{\sigma^2_{a_4}}, \beta_{\sigma^2_{a_4}})$, $\nu \sim \mathcal{N}(\alpha_\nu, \beta_\nu)$,

$\sigma^2_{k^m} \sim \text{InvGamma}(\alpha_{\sigma^2_{k^m}}, \beta_{\sigma^2_{k^m}})$, $\rho^2 \sim \text{InvGamma}(\alpha_{\rho^2}, \beta_{\rho^2})$,

$\sigma_h^2 \sim \text{InvGamma}(\alpha_{\sigma_h^2}, \beta_{\sigma_h^2})$,

where $\alpha_{a_1^m} = 1.477, \beta_{a_1^m} = 9.423, \alpha_{a_2^m} = 24.362, \beta_{a_2^m} = 12.488, \alpha_{a_3^m} = 1.031, \beta_{a_3^m} = 7.378, \alpha_{a_4} = 38.362, \beta_{a_4} = 19.058, \alpha_{k^m} = 0.362, \beta_{k^m} = 0.255, \alpha_{\sigma^2_{a_2^m}} = 2, \beta_{\sigma^2_{a_2^m}} = 12.488^2, \alpha_{\sigma^2_{a_4}} = 2, \beta_{\sigma^2_{a_4}} = 19.058^2, \alpha_{\sigma^2_{k^m}} = 2, \beta_{\sigma^2_{k^m}} = 0.255^2, \alpha_{a_1^f} = 2.093, \beta_{a_1^f} = 16.302, \alpha_{\triangle_{a_2}} = 12.080, \beta_{\triangle_{a_2}} = 11.140, \alpha_{a_3^f} = 1.031, \beta_{a_3^f} = 7.378, \alpha_{k^f} = 0.362, \beta_{k^f} = 0.255, \alpha_{\sigma^2_{\triangle_{a_2}}} = 2, \beta_{\sigma^2_{\triangle_{a_2}}} =$

$11^2, \alpha_{\sigma^2_{kf}} = 2, \beta_{\sigma^2_{kf}} = 0.255^2, \alpha_\nu = -10.414, \beta_\nu = 1.186^2, \alpha_{\rho^2} = 2, \beta_{\rho^2} = 1.186^2, \alpha_{\sigma^2_h} = 2, \beta_{\sigma^2_h} = 0.01^2$.

A.2 MCMC Convergence Diagnostics

A.2.1 Hyperparameter Diagnostics

In this section, we present the MCMC convergence diagnostics of the hyperparameters in Level 4 of the model in terms of traceplots, Raftery diagnostic statistics (Raftery and Lewis, 1992), and Gelman diagnostic statistics (Gelman and Rubin, 1992a). Table A.1 provides the Gelman and Raftery diagnostic statistics of all hyperparameters. We use 3 chains with 2000 burnin and 8000 samples without thinning for the Gelman diagnostics, and randomly choose one of the chain to perform the Raftery diagnostics. Figure A.1 shows the traceplots of all 8000 samples of hyperparameters.

Table A.1: Diagnostic statistics for hyperparameters. PSRF and 95% UCI are the point estimator and upper bound of the 95% CI of the Gelman potential scale reduction factor. Burn1, Size1, and DF1 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters $q = 0.025, r = 0.0125, s = 0.95$. Burn2, Size2, and DF2 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters $q = 0.975, r = 0.0125, s = 0.95$.

Parameters	Gelman Diag		Raftery Diag					
	PSRF	95% UCI	Burn1	Size1	DF1	Burn2	Size2	DF2
a_1^m	1	1.00	6	1318	2.20	6	1164	1.94
a_2^m	1	1.00	3	710	1.18	6	1504	2.51
a_3^m	1	1.00	6	1584	2.64	6	1424	2.37
a_4	1	1.01	8	1750	2.92	9	2028	3.38
k^m	1	1.01	10	1952	3.25	6	1236	2.06
$\sigma^2_{a_2^m}$	1	1.00	2	640	1.07	6	1730	2.88
$\sigma^2_{a_4}$	1	1.00	21	4410	7.35	12	2132	3.55
$\sigma^2_{k^m}$	1	1.00	6	1334	2.22	8	1448	2.41
a_1^f	1	1.00	2	640	1.07	6	1318	2.20
$\triangle_{a_2}$	1	1.00	4	756	1.26	6	688	1.15
a_3^f	1	1.00	8	1504	2.51	12	1852	3.09
k^f	1	1.00	5	895	1.49	2	640	1.07
$\sigma^2_{\triangle_{a_2}}$	1	1.00	3	696	1.16	4	839	1.40
$\sigma^2_{k^f}$	1	1.00	2	640	1.07	6	1376	2.29
ν	1	1.00	8	1518	2.53	8	1934	3.22
ρ^2	1	1.00	6	1270	2.21	6	1392	2.32
σ_h^2	1	1.00	10	1872	3.12	12	2337	3.90

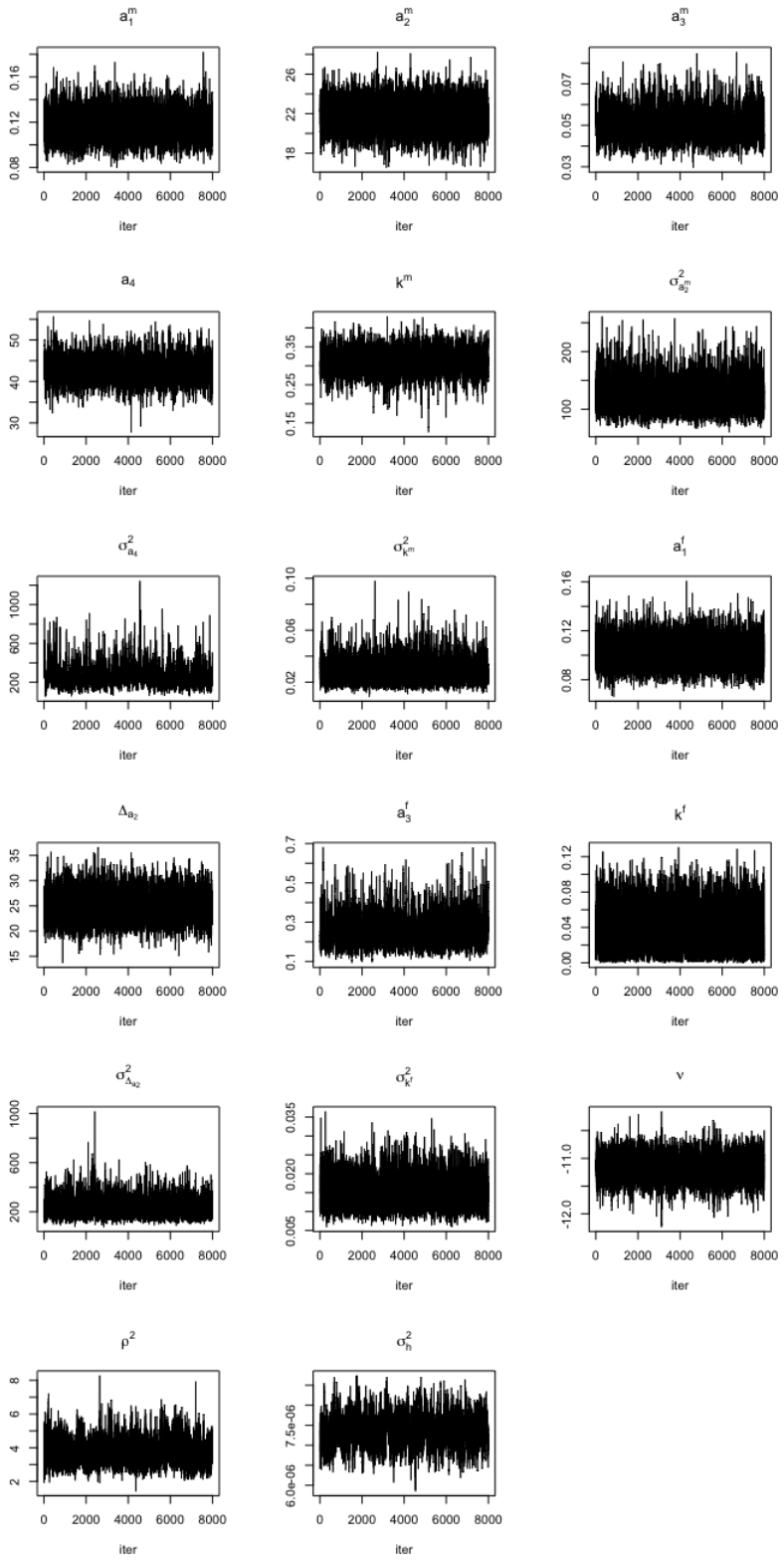


Figure A.1: Traceplots for the hyperparameters in BHM for ASAF.

A.2.2 Country-specific Parameter Diagnostics

In this section, we present the MCMC convergence diagnostics of country-specific parameters of the model in terms of traceplots, Raftery diagnostic statistics, and Gelman diagnostic statistics. Table A.2 provides the Gelman and Raftery diagnostic statistics of country-specific parameters of the United States for male and female. The chains are the same as in the previous section. Figure A.2 shows the traceplots of all 8000 samples of country-specific paremeters for male and female of the United States.

Table A.2: Diagnostic statistics for country-specific parameters for the United States. PSRF and 95% UCI are the point estimator and upper bound of the 95% CI of the Gelman potential scale reduction factor. Burn1, Size1, and DF1 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters $q = 0.025, r = 0.0125, s = 0.95$. Burn2, Size2, and DF2 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters $q = 0.975, r = 0.0125, s = 0.95$.

Parameters	Gelman Diag		Raftery Diag					
	PSRF	95% UCI	Burn1	Size1	DF1	Burn2	Size2	DF2
a_1^m	1	1.00	4	830	1.38	2	640	1.07
a_2^m	1	1.00	6	1326	2.21	4	756	1.26
a_3^m	1	1.00	4	822	1.37	2	633	1.06
a_4^m	1	1.00	2	614	1.02	4	772	1.29
k^m	1	1.00	6	1106	1.10	2	621	1.03
a_1^f	1	1.00	3	661	1.10	2	614	1.02
a_2^f	1	1.00	2	627	1.04	2	614	1.02
a_3^f	1	1.00	2	627	1.04	6	1314	2.19
a_4^f	1	1.00	3	661	1.10	4	848	1.41
k^f	1	1.00	3	668	1.11	6	1444	2.41
σ_c^2	1.01	1.03	96	15198	25.30	24	3834	6.39

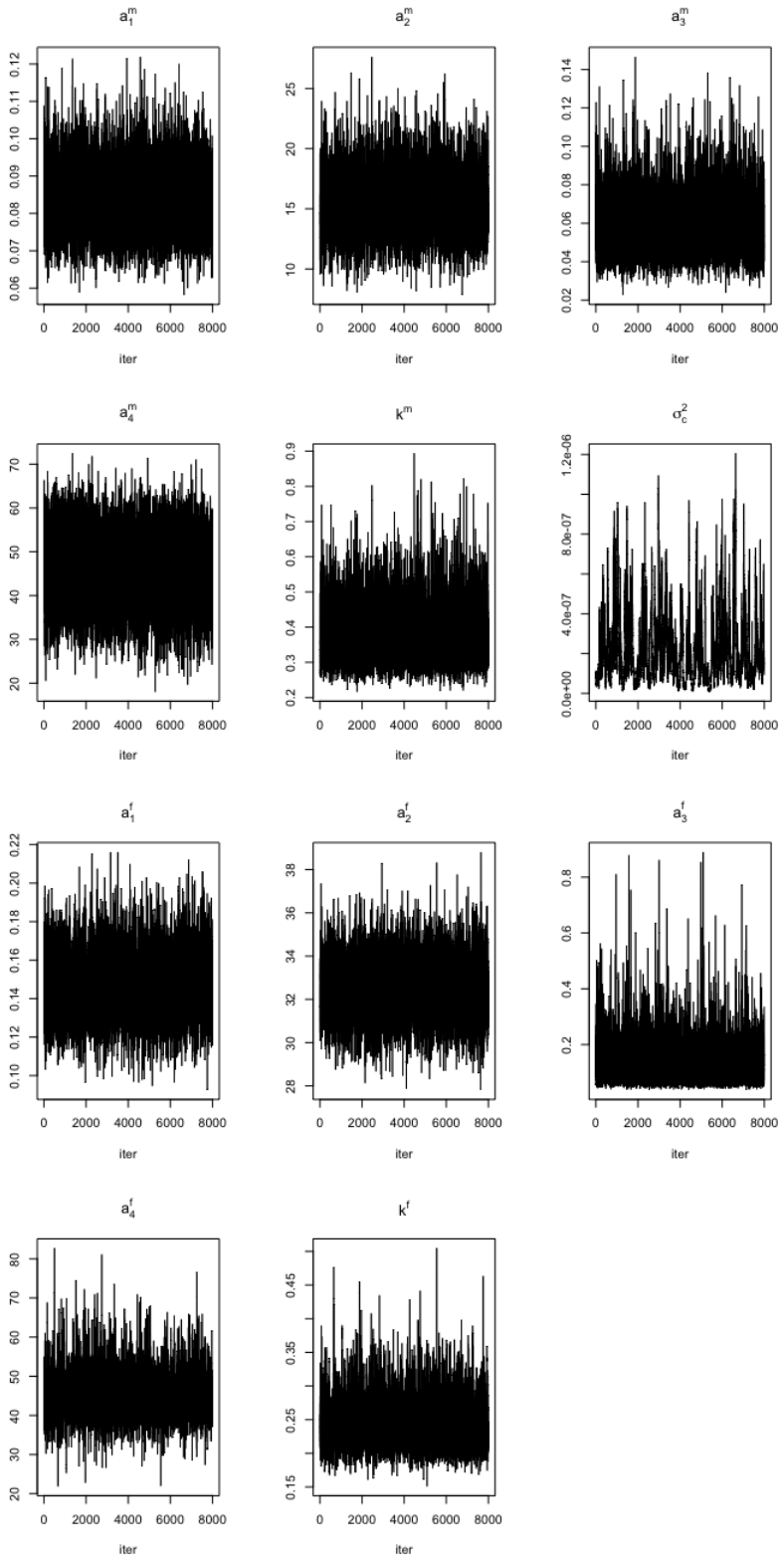


Figure A.2: Traceplots for the country-specific parameters of the United States in BHM for ASAF.

A.3 Hyperparameter Sensitivity Analysis

In this section, we present the sensitivity analysis for the hyperparameters set in $\pi(\cdot)$ on the posterior distributions of the global parameters ψ in Level 4 of our model. We use `rstansensitivity` package (Giordano, 2019) in R to perform the sensitivity analysis. The local sensitivity of the posterior mean of parameter θ under hyperparameters ζ (i.e., $\mathbb{E}(\theta|x,\zeta)$) to ζ at ζ_0 is defined as

$$S_{\zeta_0} := \frac{d\mathbb{E}(\theta|x,\zeta)}{d\zeta}\Big|_{\zeta_0},$$

where x is the observed data. See Basu et al. (1996); Gustafson (1996); Giordano et al. (2018) for more discussions on local sensitivity in Bayesian analysis. By scaling the local sensitivity to be comparable with the possible range of the posterior distribution of θ, the normalized local sensitivity is defined as

$$\tilde{S}_{\zeta_0} := \left|\frac{S_{\zeta_0}}{\mathrm{sd}(\theta|x,\zeta_0)}\right|.$$

As commented in Giordano et al. (2018), if the quantity $\tilde{S}_{\zeta_0}$ is less than 1 or if $\tilde{S}_{\zeta_0}$ is greater than 1 but the final results barely change when modifying the hyperparameters, then the model is robust. First of all, Table A.3 investigates the normalized local sensitivity of the hyperparameters set in $\pi(\cdot)$ on posterior distributions of the global parameters ψ. For most hyperparameters, the normalized local sensitivity are much smaller than 1. For those whose normalized local sensitivity are greater than 1, we conduct out-of-sample validations for three five-year period prediction with the hyperparameters changed to evaluate the actual changes on the validation results. Table A.4, A.5, A.6, and A.7 show the out-of-sample validtion results after modifying $\beta_{\sigma^2_{k_m}}$ (0.255^2 to 1), $\beta_{\sigma^2_{k_f}}$ (0.255^2 to 1), $\beta_{\sigma^2_h}$ (0.01^2 to 0.02^2), and α_{a_4} (38.362 to 20), respectively. All four cases show that the validation results barely change with the hyperparameters, and it is safe to conclude that model is robust to the current choices of hyperparameters.

Table A.3: Normalized local sensitivity of hyperparameters on the global parameters. The bold numbers are local sensitivity $\tilde{S}_{\zeta_0}$ with absolute value greater than 1.

	a_1^m	a_2^m	a_3^m	a_4	k_m	$\sigma^2_{a_2^m}$	$\sigma^2_{a_4}$	σ^2_{km}	a_1^f	$\triangle_{a_2}$	a_3^f	k_f	$\sigma^2_{\triangle_{a_2}}$	$\sigma^2_{k_f}$	ν	ρ^2	σ^2_h
$\alpha_{a_1^m}$	0.014	0.001	0.001	0.046	-0.006	-0.031	0.198	0.000	-0.000	-0.018	0.000	-0.001	-0.096	-0.000	0.003	0.002	-0.000
$\beta_{a_1^m}$	-0.002	-0.000	-0.000	-0.006	0.001	0.003	-0.023	-0.000	0.000	0.002	-0.000	0.000	0.011	0.000	-0.000	-0.000	0.000
$\alpha_{a_2^m}$	0.000	0.012	0.000	-0.001	-0.000	0.005	0.003	0.000	0.000	-0.003	-0.000	0.000	-0.002	0.000	0.000	-0.000	-0.000
$\beta_{a_2^m}$	-0.000	-0.004	-0.000	0.000	-0.000	-0.000	-0.002	0.000	-0.000	0.001	0.000	-0.000	0.002	-0.000	-0.000	0.000	0.000
$\alpha_{a_3^m}$	0.002	0.019	0.011	0.059	-0.009	0.002	0.295	0.001	-0.000	-0.034	0.002	-0.001	-0.094	-0.000	-0.001	-0.001	0.000
$\beta_{a_3^m}$	-0.000	-0.001	-0.001	-0.003	0.000	-0.000	-0.015	-0.000	0.000	0.002	-0.000	0.000	0.005	0.000	0.000	0.000	-0.000
α_{a_4}	0.000	-0.001	0.000	0.020	-0.001	-0.004	0.048	-0.000	0.000	-0.003	0.000	-0.000	-0.011	-0.000	-0.000	-0.000	0.000
β_{a_4}	0.000	-0.001	0.000	0.025	-0.001	-0.005	0.064	-0.000	0.000	-0.004	0.000	-0.000	-0.013	-0.000	-0.000	-0.000	0.000
α_{k^m}	-0.030	-0.004	-0.027	-0.617	0.168	0.290	**-1.591**	-0.032	-0.004	0.186	-0.007	0.010	0.870	0.000	-0.003	0.005	-0.000
β_{k^m}	-0.012	0.015	-0.012	-0.316	0.062	0.162	-0.610	-0.000	-0.003	0.085	-0.004	0.005	0.399	0.001	0.004	0.005	-0.000
$\alpha_{\sigma^2_{a_2^m}}$	0.001	-0.025	-0.000	0.025	-0.002	-1.104	-0.051	0.000	0.002	-0.003	0.003	-0.001	-0.136	-0.000	-0.002	0.007	0.000
$\beta_{\sigma^2_{a_2^m}}$	-0.000	0.000	-0.000	-0.000	0.000	0.009	0.000	-0.000	-0.000	0.000	-0.000	0.000	0.001	0.000	0.000	-0.000	-0.000
$\alpha_{\sigma^2_{a_4}}$	-0.008	-0.026	-0.007	-0.260	0.012	-0.066	-3.916	0.001	-0.002	-0.014	0.010	-0.004	0.144	0.000	-0.006	0.006	0.000
$\beta_{\sigma^2_{a_4}}$	0.000	0.000	0.000	0.001	-0.000	0.000	0.015	-0.000	0.000	0.000	-0.000	0.000	-0.001	-0.000	0.000	-0.000	-0.000
$\alpha_{\sigma^2_{km}}$	-0.001	-0.001	-0.002	0.068	0.016	-0.010	0.178	-0.029	-0.001	0.051	0.001	-0.001	0.084	-0.000	-0.004	0.009	-0.000
$\beta_{\sigma^2_{km}}$	0.012	-0.063	0.055	**-2.644**	-0.468	-0.067	**-10.187**	**1.004**	0.058	**-1.964**	-0.051	0.048	**-3.475**	0.004	0.123	-0.378	0.000
$\alpha_{a_1^f}$	-0.000	0.001	-0.000	0.012	-0.001	-0.057	0.024	0.000	0.013	-0.051	0.004	-0.002	-0.065	-0.000	0.002	0.001	-0.000
$\beta_{a_1^f}$	0.000	-0.000	0.000	-0.001	0.000	0.006	-0.003	-0.000	-0.001	0.005	-0.000	0.000	0.007	0.000	-0.000	-0.000	0.000
$\alpha_{\triangle_{a_2}}$	-0.000	-0.005	-0.000	-0.006	0.001	0.001	0.008	-0.000	-0.001	0.040	-0.000	0.001	0.069	0.000	-0.001	-0.000	0.000
$\beta_{\triangle_{a_2}}$	-0.001	-0.013	-0.001	-0.020	0.002	0.002	0.030	-0.001	-0.002	0.123	-0.001	0.002	0.222	0.000	-0.002	0.000	-0.000
$\alpha_{a_3^f}$	0.000	0.002	0.001	0.003	-0.002	-0.074	-0.286	-0.000	0.004	-0.040	0.069	-0.006	-0.263	-0.002	-0.002	0.004	0.000
$\beta_{a_3^f}$	-0.000	0.001	-0.000	-0.000	0.000	0.019	0.070	0.000	-0.001	0.008	-0.018	0.001	0.059	0.000	0.001	-0.002	-0.000
α_{k^f}	-0.005	0.052	-0.002	-0.071	0.008	0.134	0.500	0.002	-0.007	0.249	-0.022	0.110	0.909	-0.022	0.020	0.010	-0.000
β_{k^f}	0.012	-0.119	0.004	0.151	-0.019	-0.289	**-1.155**	-0.005	0.015	-0.551	0.049	-0.249	**-1.948**	0.049	-0.048	-0.027	0.000
$\alpha_{\sigma^2_{\triangle_{a_2}}}$	0.003	0.013	0.003	0.063	-0.006	-0.129	0.217	0.001	0.002	-0.184	0.008	-0.007	**-2.611**	-0.002	-0.004	0.001	0.000
$\beta_{\sigma^2_{\triangle_{a_2}}}$	-0.000	-0.000	-0.000	-0.000	0.000	0.001	-0.001	-0.000	-0.000	0.001	-0.000	0.000	0.012	0.000	0.000	-0.000	-0.000
$\alpha_{\sigma^2_{k^f}}$	0.000	-0.001	0.000	0.052	0.000	-0.021	0.044	-0.000	0.001	-0.018	0.007	0.021	-0.214	-0.016	0.002	0.008	0.000
$\beta_{\sigma^2_{k^f}}$	-0.025	-0.139	-0.021	**-3.454**	-0.022	**1.940**	**-2.054**	0.029	-0.042	**1.225**	-0.486	**-1.595**	**14.056**	**1.093**	-0.124	-0.655	-0.000
α_ν	0.001	0.013	-0.000	-0.012	-0.000	0.018	0.034	0.001	0.001	-0.014	-0.001	0.002	0.019	-0.000	0.079	-0.002	-0.000
β_ν	-0.001	-0.012	0.000	0.014	0.000	-0.029	-0.013	-0.001	-0.001	0.019	0.001	-0.001	-0.016	0.000	-0.098	0.005	0.000
α_{ρ^2}	-0.000	0.003	0.000	0.010	-0.000	0.051	0.039	0.000	-0.000	0.001	-0.002	-0.001	0.017	0.000	0.002	-0.165	-0.000
β_{ρ^2}	0.000	-0.001	-0.000	-0.003	-0.000	-0.014	-0.013	-0.000	0.000	-0.000	0.000	0.000	-0.005	-0.000	-0.001	0.046	0.000
$\alpha_{\sigma^2_h}$	0.000	0.001	-0.000	-0.000	0.001	0.016	0.043	-0.000	0.001	-0.000	-0.000	0.000	0.021	0.000	0.003	-0.007	-0.000
$\beta_{\sigma^2_h}$	**-37.688**	**-116.830**	**16.011**	**17.100**	**-82.440**	**-2098.717**	**-5650.674**	**27.425**	**-71.201**	**10.789**	**37.786**	**-66.317**	**-2863.039**	**-1.673**	**-434.118**	**958.762**	**4.054**

Table A.4: Out-of-sample validation results of ASAF for both male and female with $\beta_{\sigma^2_{k_m}}$ changed. "Bayes (mod)" is the BHM with changed hyperparameter.

Gender	Train	Test	num	Method	MAE	Coverage		
						80%	90%	95%
Male	1950-2000	2000-2015	63	Bayes(mod)	0.016	0.63	0.76	0.85
				Bayes	0.016	0.65	0.76	0.84
Female	1950-2000	2000-2015	63	Bayes(mod)	0.011	0.80	0.89	0.94
				Bayes	0.011	0.81	0.90	0.95

Table A.5: Out-of-sample validation results of ASAF for both male and female with $\beta_{\sigma^2_{k_f}}$ changed. "Bayes (mod)" is the BHM with changed hyperparameter.

Gender	Train	Test	num	Method	MAE	Coverage		
						80%	90%	95%
Male	1950-2000	2000-2015	63	Bayes(mod)	0.016	0.64	0.77	0.85
				Bayes	0.016	0.65	0.76	0.84
Female	1950-2000	2000-2015	63	Bayes(mod)	0.011	0.82	0.90	0.95
				Bayes	0.011	0.81	0.90	0.95

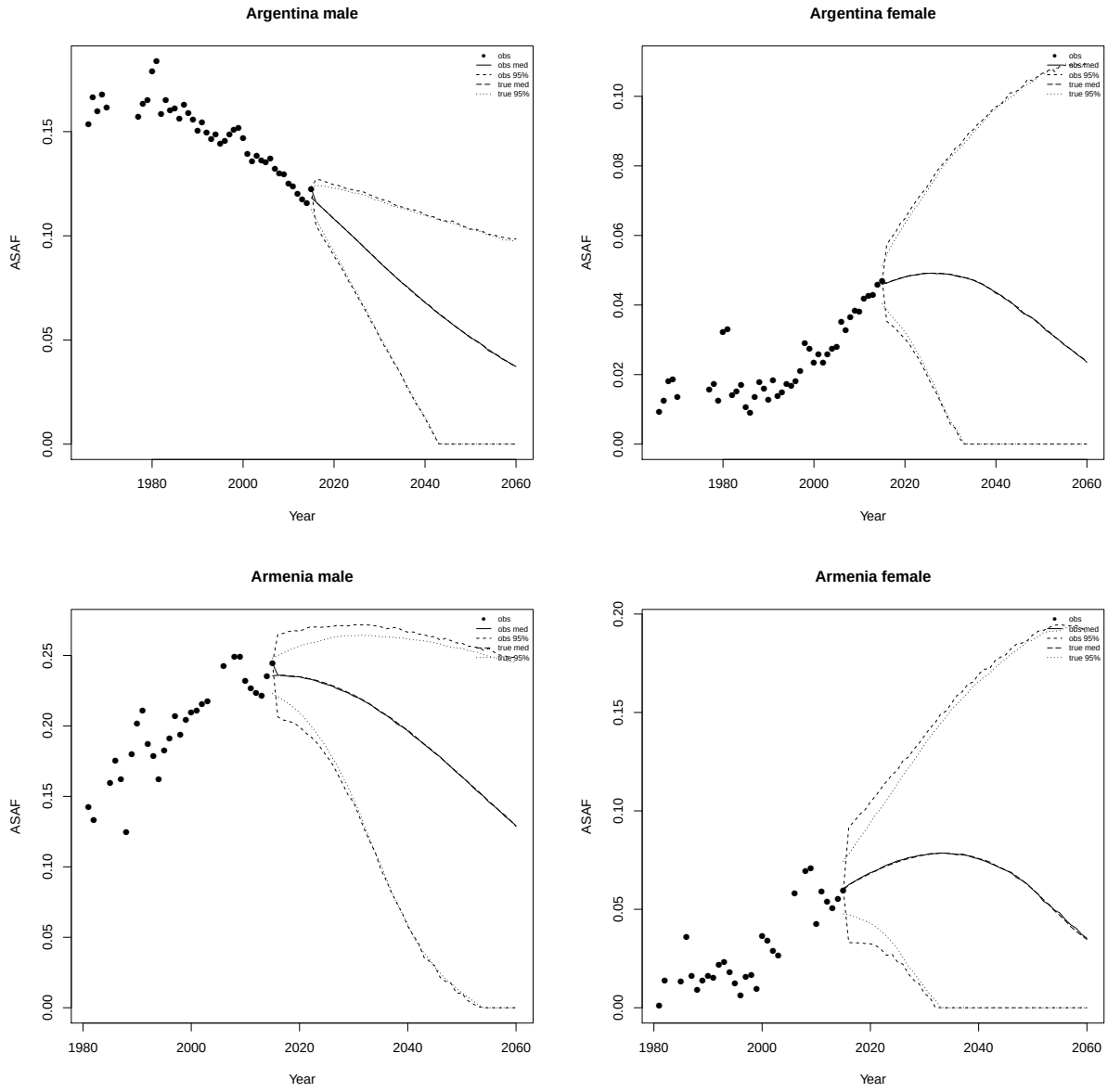
Table A.6: Out-of-sample validation results of ASAF for both male and female with $\beta_{\sigma_h^2}$ changed. "Bayes (mod)" is the BHM with changed hyperparameter.

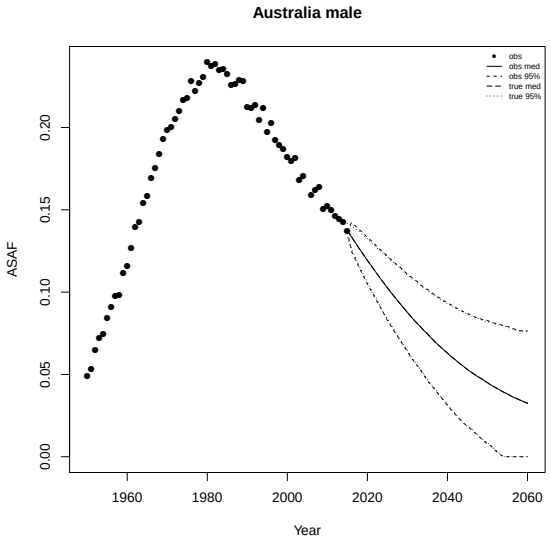
Gender	Train	Test	num	Method	MAE	Coverage		
						80%	90%	95%
Male	1950-2000	2000-2015	63	Bayes(mod)	0.016	0.68	0.80	0.88
				Bayes	0.016	0.65	0.76	0.84
Female	1950-2000	2000-2015	63	Bayes(mod)	0.011	0.82	0.90	0.95
				Bayes	0.011	0.81	0.90	0.95

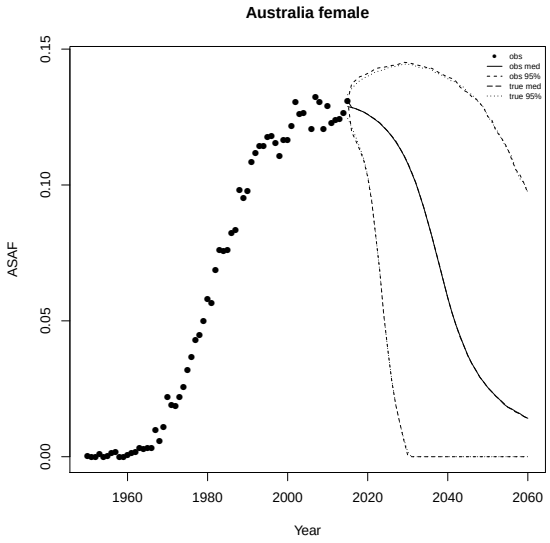
Table A.7: Out-of-sample validation results of ASAF for both male and female with α_{a_4} changed. "Bayes (mod)" is the BHM with changed hyperparameter.

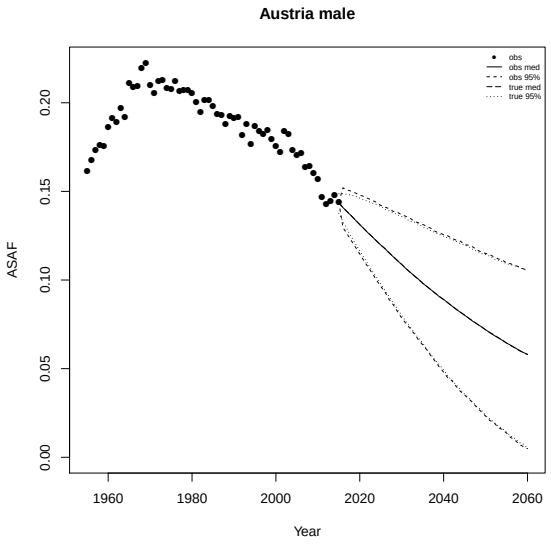
Gender	Train	Test	num	Method	MAE	Coverage		
						80%	90%	95%
Male	1950-2000	2000-2015	63	Bayes(mod)	0.016	0.65	0.76	0.85
				Bayes	0.016	0.65	0.76	0.84
Female	1950-2000	2000-2015	63	Bayes(mod)	0.011	0.81	0.89	0.95
				Bayes	0.011	0.81	0.90	0.95

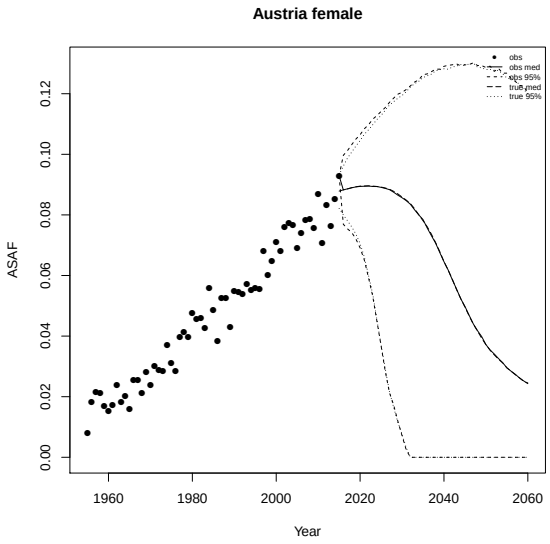
A.4 All-age Smoking Attribuable Fraction Projection to 2050 for Over 60 Countries

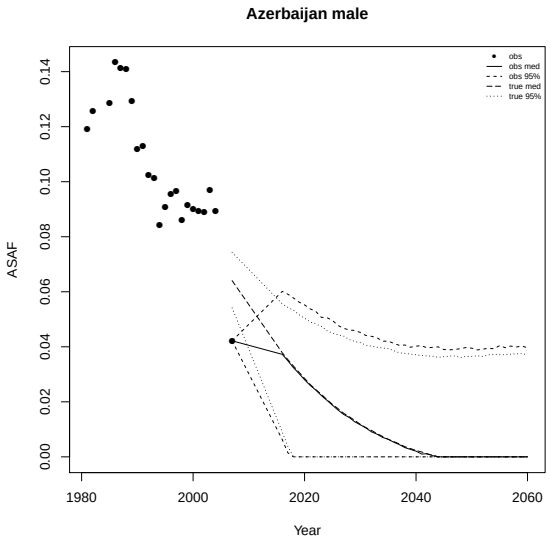


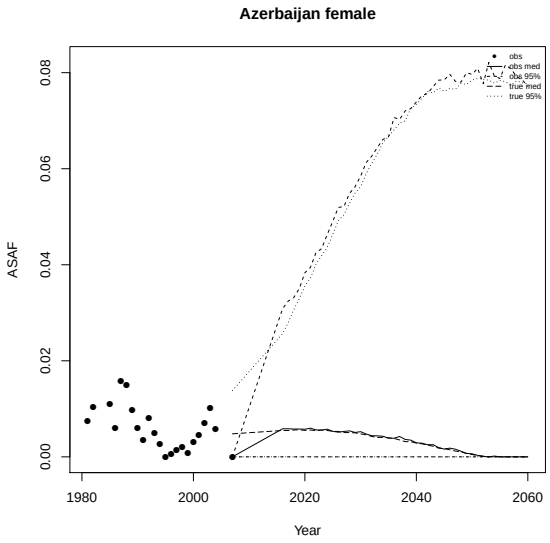

Australia male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1960
1980
2000
2020
2040
2060
Year


Australia female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
1960
1980
2000
2020
2040
2060
Year

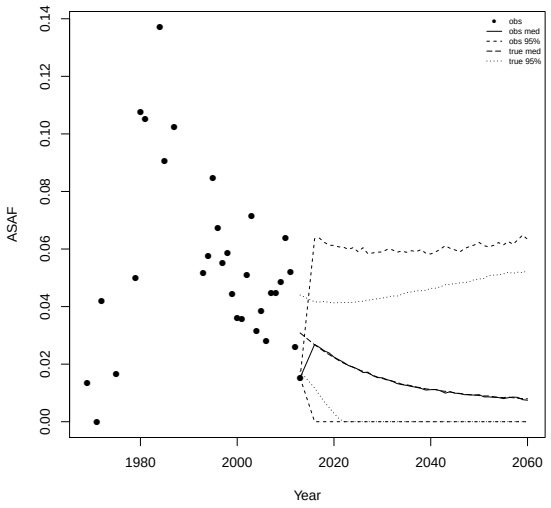

Austria male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1960
1980
2000
2020
2040
2060
Year


Austria female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1960
1980
2000
2020
2040
2060
Year

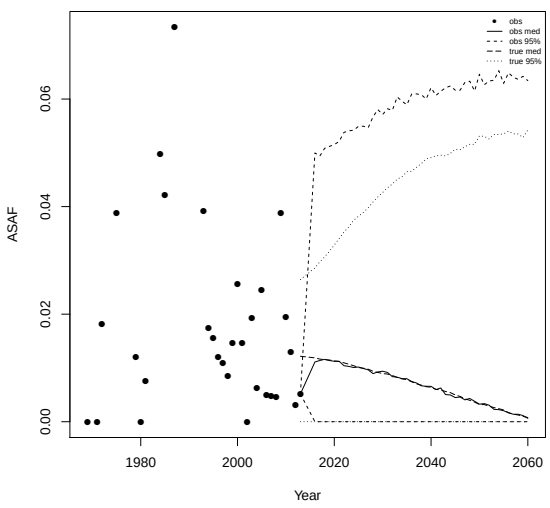

Azerbaijan male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1980
2000
2020
2040
2060
Year


Azerbaijan female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
1980
2000
2020
2040
2060
Year

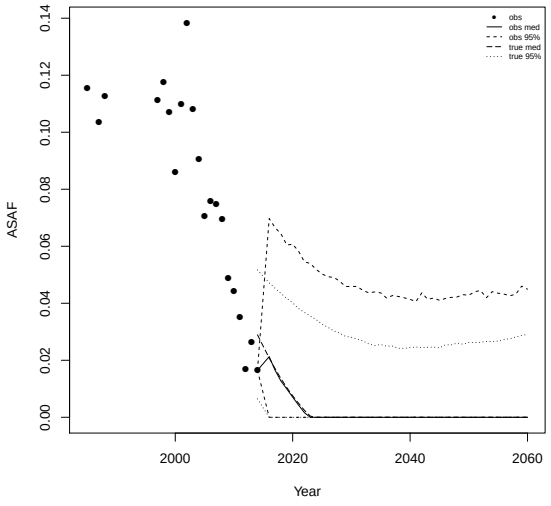
Bahamas male



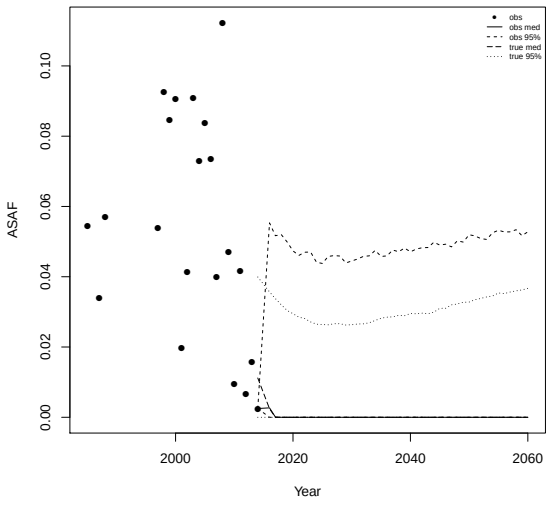
Bahamas female



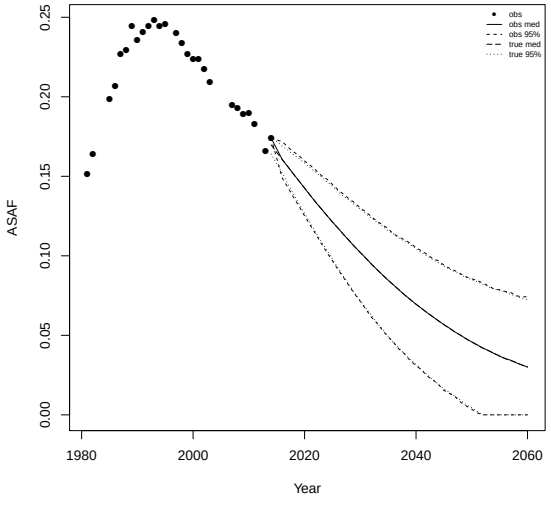
Bahrain male



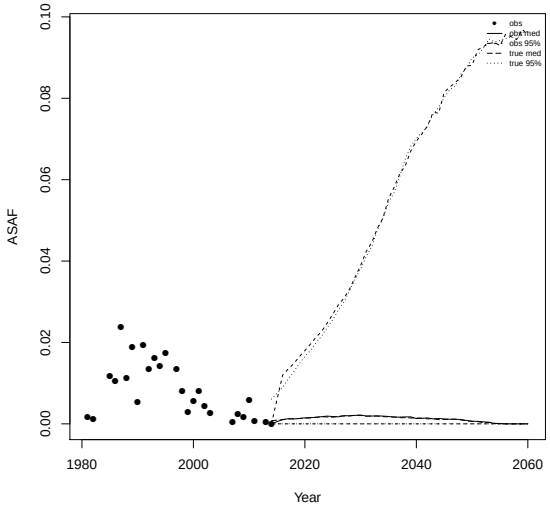
Bahrain female

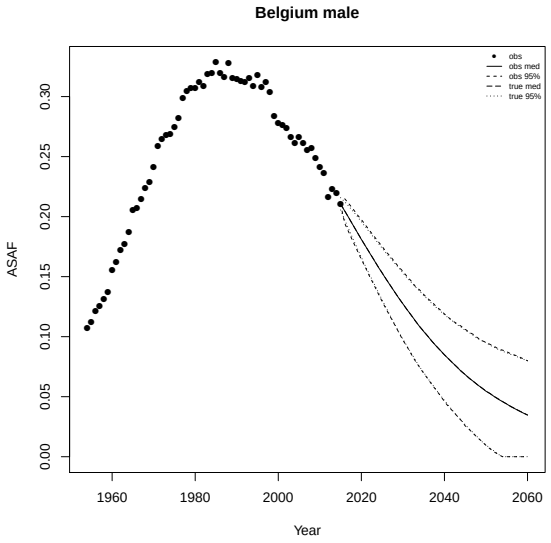


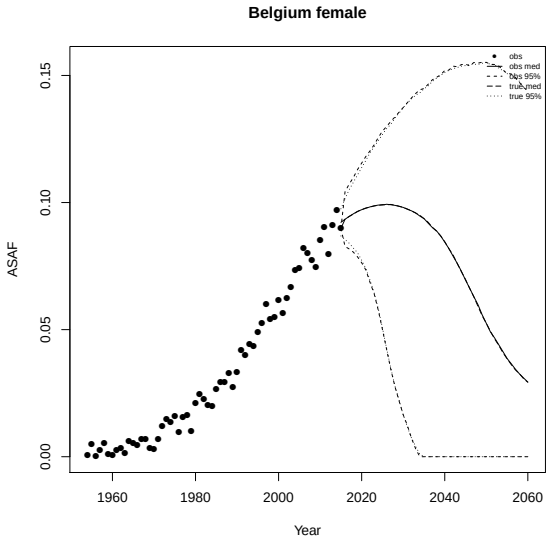
Belarus male

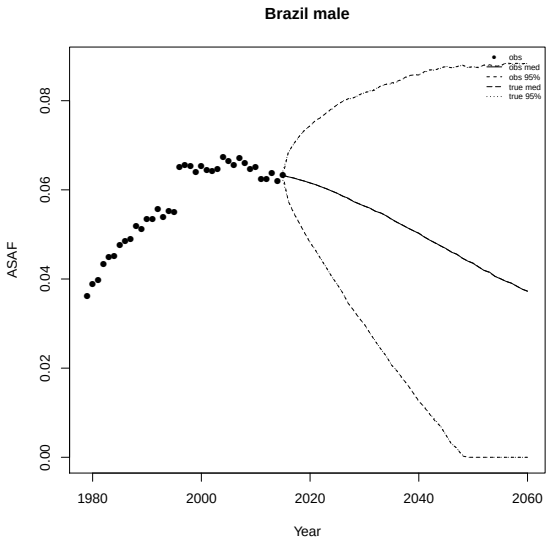


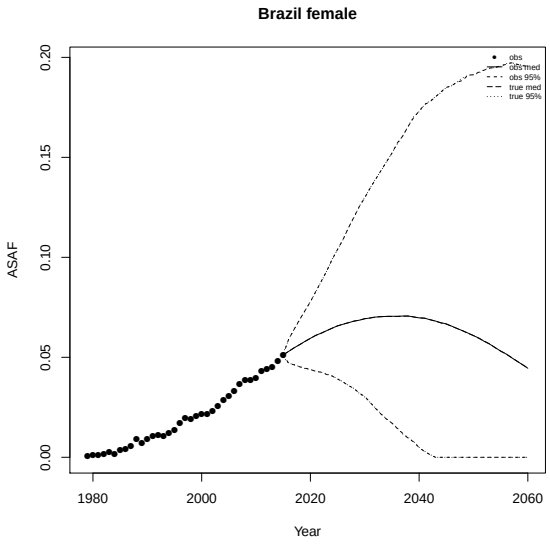
Belarus female

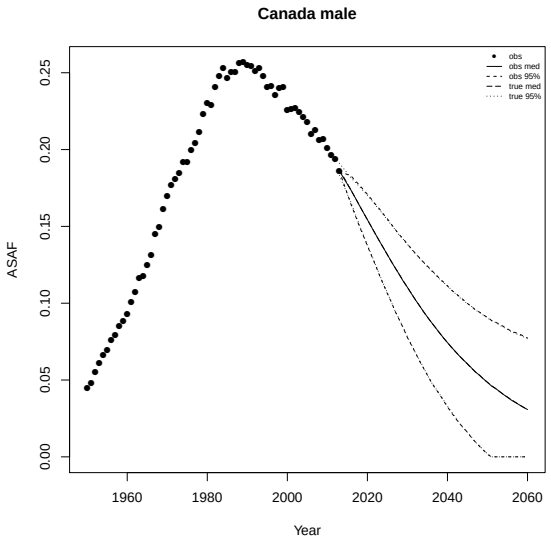


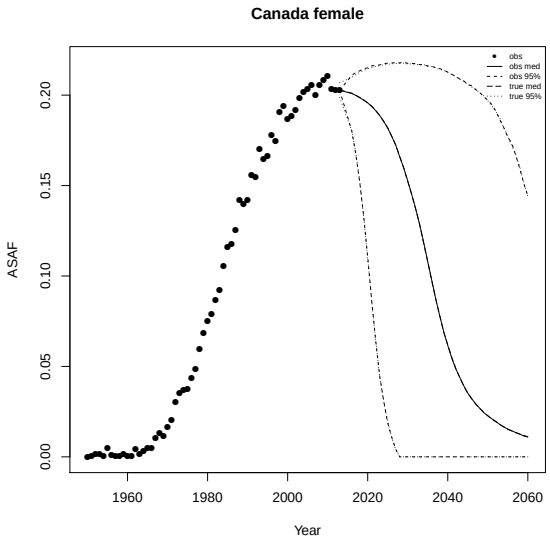

Belgium male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1960
1980
2000
2020
2040
2060
Year

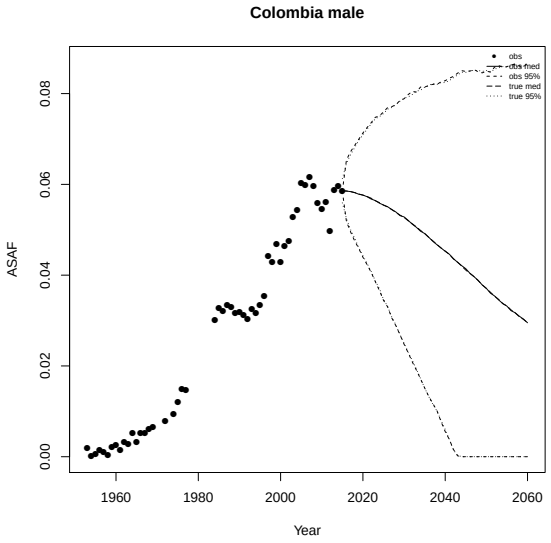

Belgium female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
1960
1980
2000
2020
2040
2060
Year

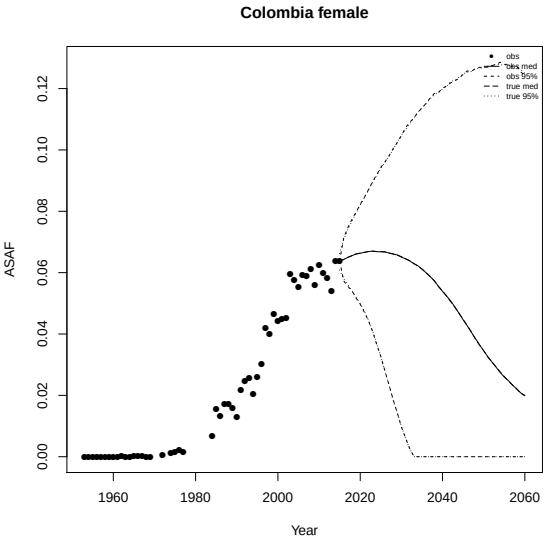

Brazil male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
1980
2000
2020
2040
2060
Year

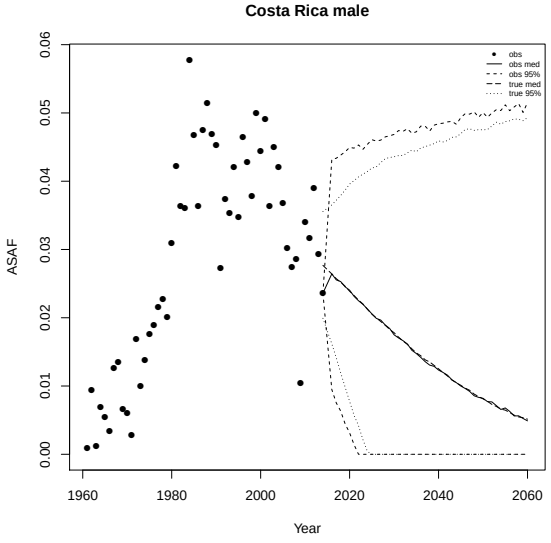

Brazil female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1980
2000
2020
2040
2060
Year

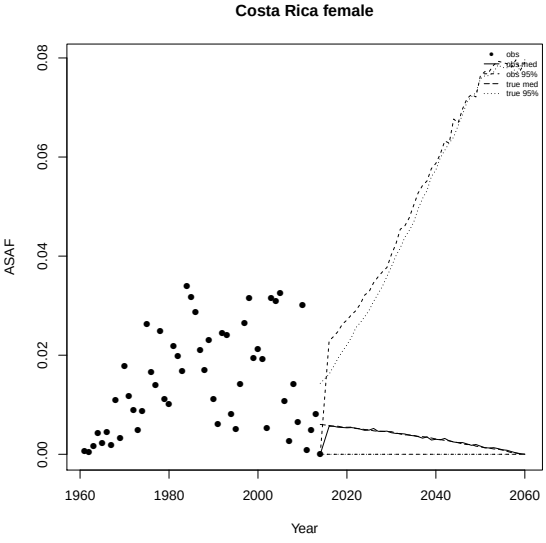

Canada male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
1960
1980
2000
2020
2040
2060
Year

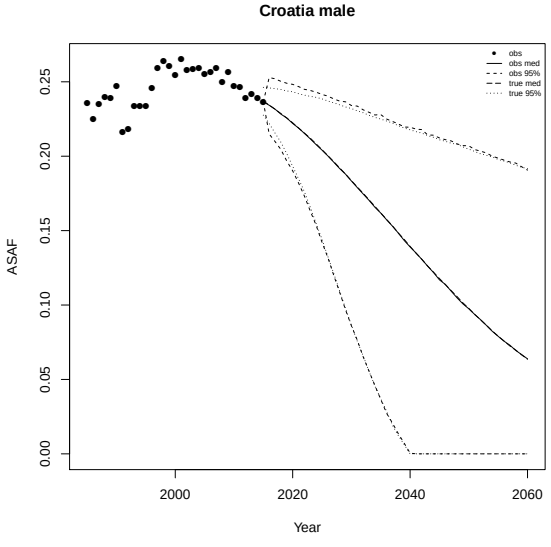

Canada female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1960
1980
2000
2020
2040
2060
Year

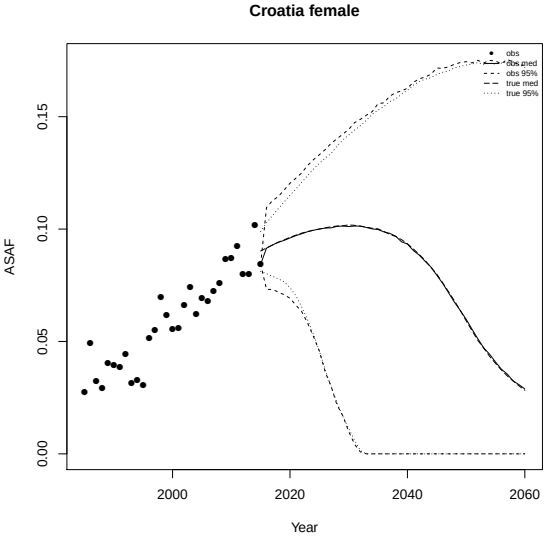

Colombia male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.02
0.04
0.06
0.08
1960
1980
2000
2020
2040
2060

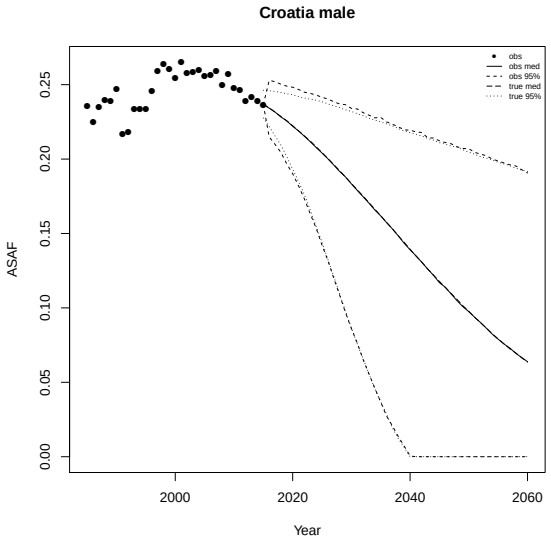

Colombia female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1960
1980
2000
2020
2040
2060

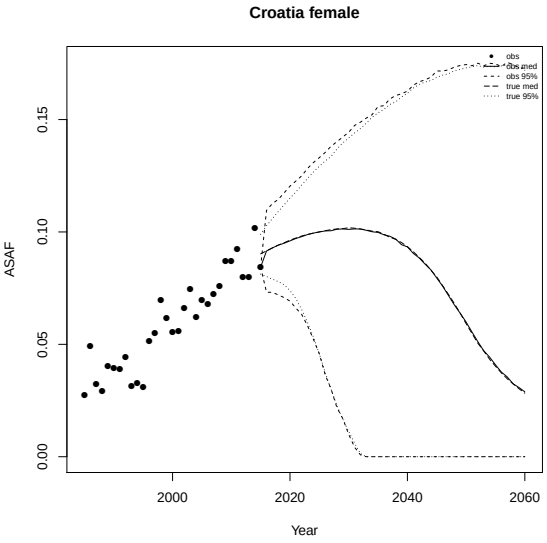

Costa Rica male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.01
0.02
0.03
0.04
0.05
0.06
1960
1980
2000
2020
2040
2060

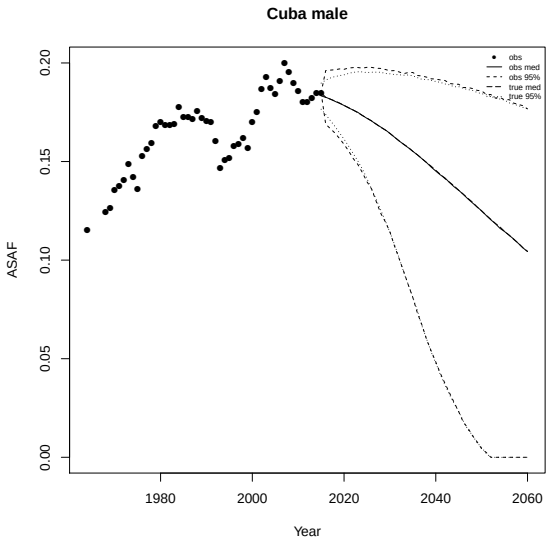

Costa Rica female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.02
0.04
0.06
0.08
1960
1980
2000
2020
2040
2060

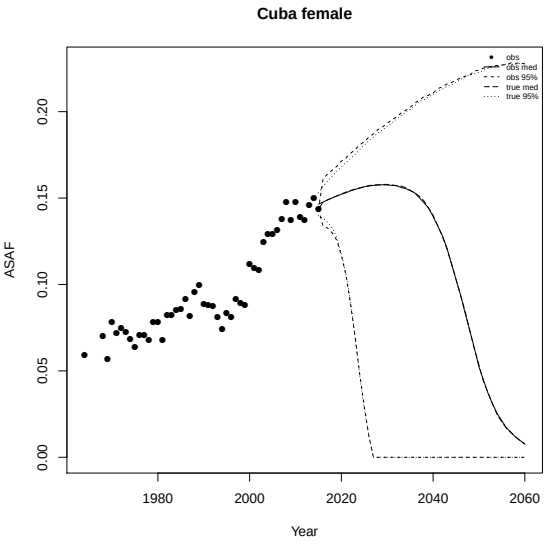

Croatia male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.05
0.10
0.15
0.20
0.25
2000
2020
2040
2060

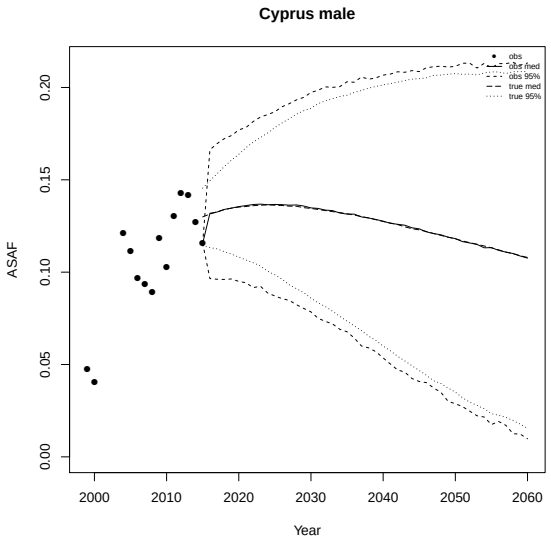

Croatia female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.05
0.10
0.15
2000
2020
2040
2060

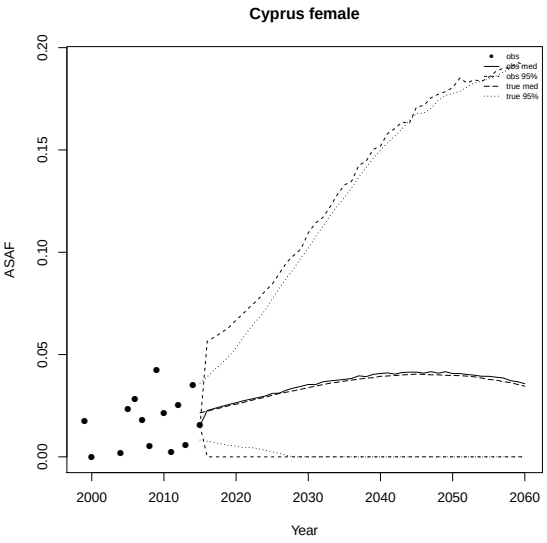

Croatia male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

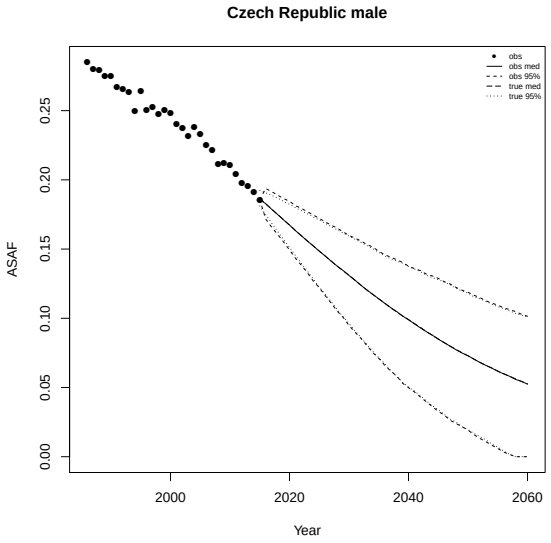

Croatia female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

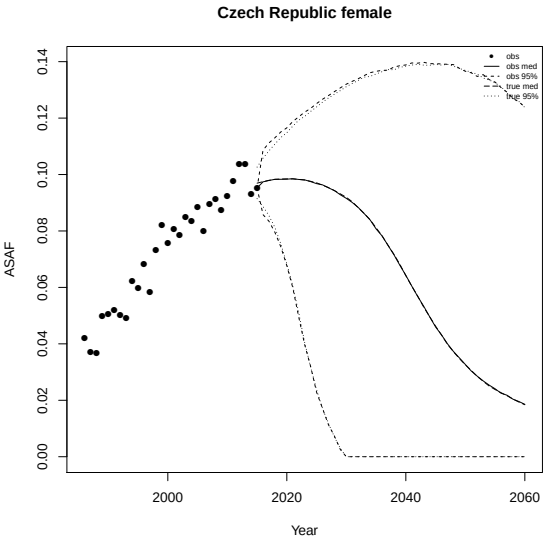

Cuba male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

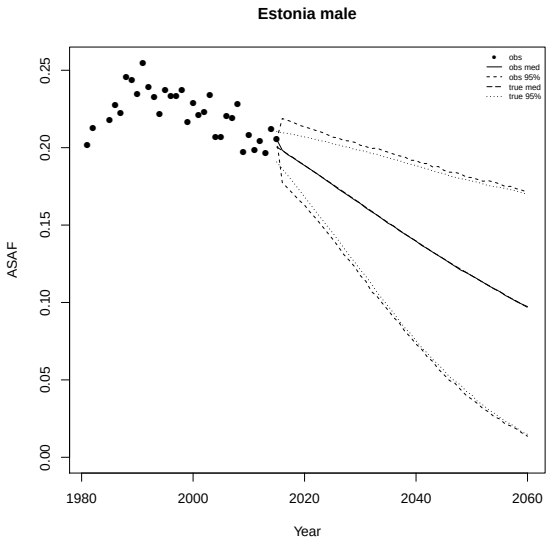

Cuba female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

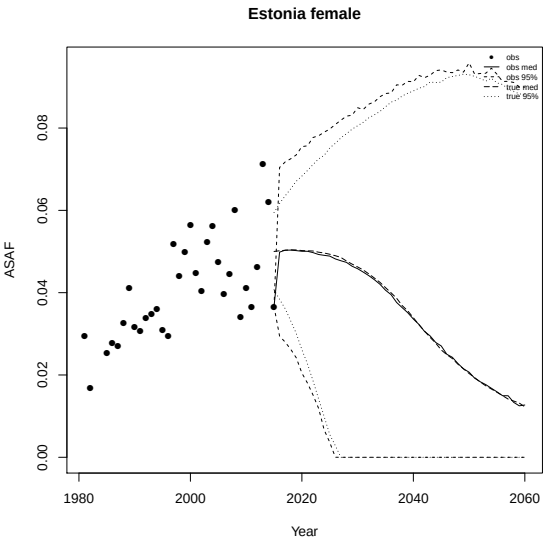

Cyprus male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

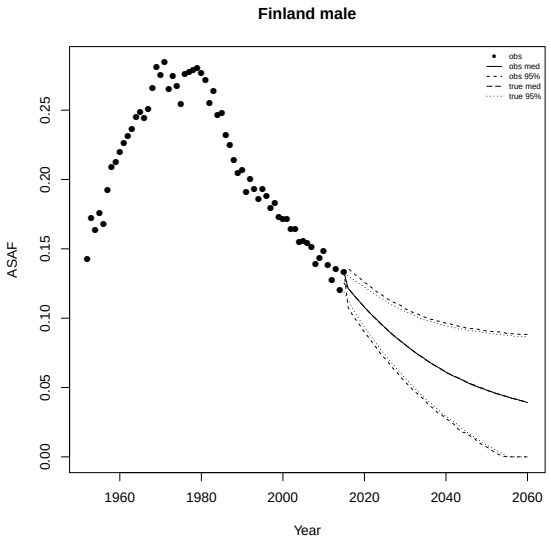

Cyprus female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

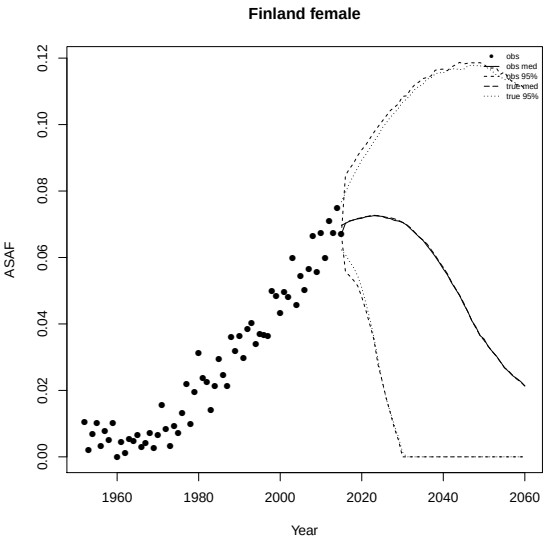

Czech Republic male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.05
0.10
0.15
0.20
0.25
2000
2020
2040
2060

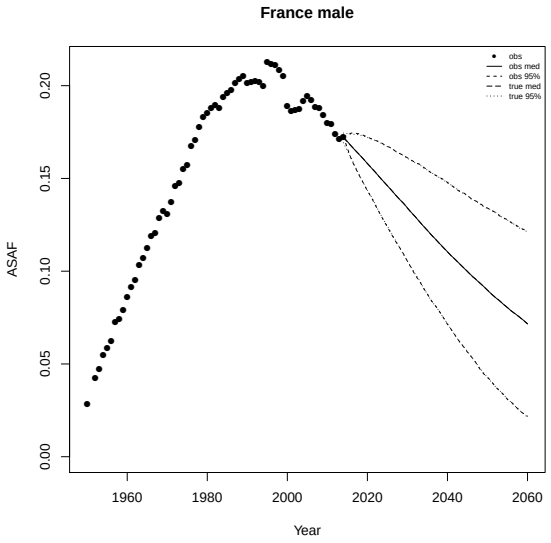

Czech Republic female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
2000
2020
2040
2060

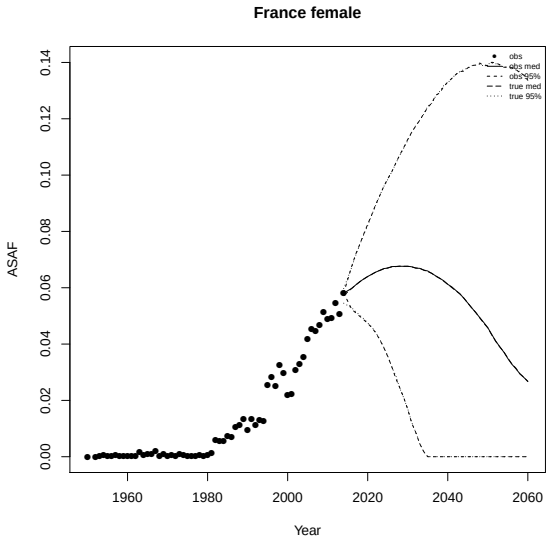

Estonia male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.05
0.10
0.15
0.20
0.25
1980
2000
2020
2040
2060

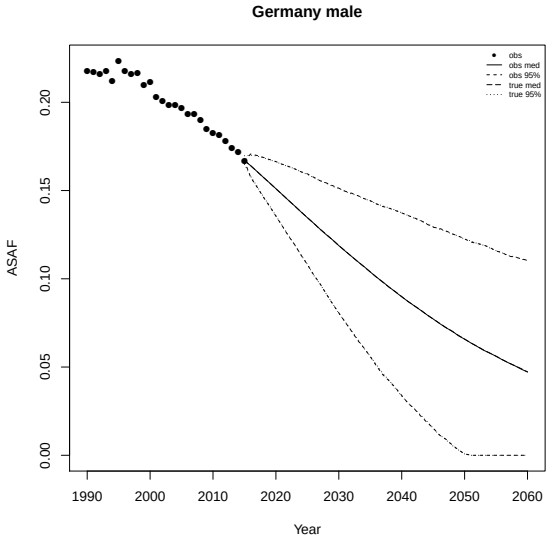

Estonia female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.02
0.04
0.06
0.08
1980
2000
2020
2040
2060

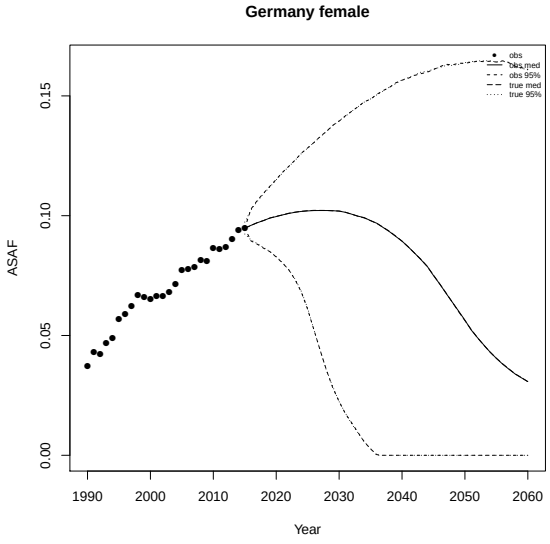

Finland male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.05
0.10
0.15
0.20
0.25
1960
1980
2000
2020
2040
2060

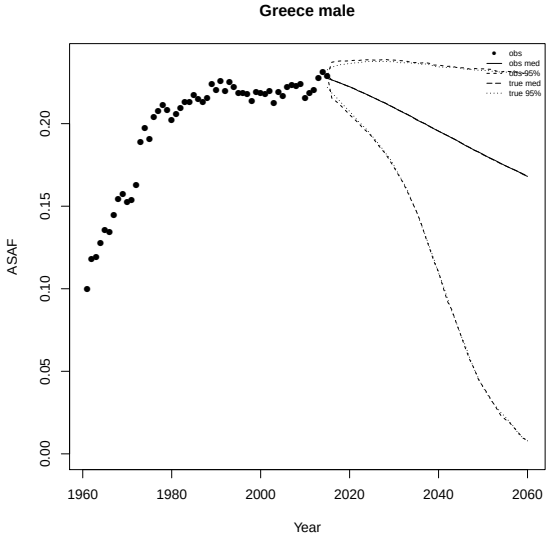

Finland female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1960
1980
2000
2020
2040
2060

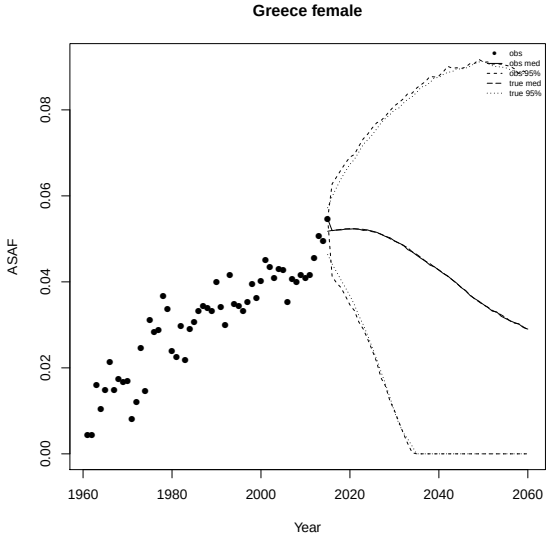

France male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1960
1980
2000
2020
2040
2060
Year

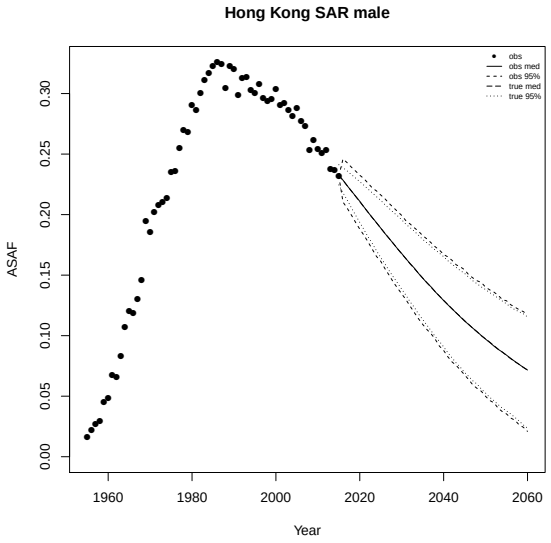

France female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1960
1980
2000
2020
2040
2060
Year

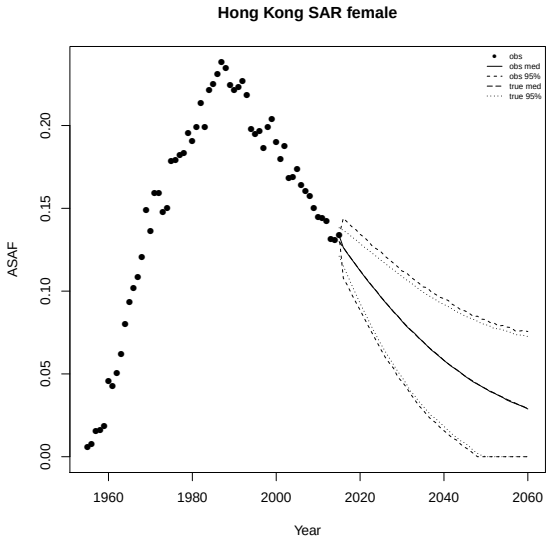

Germany male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1990
2000
2010
2020
2030
2040
2050
2060
Year

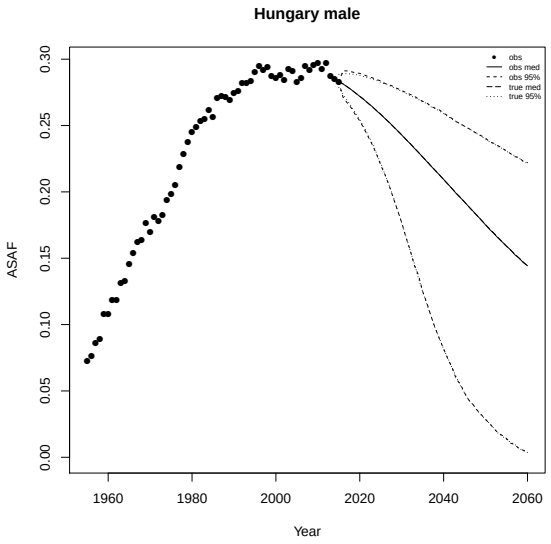

Germany female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
1990
2000
2010
2020
2030
2040
2050
2060
Year

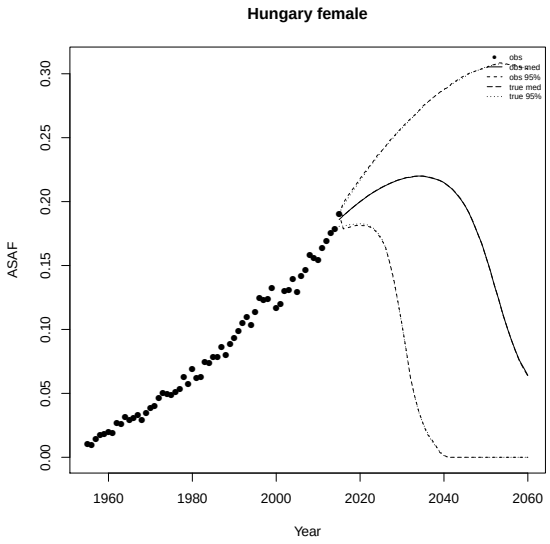

Greece male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1960
1980
2000
2020
2040
2060
Year

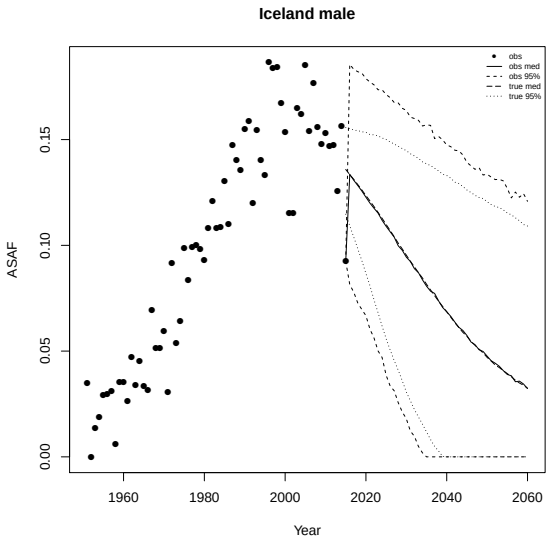

Greece female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
1960
1980
2000
2020
2040
2060
Year

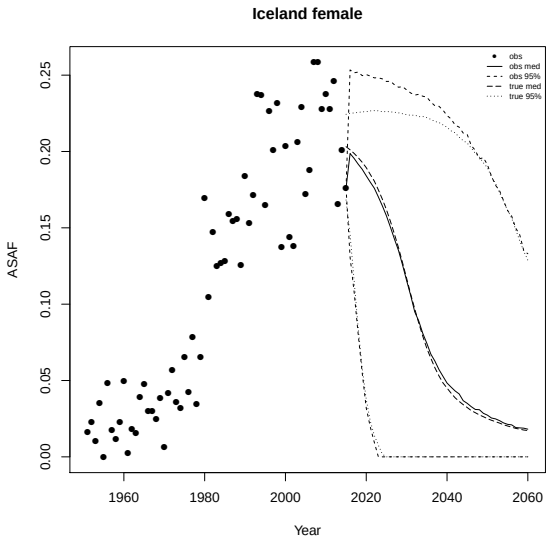

Hong Kong SAR male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1960
1980
2000
2020
2040
2060
Year

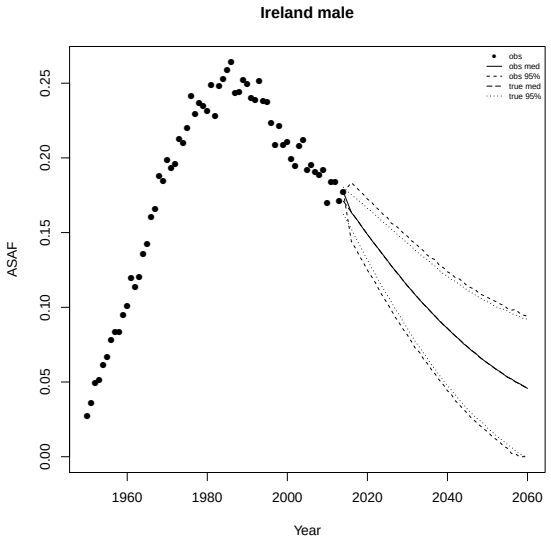

Hong Kong SAR female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1960
1980
2000
2020
2040
2060
Year

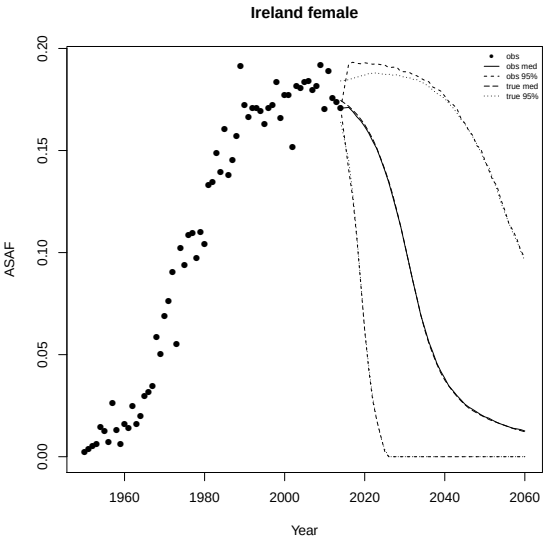

Hungary male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1960
1980
2000
2020
2040
2060
Year

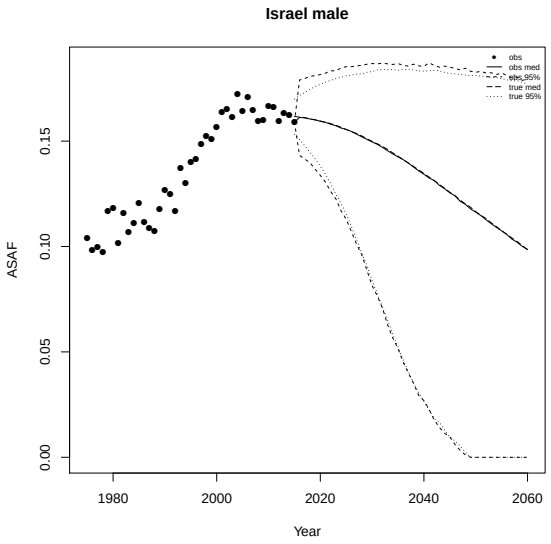

Hungary female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1960
1980
2000
2020
2040
2060
Year

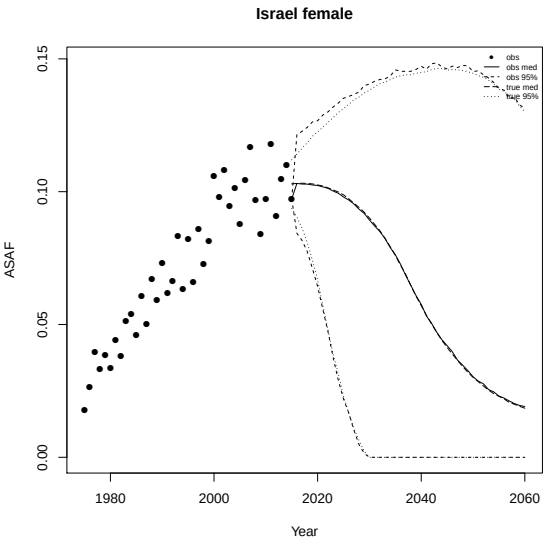

Iceland male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
1960
1980
2000
2020
2040
2060
Year

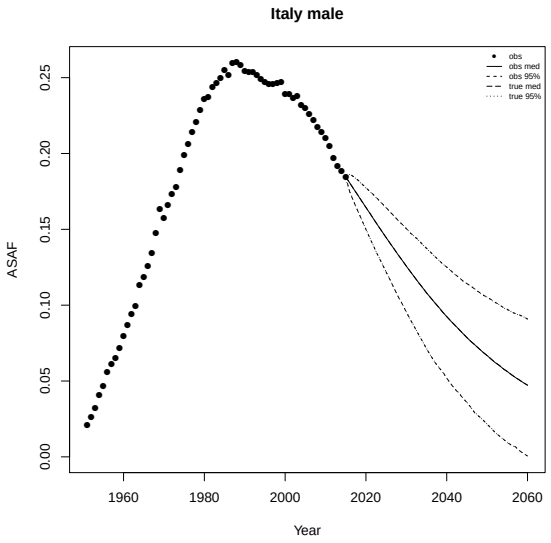

Iceland female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
1960
1980
2000
2020
2040
2060
Year

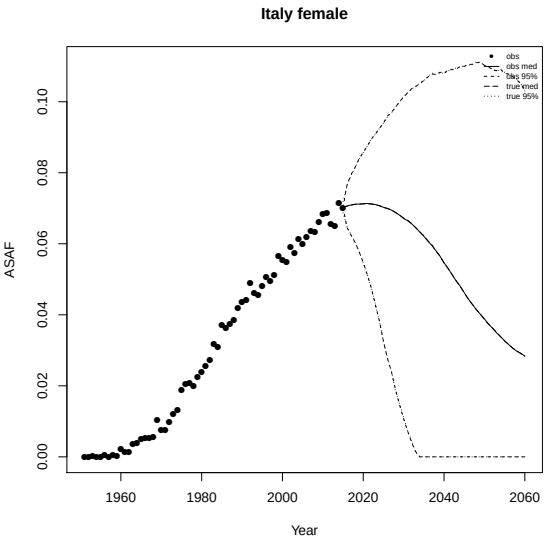

Ireland male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

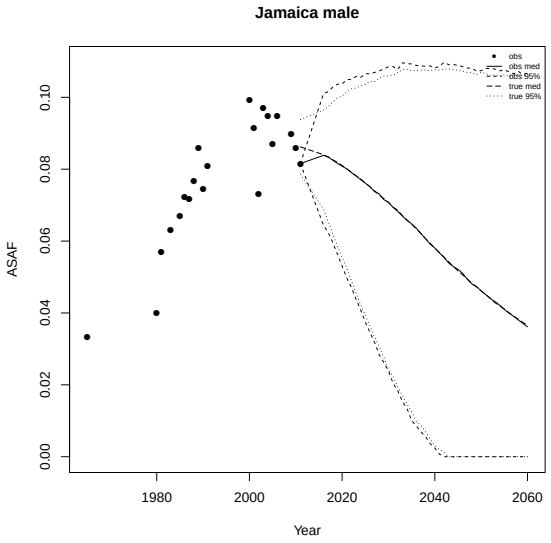

Ireland female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

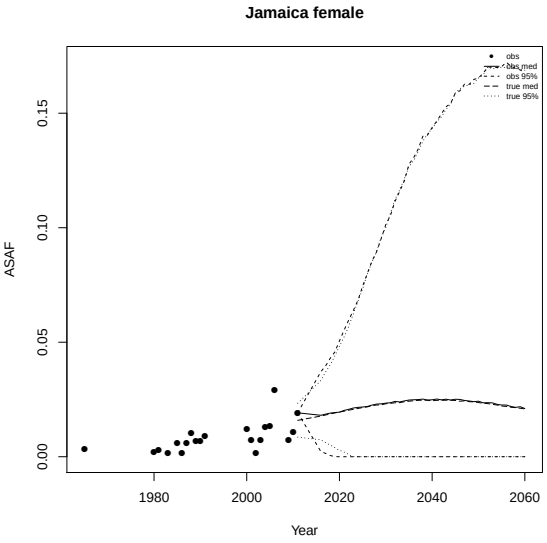

Israel male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

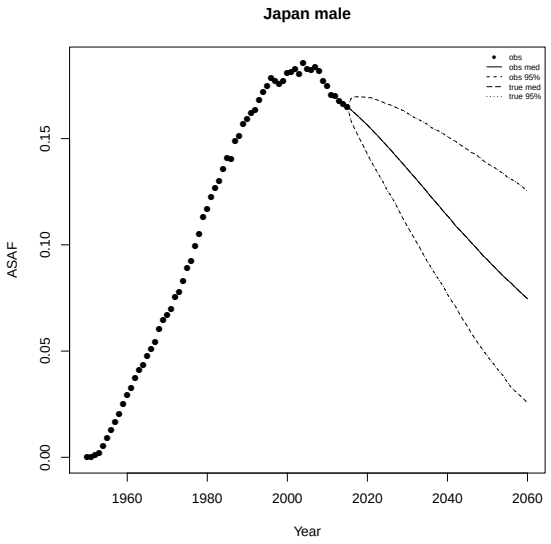

Israel female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

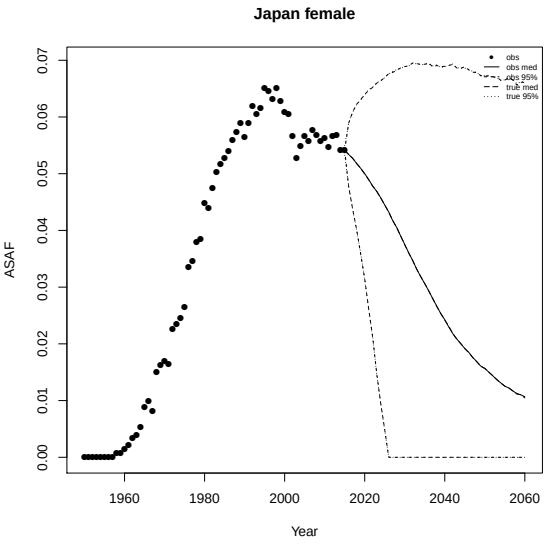

Italy male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

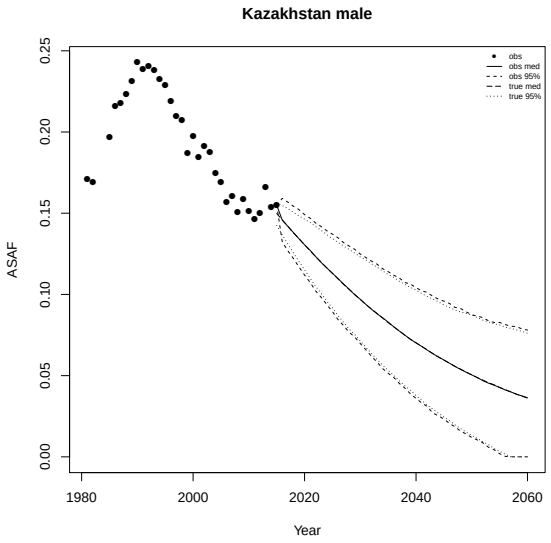

Italy female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

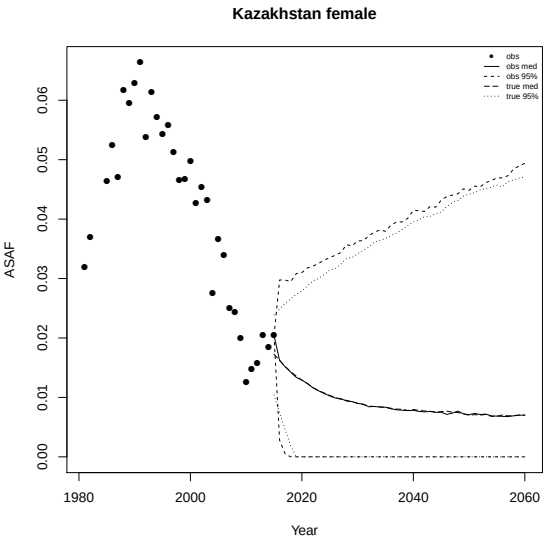

Jamaica male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

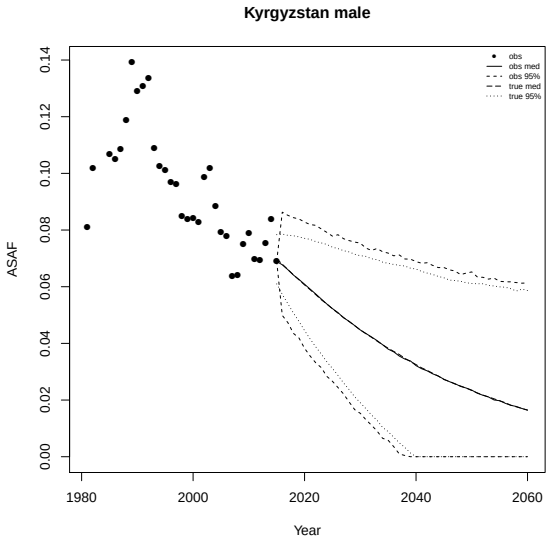

Jamaica female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

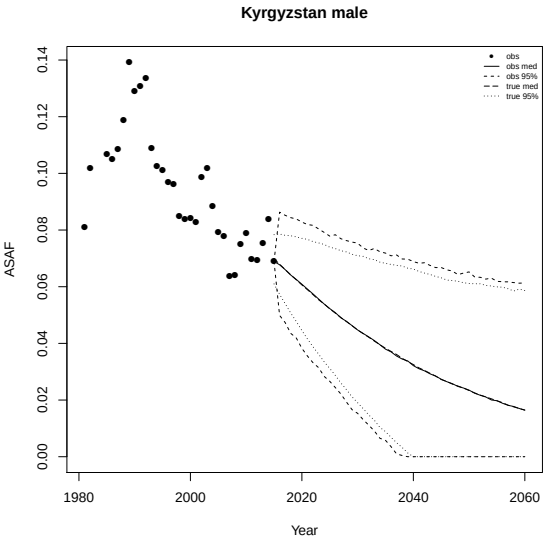

Japan male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

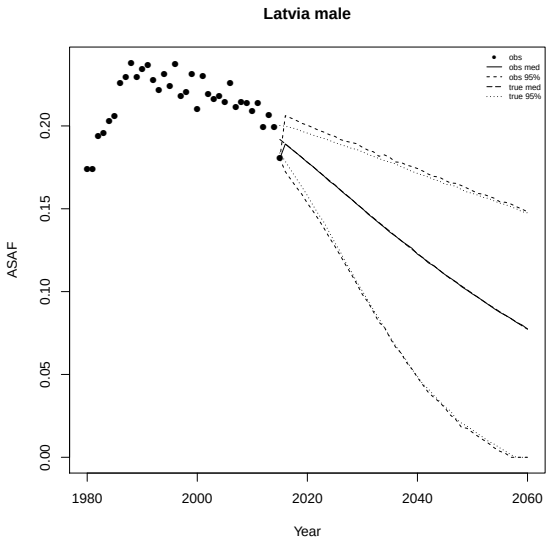

Japan female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

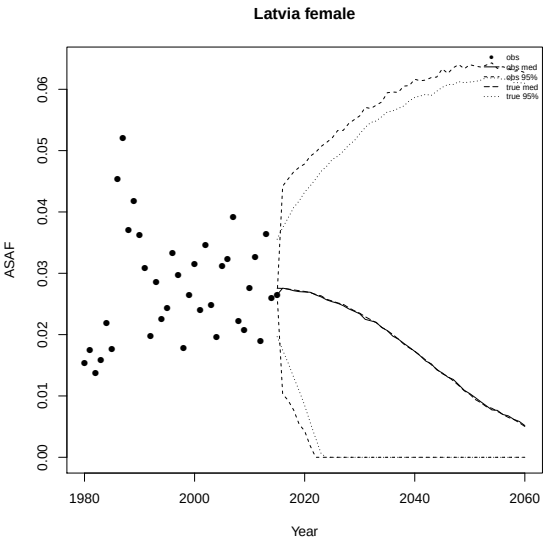

Kazakhstan male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

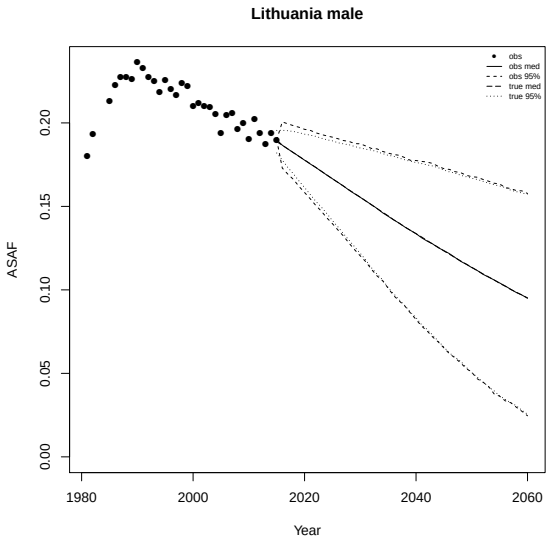

Kazakhstan female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

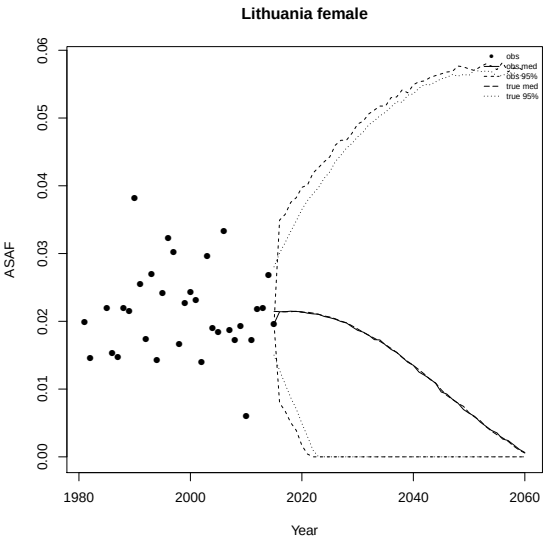

Kyrgyzstan male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1980
2000
2020
2040
2060
Year

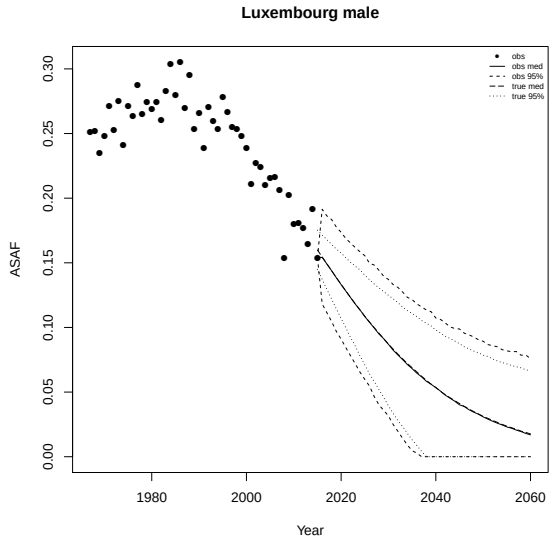

Kyrgyzstan male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1980
2000
2020
2040
2060
Year

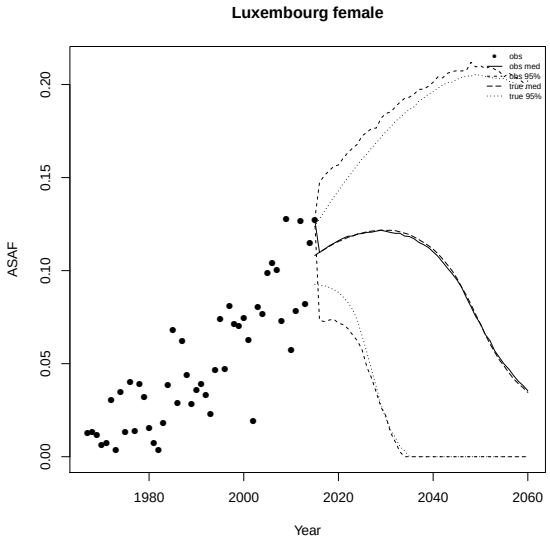

Latvia male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1980
2000
2020
2040
2060
Year

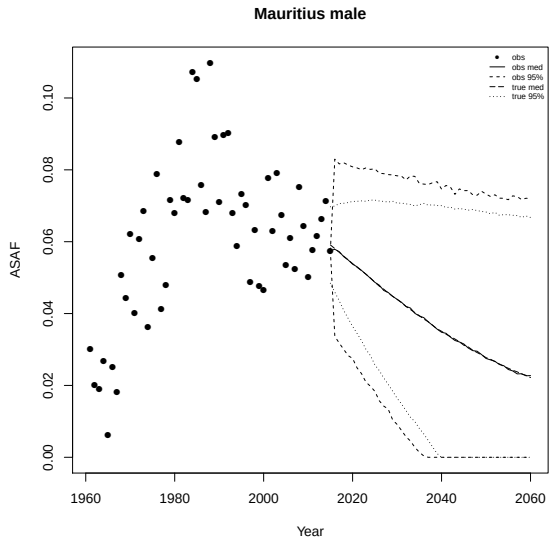

Latvia female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.01
0.02
0.03
0.04
0.05
0.06
1980
2000
2020
2040
2060
Year


Lithuania male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
1980
2000
2020
2040
2060
Year

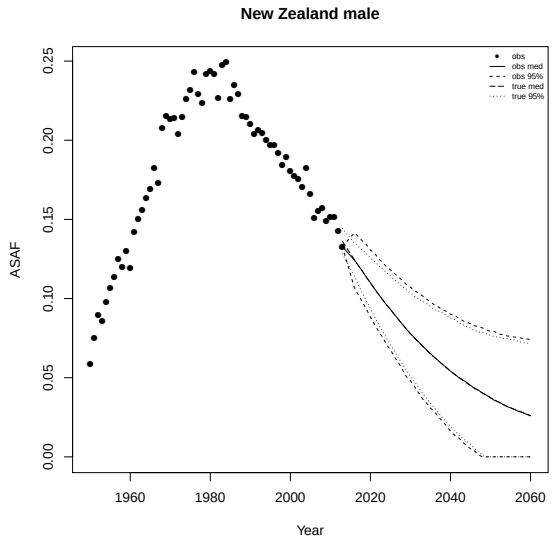

Lithuania female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.01
0.02
0.03
0.04
0.05
0.06
1980
2000
2020
2040
2060
Year

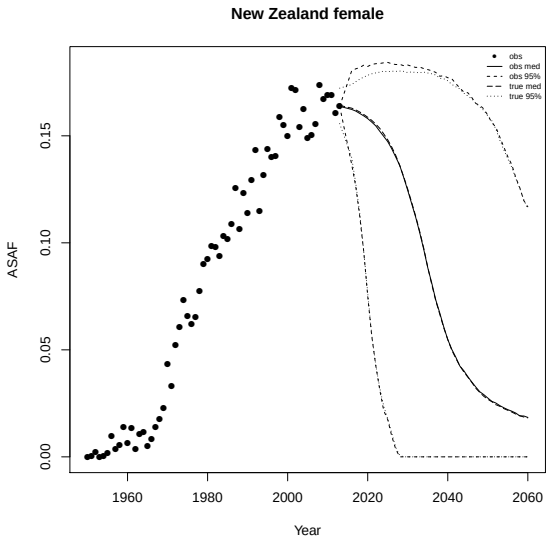

Luxembourg male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

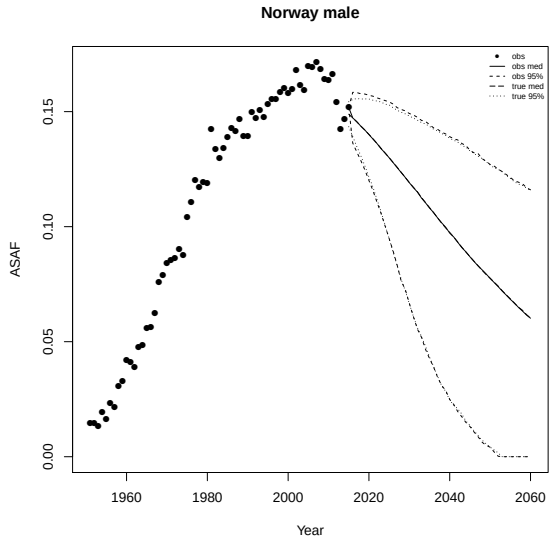

Luxembourg female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

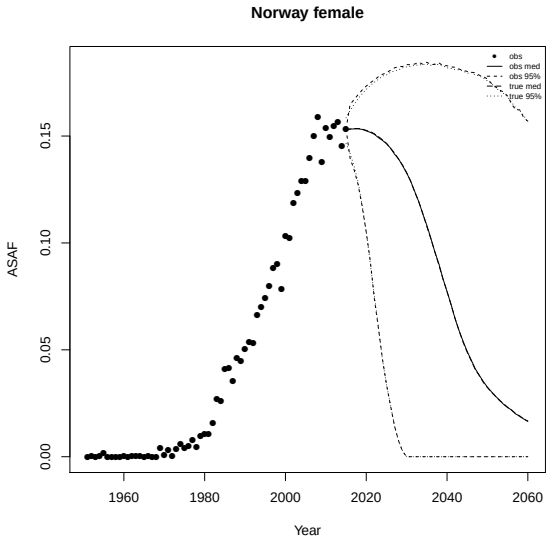

Mauritius male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

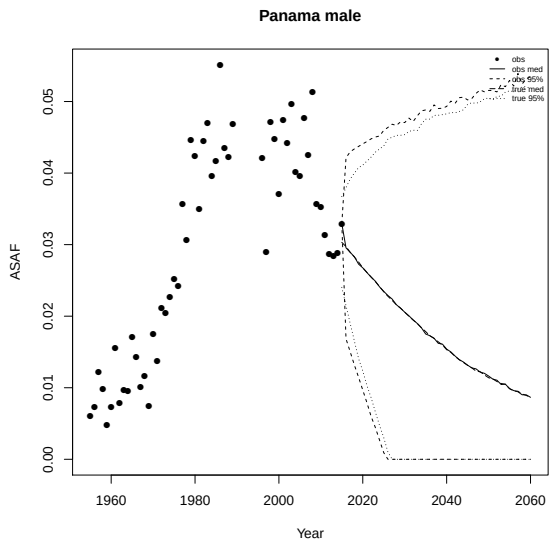

Mauritius female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

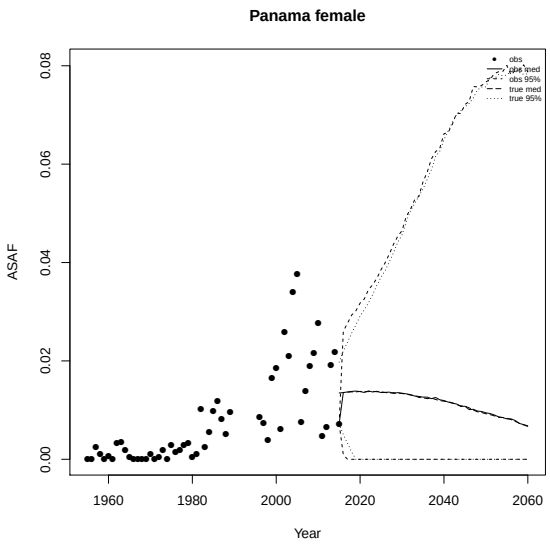

New Zealand male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

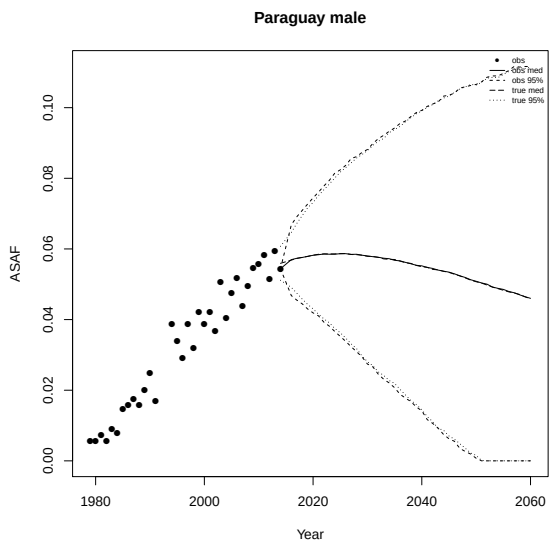

New Zealand female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

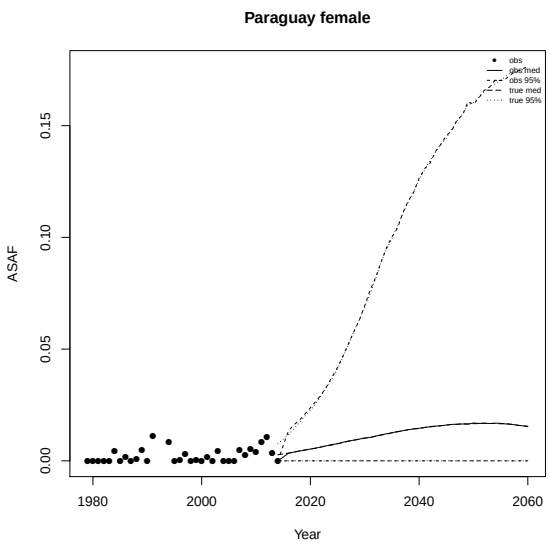

Norway male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

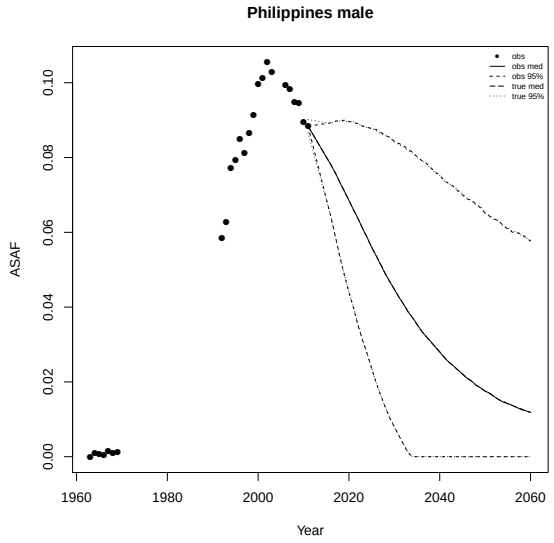

Norway female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

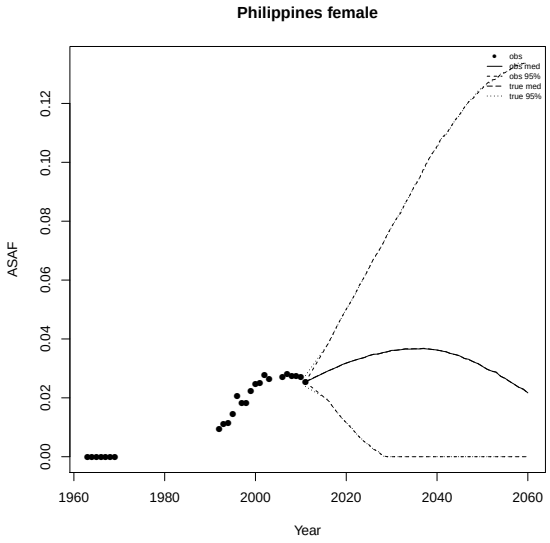

Panama male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

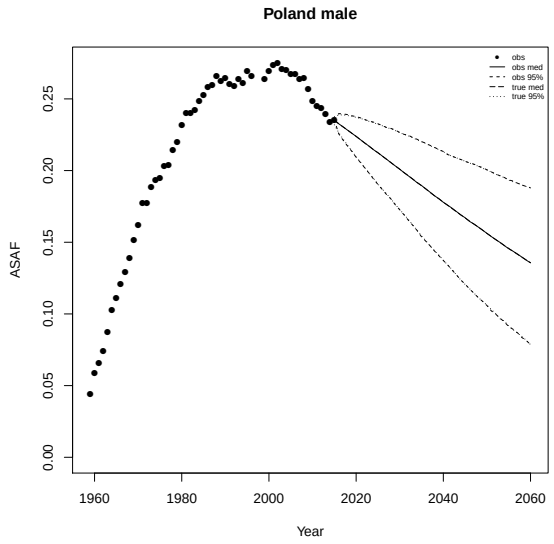

Panama female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

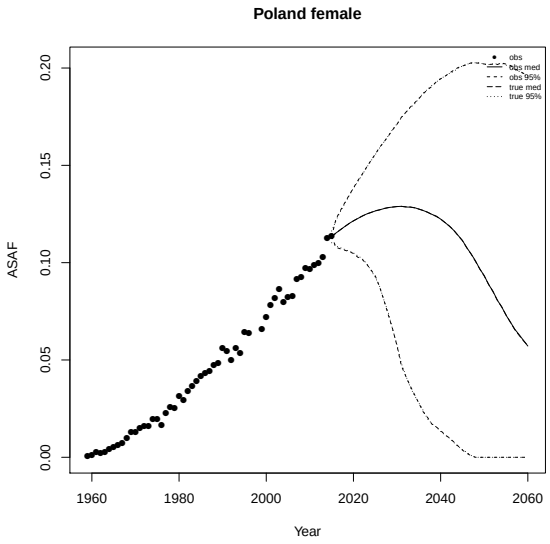

Paraguay male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

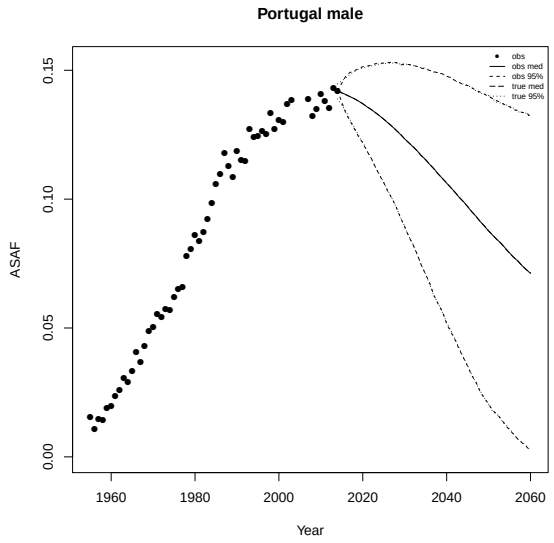

Paraguay female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

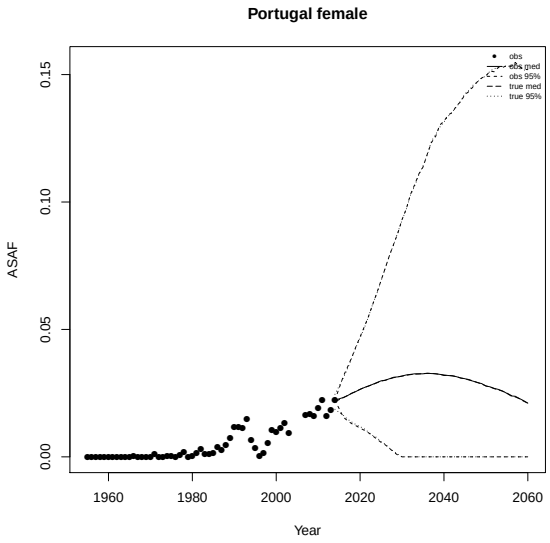

Philippines male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

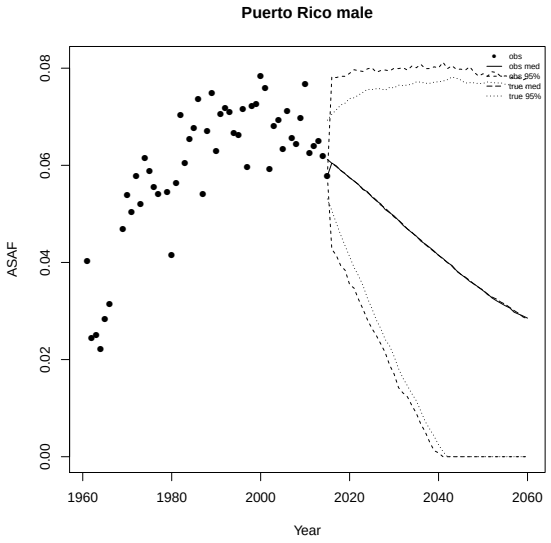

Philippines female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

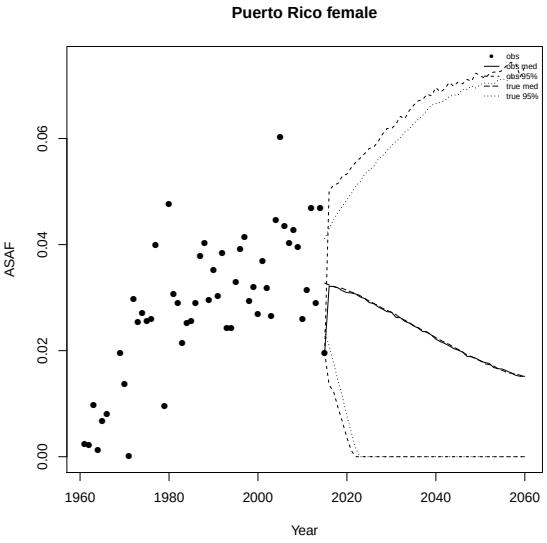

Poland male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

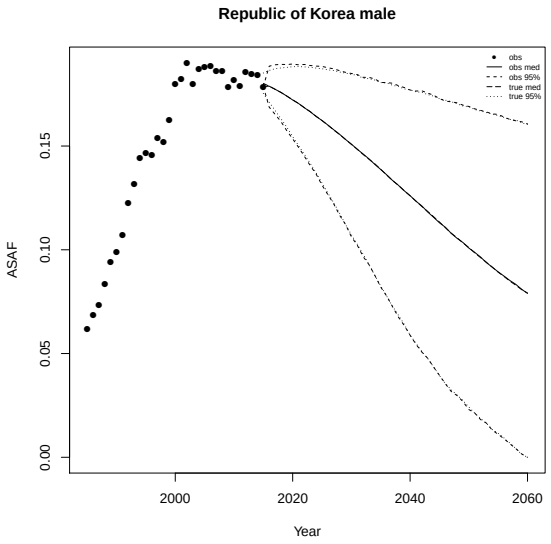

Poland female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

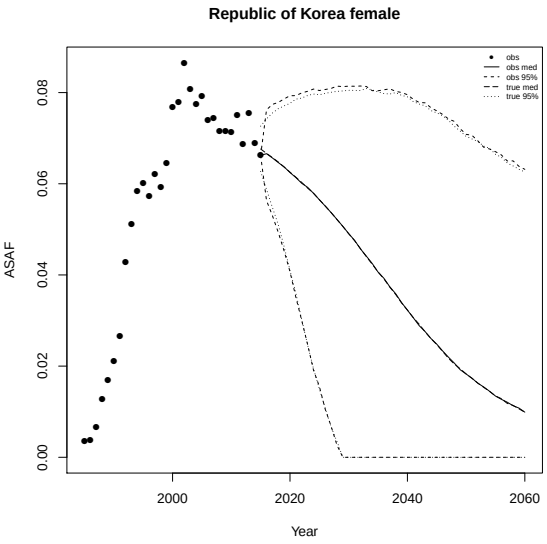

Portugal male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

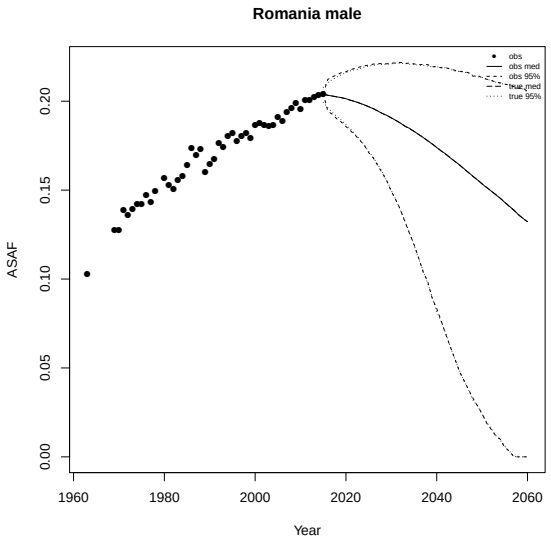

Portugal female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

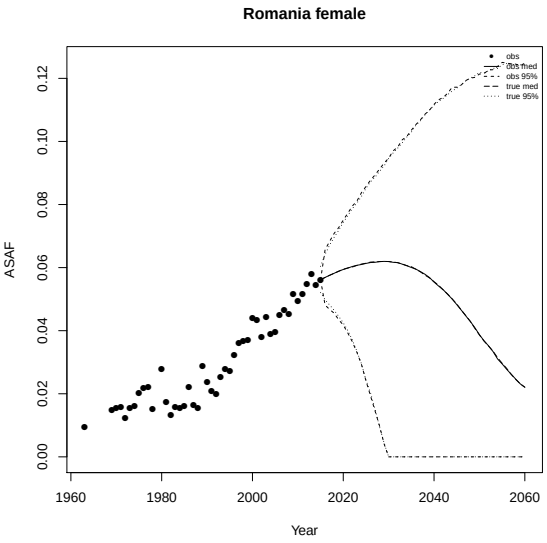

Puerto Rico male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

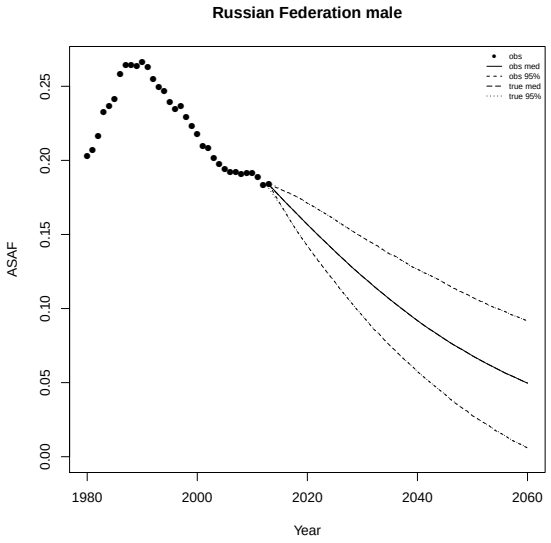

Puerto Rico female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

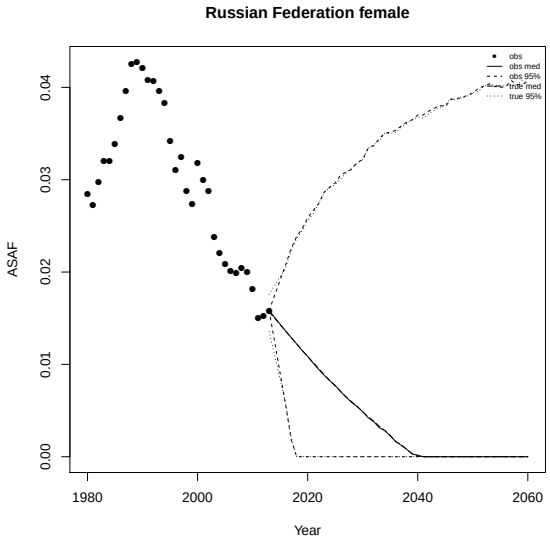

Republic of Korea male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

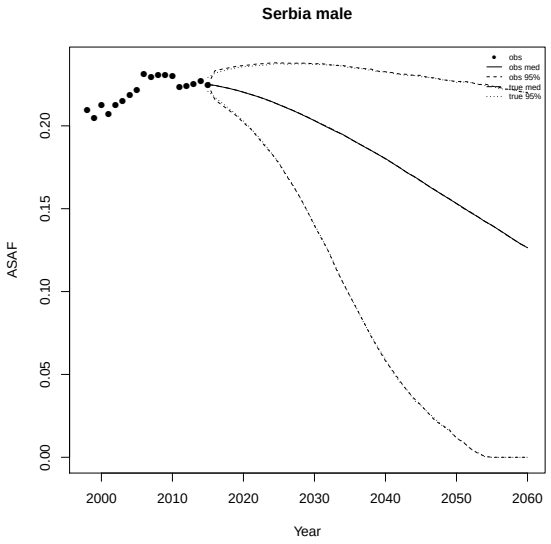

Republic of Korea female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

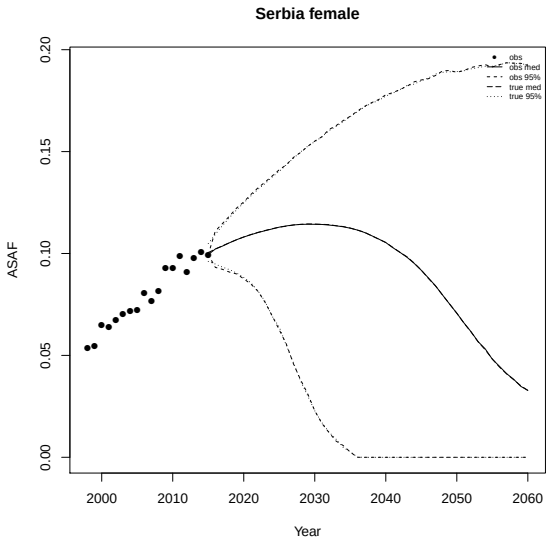

Romania male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

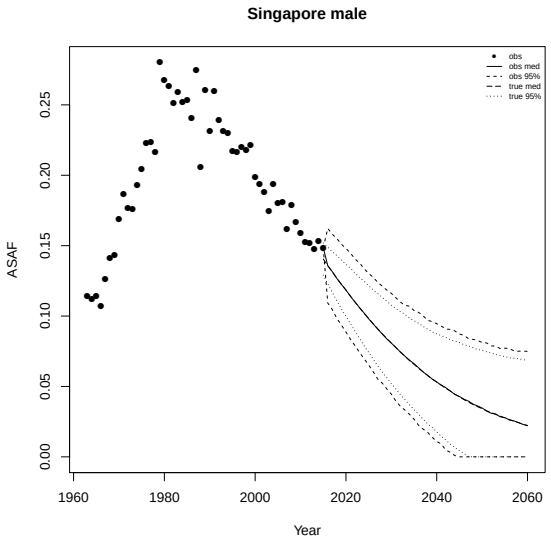

Romania female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

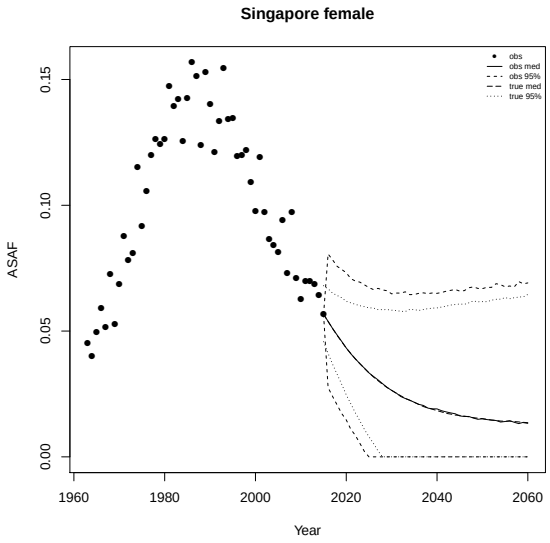

Russian Federation male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
1980
2000
2020
2040
2060
Year

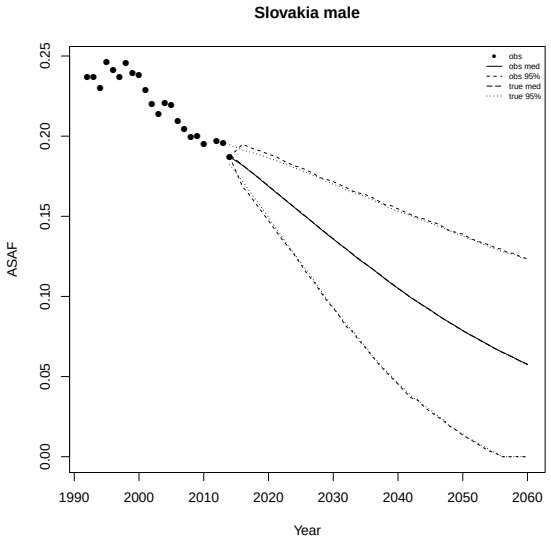

Russian Federation female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.01
0.02
0.03
0.04
1980
2000
2020
2040
2060
Year

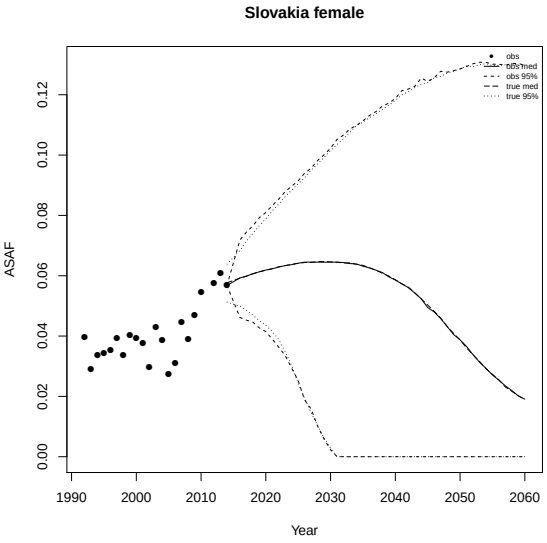

Serbia male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
2000
2010
2020
2030
2040
2050
2060
Year

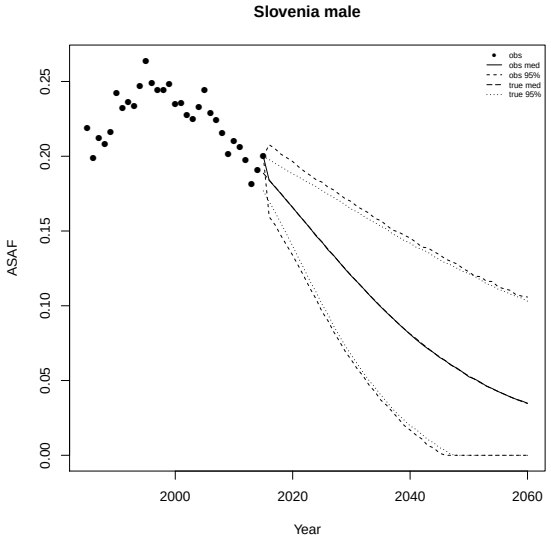

Serbia female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
2000
2010
2020
2030
2040
2050
2060
Year

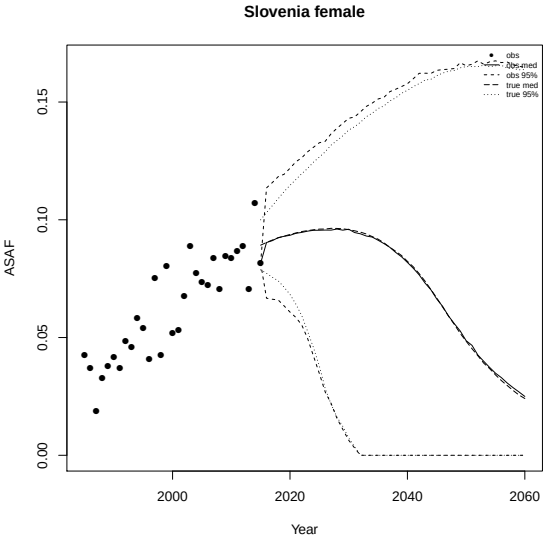

Singapore male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
1960
1980
2000
2020
2040
2060
Year

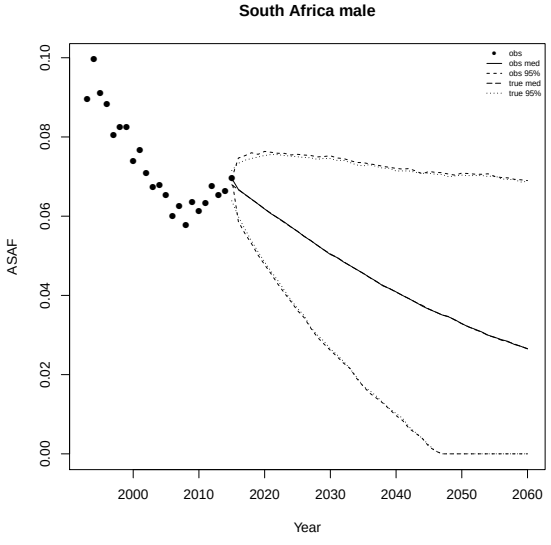

Singapore female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
1960
1980
2000
2020
2040
2060
Year

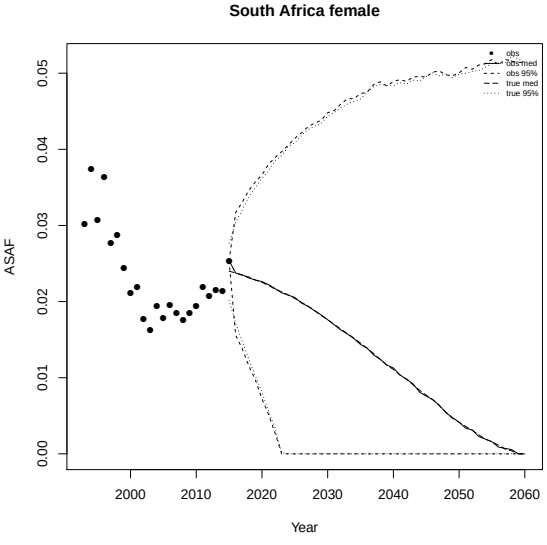

Slovakia male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

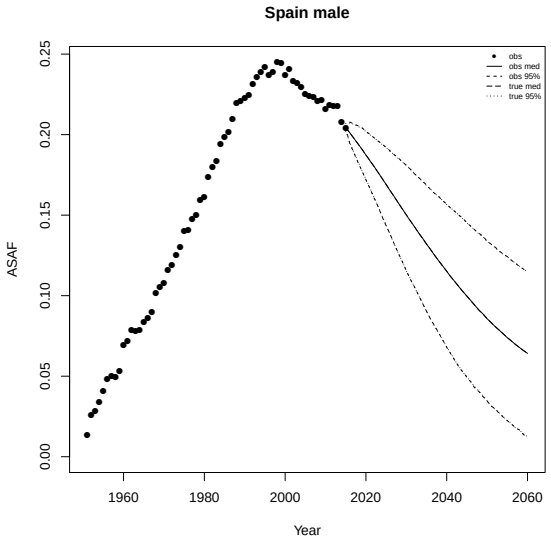

Slovakia female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

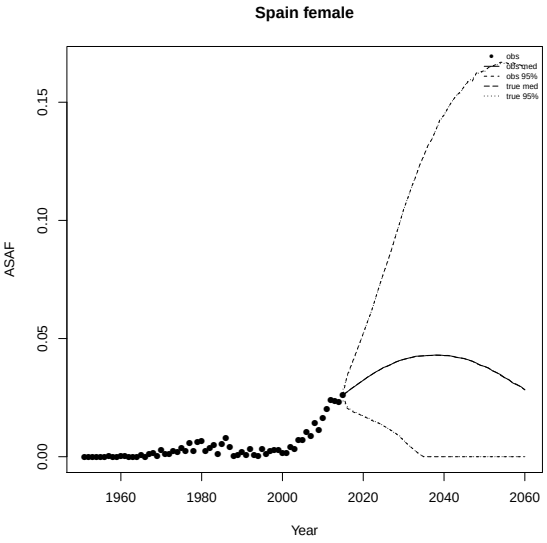

Slovenia male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

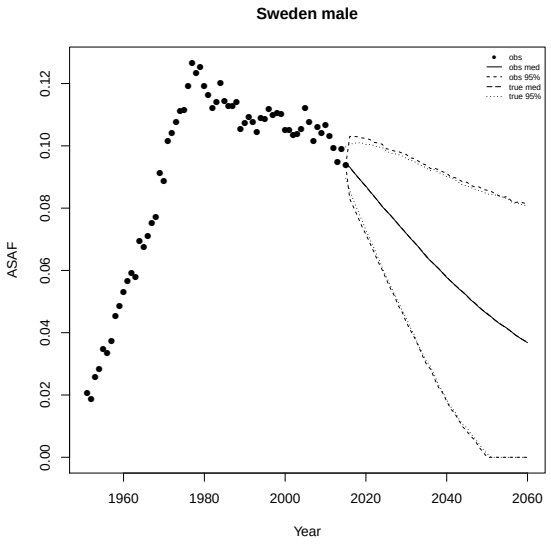

Slovenia female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

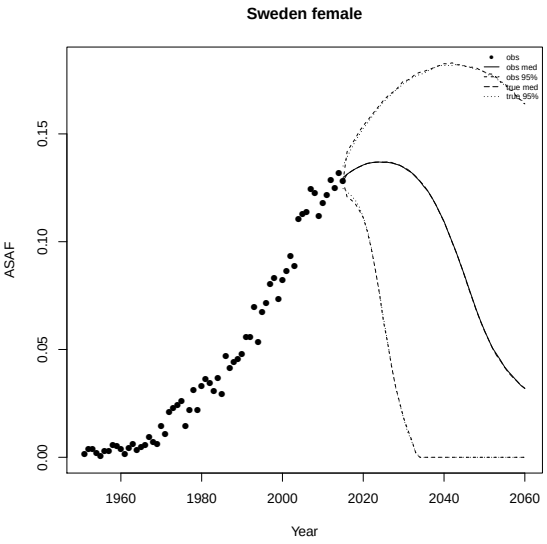

South Africa male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

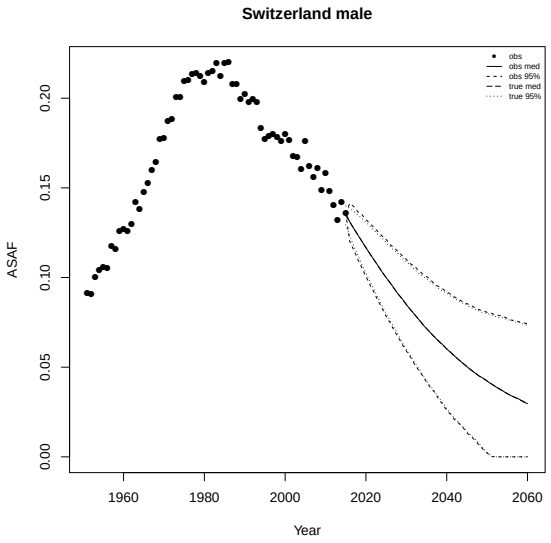

South Africa female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

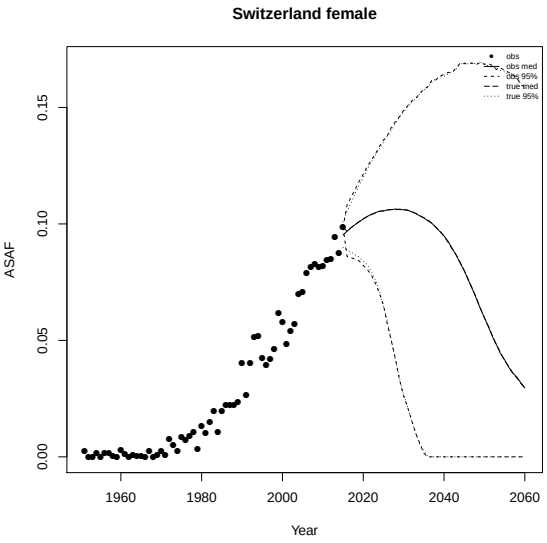

Spain male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

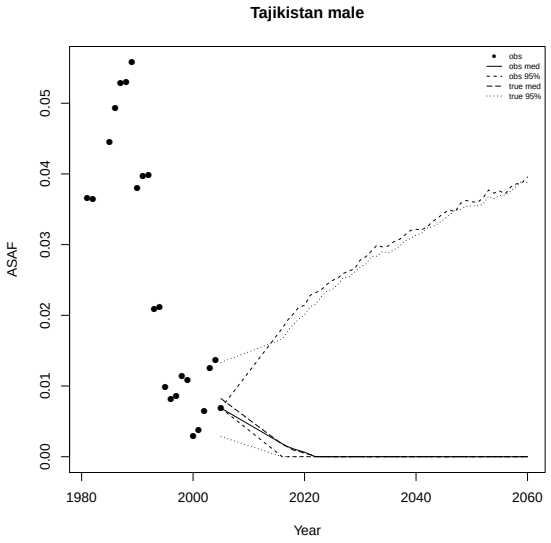

Spain female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

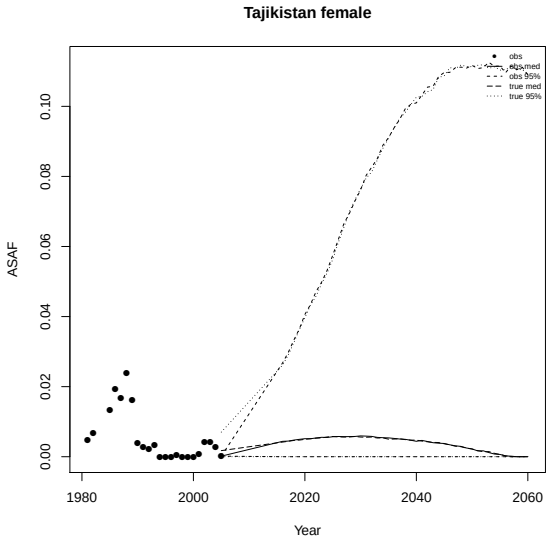

Sweden male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

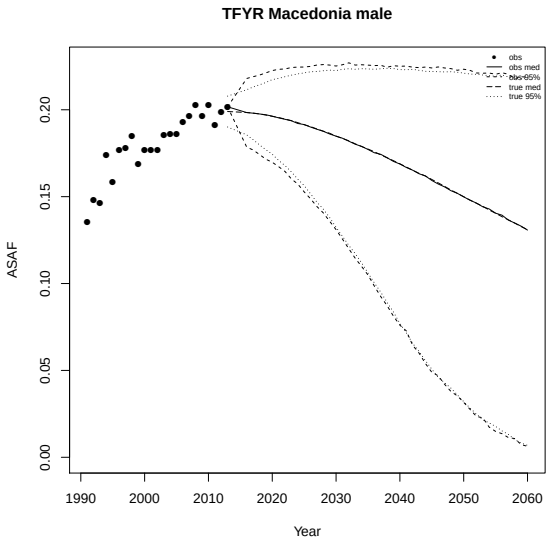

Sweden female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

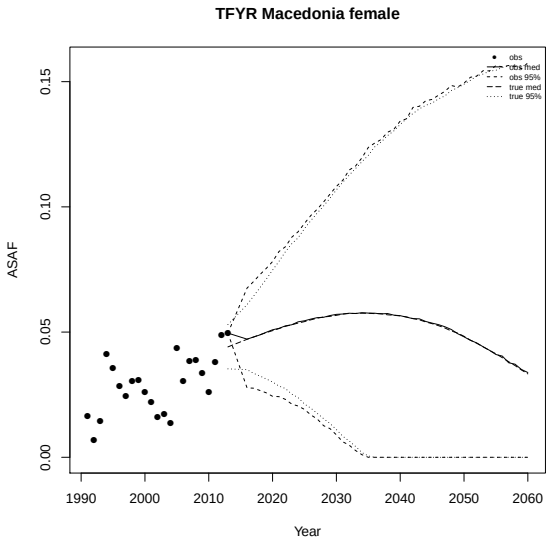

Switzerland male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

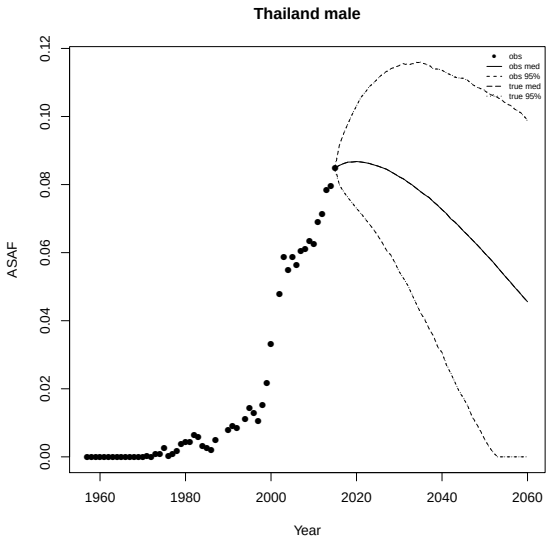

Switzerland female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

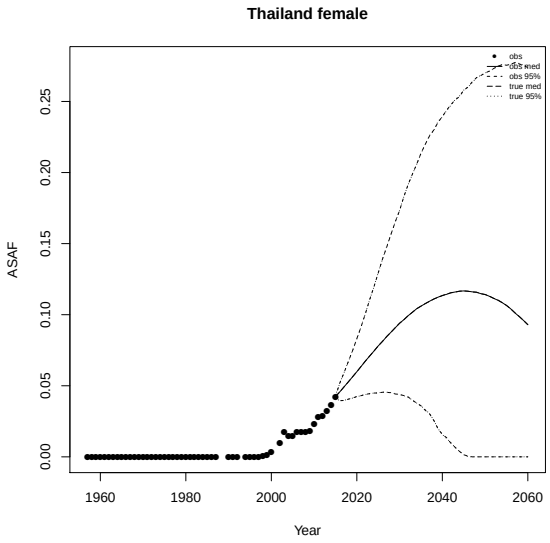

Tajikistan male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

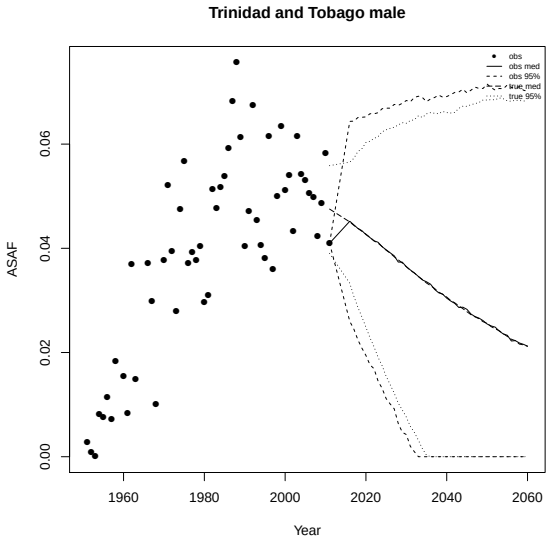

Tajikistan female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

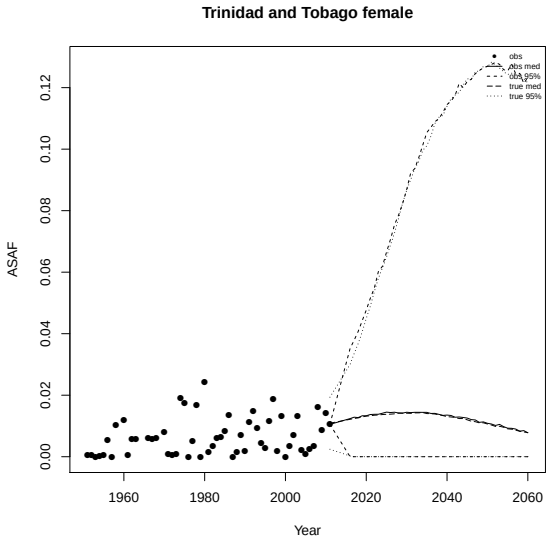

TFYR Macedonia male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

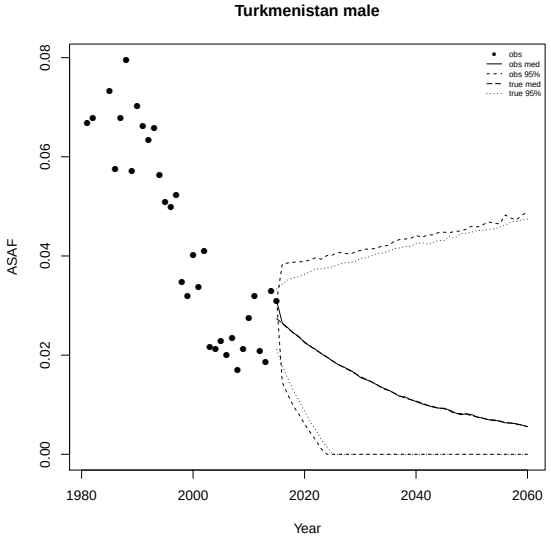

TFYR Macedonia female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

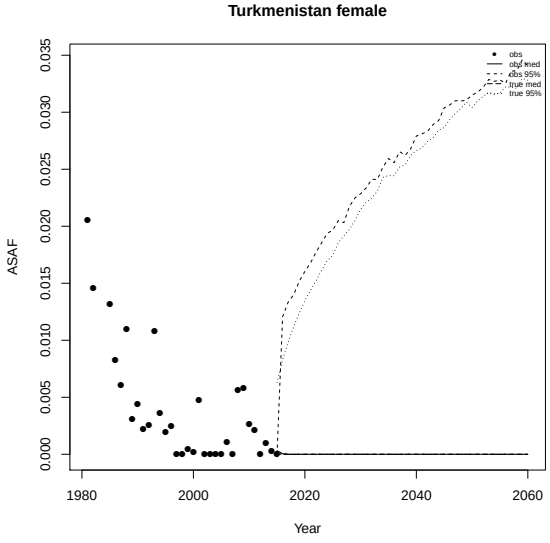

Thailand male
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

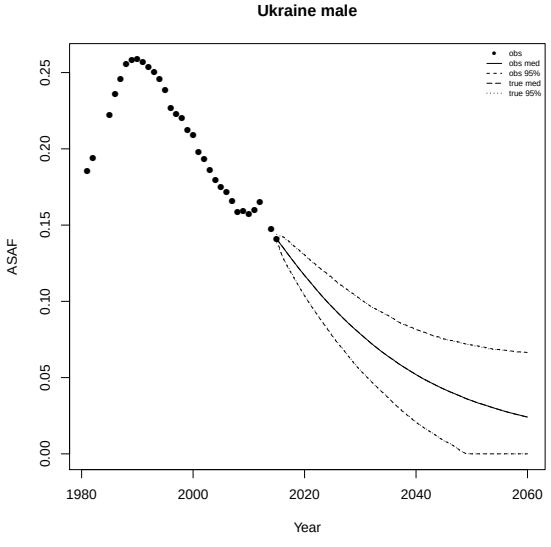

Thailand female
obs
obs med
obs 95%
true med
true 95%
ASAF
Year

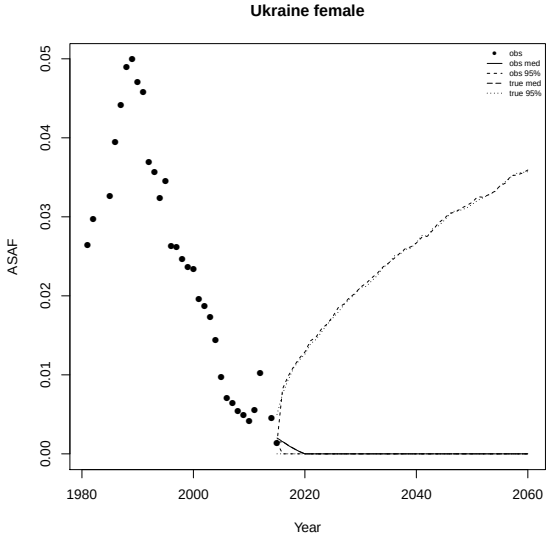

Trinidad and Tobago male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
1960
1980
2000
2020
2040
2060
Year

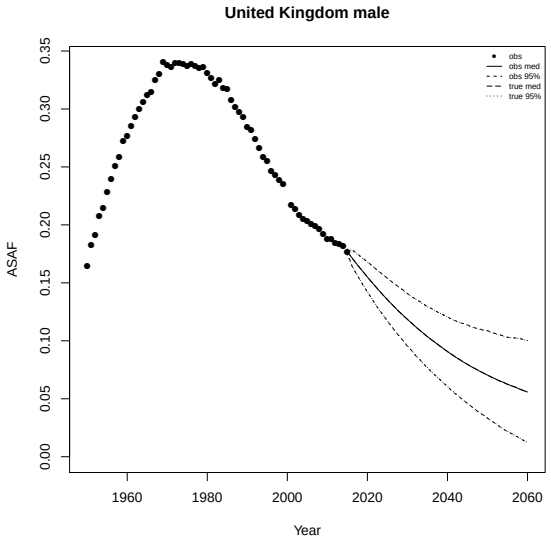

Trinidad and Tobago female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
0.10
0.12
1960
1980
2000
2020
2040
2060
Year

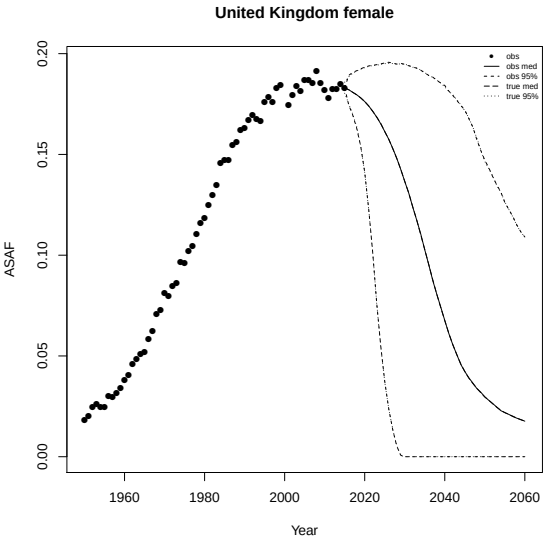

Turkmenistan male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.02
0.04
0.06
0.08
1980
2000
2020
2040
2060
Year

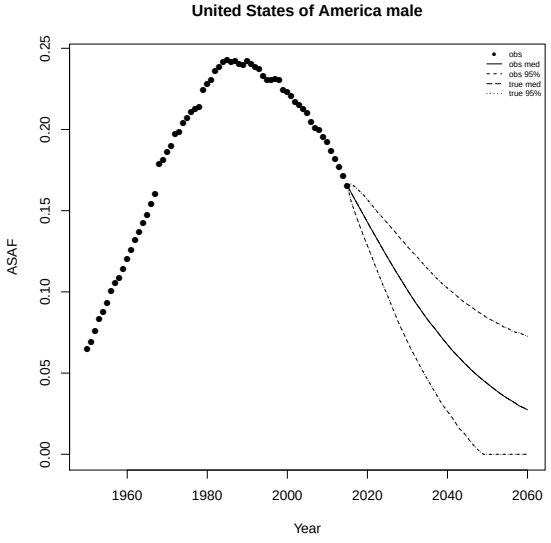

Turkmenistan female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
1980
2000
2020
2040
2060
Year

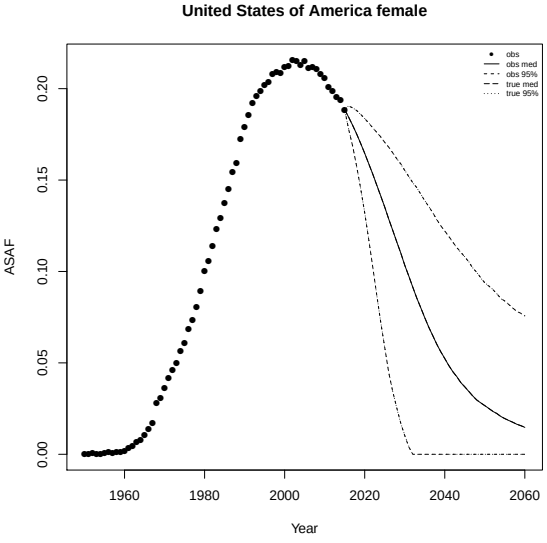

Ukraine male
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.05
0.10
0.15
0.20
0.25
1980
2000
2020
2040
2060
Year

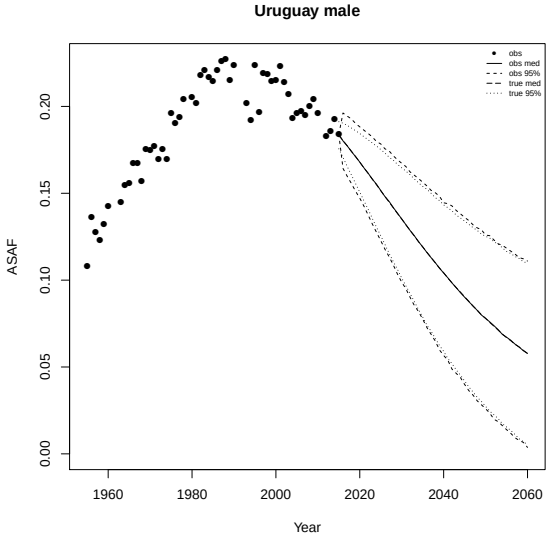

Ukraine female
obs
obs med
obs 95%
true med
true 95%
ASAF
0.00
0.01
0.02
0.03
0.04
0.05
1980
2000
2020
2040
2060
Year

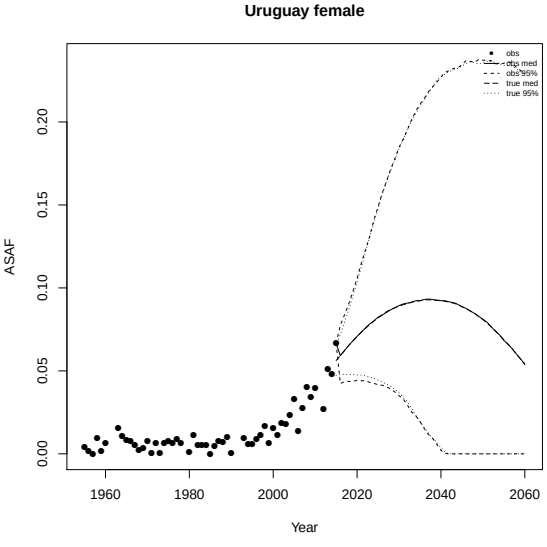

United Kingdom male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

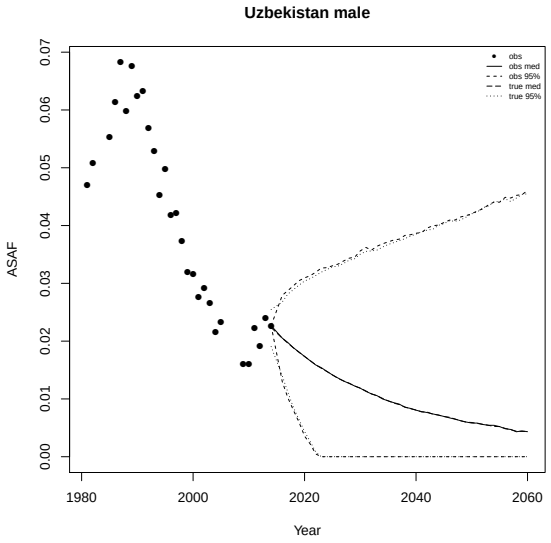

United Kingdom female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

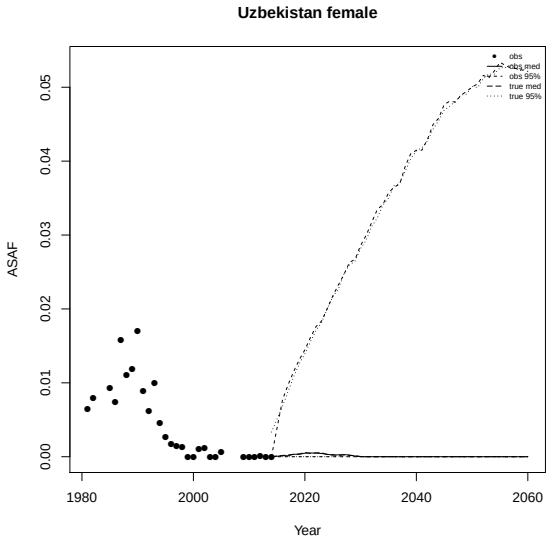

United States of America male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

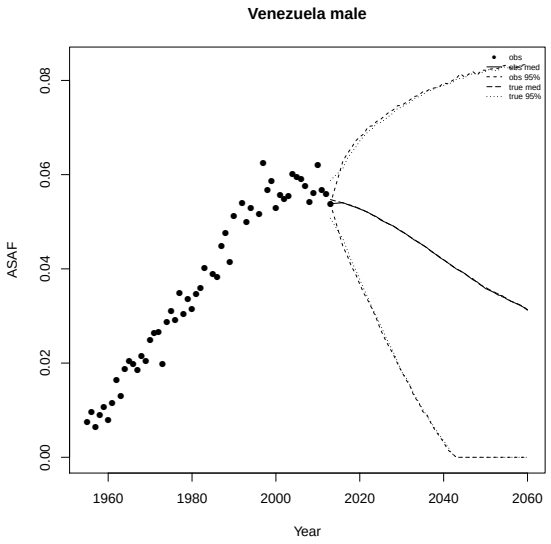

United States of America female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

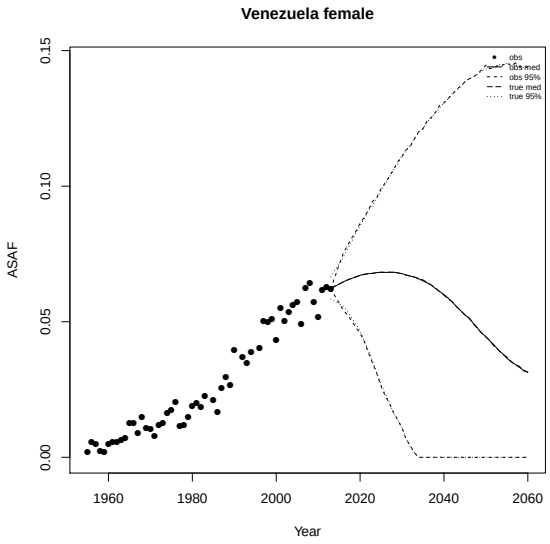

Uruguay male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%


Uruguay female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%


Uzbekistan male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%


Uzbekistan female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%


Venezuela male
ASAF
Year
obs
obs med
obs 95%
true med
true 95%


Venezuela female
ASAF
Year
obs
obs med
obs 95%
true med
true 95%

Appendix B

APPENDICES TO CHAPTER 3

B.1 Full Model Specification

We first describe the estimating and projection of the full model.

1. Estimate and forecast the male ASSAF using the 3-level Bayesian hierarchical model described in Section 3.2.3, and generate 30 samples from the posterior distributions of the mean of ASSAF of over 60 clear-pattern countries for all 13 five-year estimation periods and all 9 five-year periods forecast period;

2. For each country, generate 30 samples of male e_0^{NS} based on the ASSAF samples drawn in Step 2 for all 13 five-year estimation periods, and for each of the 30 samples, forecast male e_0^{NS} of over 60 countries for all 9 five-year periods using the 3-level Bayesian hierarchical model described in Section 3.2.4;

3. For each country, forecast male e_0 based on the method described in Section 3.2.5 for each of the 30 samples, and combine trajectories from all 30 samples to get the full posterior predictive distribution of male e_0;

4. For each country, apply the gap model described in Section 3.2.5 to the combined trajectories of male e_0 to get the full posterior predictive distribution of female e_0.

The details of the Bayesian hierarchical model for modeling age-specific smoking at-

tributable fraction (ASSAF) described in Section 3.2.3 are as follows.

Level 1: $y_{x,t}^{\ell} \overset{\text{ind}}{\sim} \mathcal{N}(\xi_x^{\ell}\tau_{t-x}^{\ell}\mathbf{1}_{x\neq 80} + \xi_x^{\ell}\tilde{\tau}_{t-x}^{\ell}\mathbf{1}_{x=80},\ \sigma_{\ell}^2)$;

Level 2: $\xi_{40}^{\ell} = 1$, $\xi_x^{\ell}|\mu_x^{[\xi]}, \sigma_x^{2[\xi]} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_x^{[\xi]}, \sigma_x^{2[\xi]})$ for all x except 40,

$\tau_c^{\ell}|\theta^{\ell}, \sigma^{2[\tau]} \overset{\text{ind}}{\sim} \mathcal{N}(g(c|\theta^{\ell}),\ \sigma^{2[\tau]})$, $\tilde{\tau}_c^{\ell}|\tilde{\theta}^{\ell}, \sigma^{2[\tau]} \overset{\text{ind}}{\sim} \mathcal{N}(g(c|\tilde{\theta}^{\ell}), \sigma^{2[\tau]})$ for $c = t - x$,

$\triangle_1^{\ell}|\mu_{\triangle_1} \overset{\text{i.i.d}}{\sim} \mathcal{G}(2, 2/\mu_{\triangle_1})$, $\triangle_2^{\ell}|\mu_{\triangle_2}, \sigma_{\triangle_2}^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_{\triangle_2}, \sigma_{\triangle_2}^2)$,

$\triangle_3^{\ell}|\mu_{\triangle_3} \overset{\text{i.i.d}}{\sim} \mathcal{G}(2, 2/\mu_{\triangle_3})$, $\triangle_4^{\ell}|\mu_{\triangle_4}, \sigma_{\triangle_4}^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_{\triangle_4}, \sigma_{\triangle_4}^2)$,

$k^{\ell}|\mu_k, \sigma_k^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_k, \sigma_k^2)$, $\delta^{\ell}|\mu_{\delta}, \sigma_{\delta}^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_{\delta}, \sigma_{\delta}^2)$,

$\sigma_{\ell}^2|\sigma^2 \overset{\text{i.i.d}}{\sim} \mathcal{IG}(2, \sigma^2)$;

Level 3: $\mu_x^{[\xi]} \overset{\text{i.i.d}}{\sim} \mathcal{N}(1, 5)$, $\sigma_x^{2[\xi]} \overset{\text{i.i.d}}{\sim} \mathcal{IG}(2, 5)$,

$\sigma^2 \sim \mathcal{IG}(2, 0.01)$, $\sigma^{2[\tau]} \sim \mathcal{IG}(2, 0.01)$,

$\mu_{\triangle_1} \sim \mathcal{G}(2, 0.1)$, $\mu_{\triangle_2} \sim \mathcal{N}(20, 1000)$,

$\mu_{\triangle_3} \sim \mathcal{G}(2, 0.1)$, $\mu_{\triangle_4} \sim \mathcal{N}(20, 1000)$,

$\mu_k \sim \mathcal{N}(0.3, 0.25)$, $\mu_{\delta} \sim \mathcal{N}(0, 100)$,

$\sigma_{\triangle_2}^2 \sim \mathcal{IG}(2, 1000)$, $\sigma_{\triangle_4}^2 \sim \mathcal{IG}(2, 1000)$,

$\sigma_k^2 \sim \mathcal{IG}(2, 0.25)$, $\sigma_{\delta}^2 \sim \mathcal{IG}(2, 100)$,

where $\theta^{\ell} := (\triangle_1^{\ell}, \triangle_2^{\ell}, \triangle_3^{\ell}, \triangle_4^{\ell}, k^{\ell})$, $\tilde{\theta}^{\ell} := (\triangle_1^{\ell}, \triangle_2^{\ell}, \triangle_3^{\ell}, \triangle_4^{\ell} + \delta^{\ell}, k^{\ell})$, and

$$g(c|\theta) = \frac{k}{1 + \exp\{-\triangle_1(c - 1873 - \triangle_2)\}} - \frac{k}{1 + \exp\{-\triangle_3(c - 1873 - \triangle_2 - \triangle_4)\}}.$$

The details of the Bayesian hierarchical model for modeling non-smoking life expectancy

(e_0^{NS}) described in Section 3.2.4 are as follows.

Level 1: $e_{0,\ell,t}^{NS} \overset{\text{ind}}{\sim} \mathcal{N}(e_{0,\ell,t-1}^{NS} + \tilde{g}(e_{0,\ell,t-1}^{NS}|\zeta^{\ell}),\ (\omega^{\ell} \cdot \phi(e_{0,\ell,t-1}^{NS}))^2)$;

Level 2: $a_i^{\ell}|\mu_{a_i}, \sigma_{a_i}^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}_{[0,100]}(\mu_{a_i}, \sigma_{a_i}^2)$, $i = 1, \cdots, 4$,

$w^{\ell}|\mu_w, \sigma_w^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}_{[0,15]}(\mu_w, \sigma_w^2)$, $z^{\ell}|\mu_z, \sigma_z^2 \overset{\text{i.i.d}}{\sim} \mathcal{N}_{[0,1.15]}(\mu_z, \sigma_z^2)$,

$\omega^{\ell} \overset{\text{i.i.d}}{\sim} \mathcal{U}_{[0,10]}$;

Level 3: $\mu_{a_1} \sim \mathcal{N}(15.77, 15.6^2)$, $\mu_{a_2} \sim \mathcal{N}(40.97, 23.5^2)$,

$\mu_{a_3} \sim \mathcal{N}(0.21, 14.5^2)$, $\mu_{a_4} \sim \mathcal{N}(19.82, 14.7^2)$,

$\mu_w \sim \mathcal{N}(2.93, 3.5^2)$, $\mu_z \sim \mathcal{N}(0.40, 0.6^2)$,

$\sigma_{a_1}^2 \sim \mathcal{IG}(2, 15.6^2)$, $\sigma_{a_2}^2 \sim \mathcal{IG}(2, 14.5^2)$,

$\sigma_{a_3}^2 \sim \mathcal{IG}(2, 14.7^2)$, $\sigma_{a_4}^2 \sim \mathcal{IG}(2, 3.5^2)$,

$\sigma_w^2 \sim \mathcal{IG}(2, 0.6^2)$, $\sigma_z^2 \sim \mathcal{IG}(2, 0.6^2)$,

where $\zeta := (a_1, a_2, a_3, a_4, w, z)$ and

$$\tilde{g}(e_0^{NS}|\zeta) := \frac{w}{1 + \exp\{-\frac{4.4}{a_2}(e_0^{NS} - a_1 - 0.5a_2)\}} + \frac{z - w}{1 + \exp\{-\frac{4.4}{a_4}(e_0^{NS} - \sum_{i=1}^{3} a_i - 0.5a_4)\}}.$$

B.2 MCMC Convergence Diagnostics

First of all, we check the convergence of BHM for ASSAF based on trace plots and Raftery diagnostics Raftery and Lewis (1992) for global parameters in Level 3. We check one chain with 2,000 burnin and 100,000 samples with 20 thinning period. Table B.1 shows the summarizing statistics of the diagnostics. Fig. B.1 shows the trace plots of all 3,000 samples of global parameters. Second, we check the convergence of BHM for e_0^{NS} based on trace plots and Raftery diagnostics Raftery and Lewis (1992) for global parameters in Level 3. We check one of the 30 samples with 1,000 burnin and 100,000 samples with 50 thinning period. Table B.2 shows the summarizing statistics of the diagnostics. Fig. B.2 shows the trace plots of all 1,000 samples of global parameters.

Table B.1: Diagnostic statistics for global parameters in BHM for ASSAF. Burn1, Size1, and DF1 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters q = 0.025, r = 0.0125, s = 0:95. Burn2, Size2, and DF2 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters q = 0.975, r = 0.0125, s = 0.95.

Parameters	Burn1	Size1	DF1	Burn2	Size2	DF2
$\mu_{40}^{2[\beta]}$	-	-	-	-	-	-
$\mu_{45}^{2[\beta]}$	2	606	1.01	2	631	1.05
$\mu_{50}^{2[\beta]}$	2	641	1.07	2	581	0.97
$\mu_{55}^{2[\beta]}$	2	577	0.96	2	631	1.05
$\mu_{60}^{2[\beta]}$	2	616	1.03	2	591	0.98
$\mu_{65}^{2[\beta]}$	2	601	1.00	2	621	1.03
$\mu_{70}^{2[\beta]}$	2	616	1.03	2	621	1.03
$\mu_{75}^{2[\beta]}$	2	587	0.98	2	611	1.02
$\mu_{80}^{2[\beta]}$	2	626	1.04	3	664	1.11

Continued on next page

Table B.1 – *Continued from previous page*

Parameters	Burn1	Size1	DF1	Burn2	Size2	DF2
$\sigma_{40}^{2[\beta]}$	-	-	-	-	-	-
$\sigma_{45}^{2[\beta]}$	3	669	1.12	3	653	1.09
$\sigma_{50}^{2[\beta]}$	2	601	1.00	2	601	1.00
$\sigma_{55}^{2[\beta]}$	2	591	0.98	2	621	1.03
$\sigma_{60}^{2[\beta]}$	2	606	1.01	2	572	0.95
$\sigma_{65}^{2[\beta]}$	2	611	1.02	2	611	1.02
$\sigma_{70}^{2[\beta]}$	1	595	0.99	2	591	0.98
$\sigma_{75}^{2[\beta]}$	3	648	1.08	2	641	1.07
$\sigma_{80}^{2[\beta]}$	2	621	1.03	3	676	1.13
σ	2	591	0.98	3	658	1.10
$\sigma^{2[\tau]}$	2	591	0.98	3	648	1.08
$\mu_{\triangle_1}$	6	1308	2.18	9	2040	3.40
$\mu_{\triangle_2}$	8	1610	2.68	2	641	1.07
$\sigma_{\triangle_2}^2$	12	2160	3.60	8	1586	2.64
$\mu_{\triangle_3}$	4	790	1.32	3	686	1.14
$\mu_{\triangle_4}$	12	1980	3.30	15	2211	3.68
$\sigma_{\triangle_4}^2$	6	1701	2.84	8	1656	2.76
μ_k	6	1432	2.39	8	1456	2.43
σ_k^2	4	1236	2.06	4	771	1.28
μ_δ	18	3090	5.15	24	4912	8.19
σ_δ^2	15	2499	4.16	30	6955	11.60

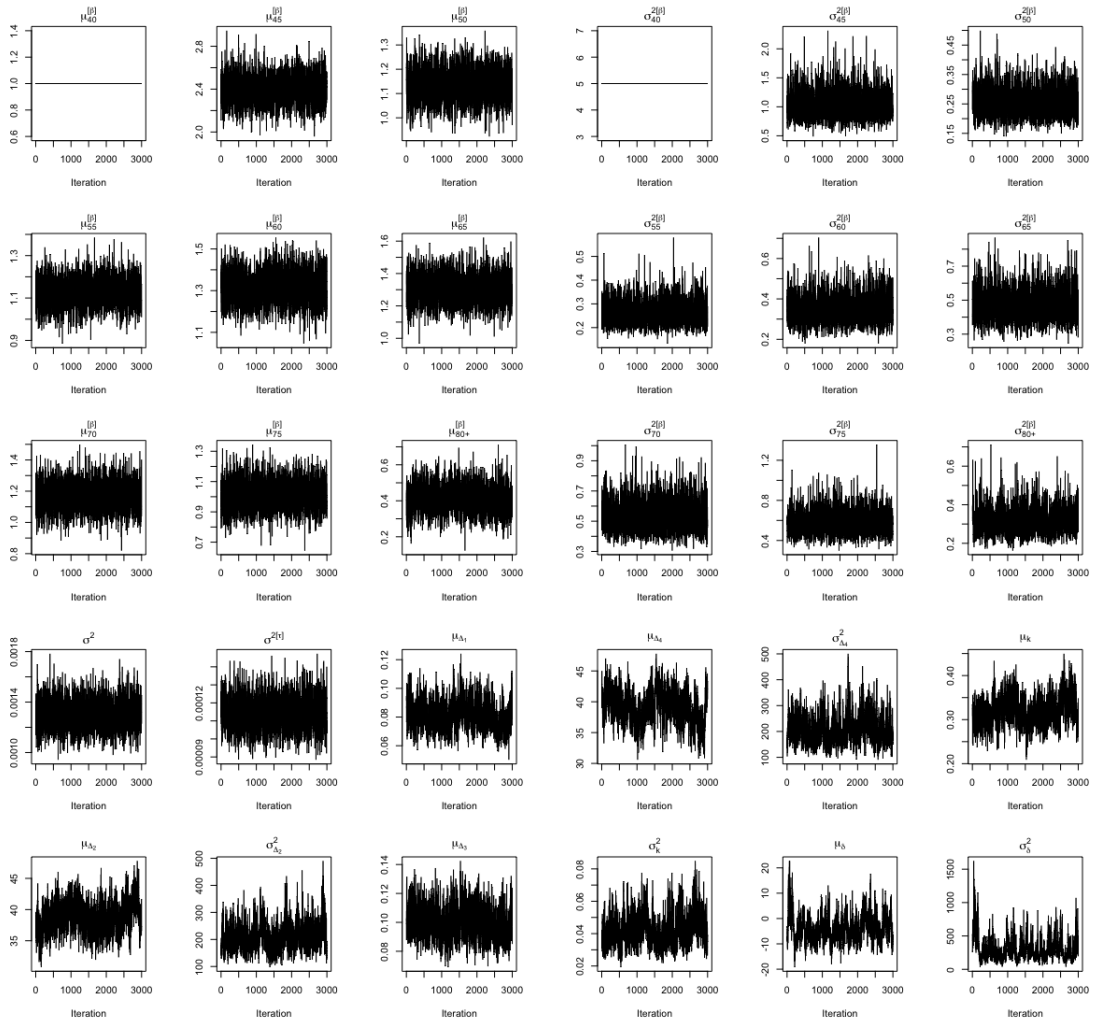


Figure B.1: Traceplots for the hyperparameters in BHM for ASSAF.

Table B.2: Diagnostic statistics for global parameters in BHM for e_0^{NS}. Burn1, Size1, and DF1 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters q = 0.025, r = 0.0125, s = 0:95. Burn2, Size2, and DF2 are the length of burn-in, requied sample size, and dependent factor of Raftery diagnostics with parameters q = 0.975, r = 0.0125, s = 0.95.

Parameters	Burn1	Size1	DF1	Burn2	Size2	DF2
μ_{a_1}	164	14734	24.60	3	703	1.17
$\sigma^2_{a_1}$	2	596	0.99	3	648	1.08
μ_{a_2}	2	572	0.95	81	10795	18.00
$\sigma^2_{a_2}$	3	648	1.08	3	648	1.08
μ_{a_3}	2	572	0.95	7	1179	1.96
$\sigma^2_{a_3}$	2	596	0.99	2	621	1.03
μ_{a_4}	3	703	1.17	6	1005	1.68
$\sigma^2_{a_4}$	3	675	1.12	5	867	1.44
μ_w	3	648	1.08	2	596	0.99
σ^2_w	2	572	0.95	2	596	0.99
μ_z	3	662	1.10	2	572	0.95
σ^2_z	2	596	0.99	3	689	1.15

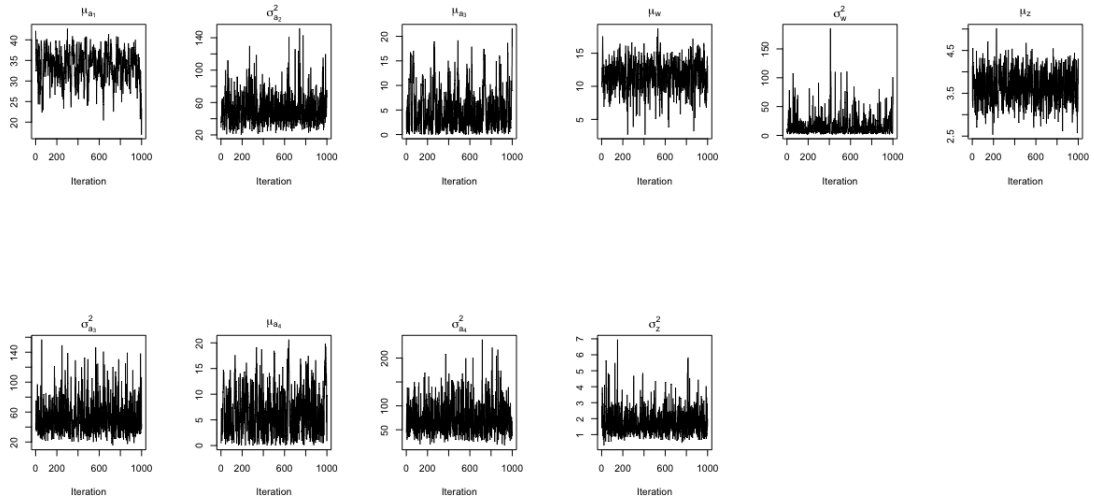
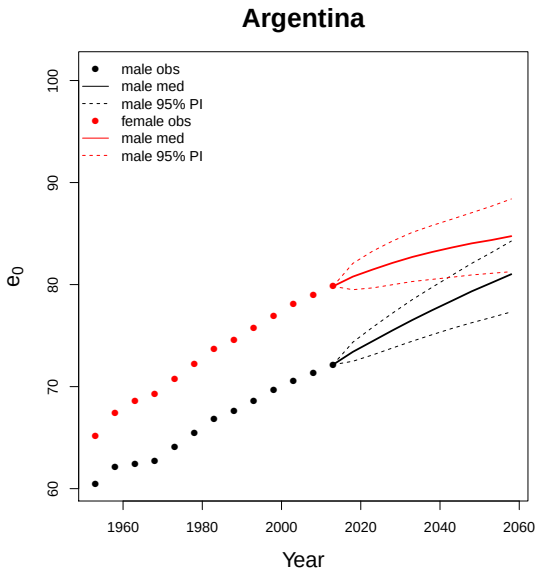
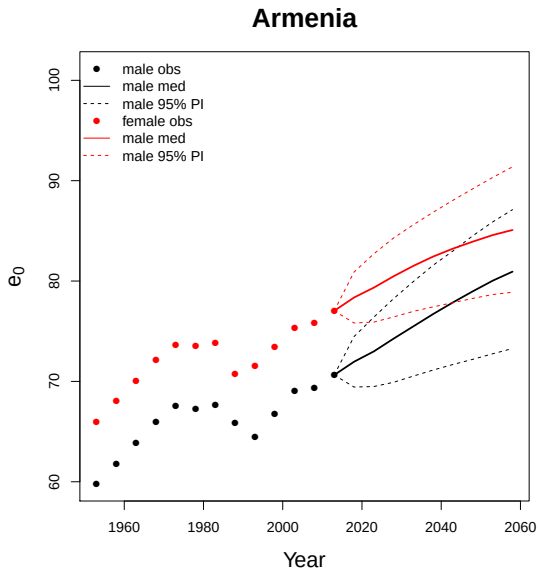
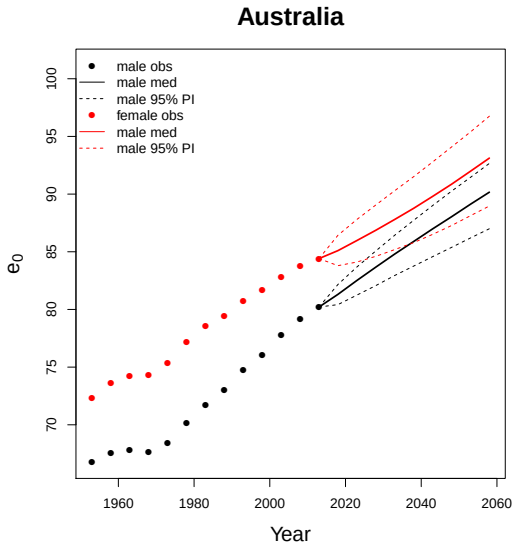


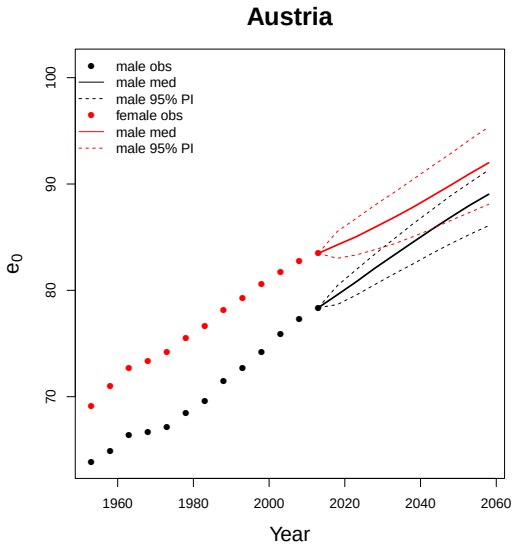
Figure B.2: Traceplots for the hyperparameters in BHM for e_0^{NS}.

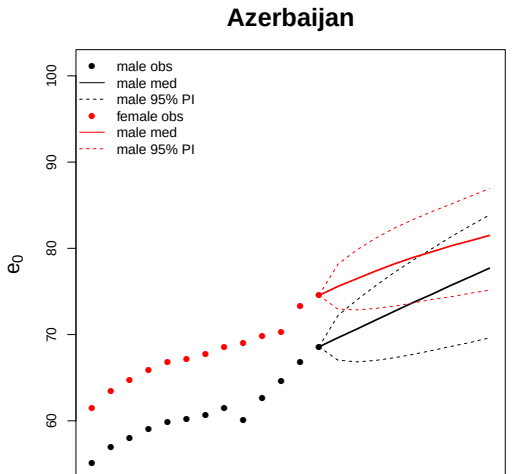
B.3 Life Expectancy at Birth Projection to 2060 for over 60 countries

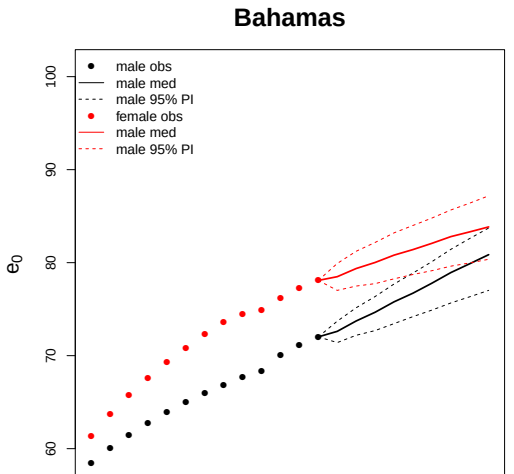


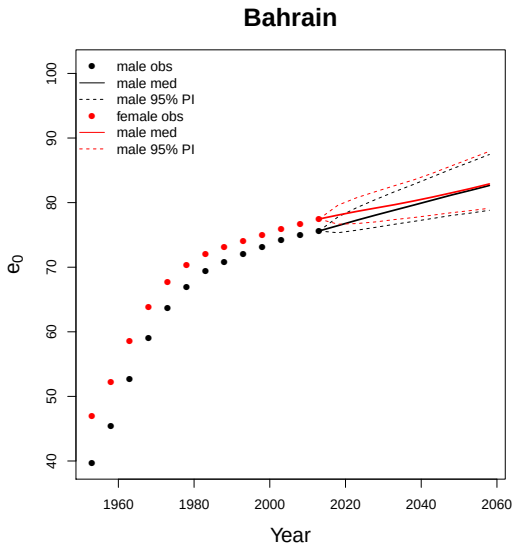


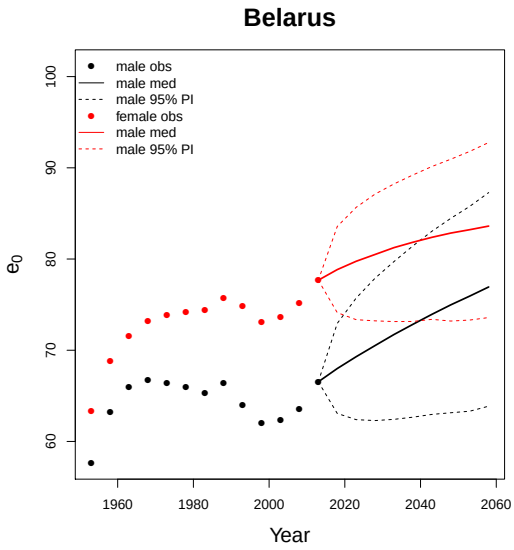


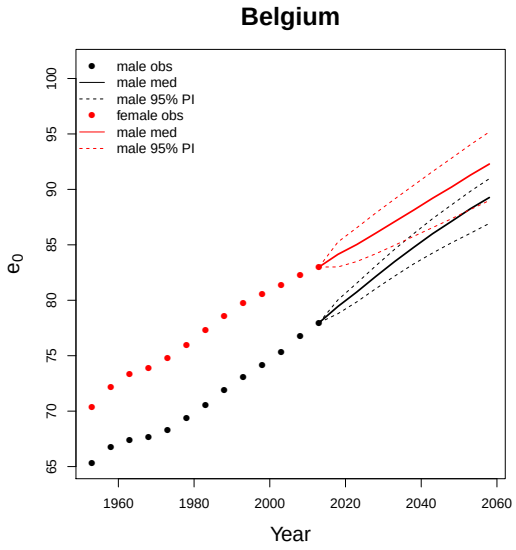


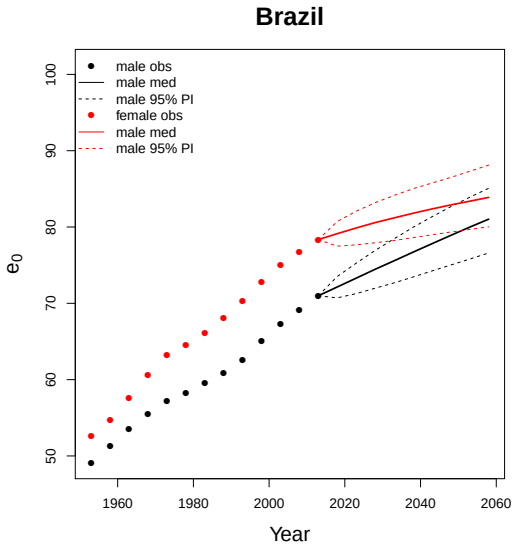


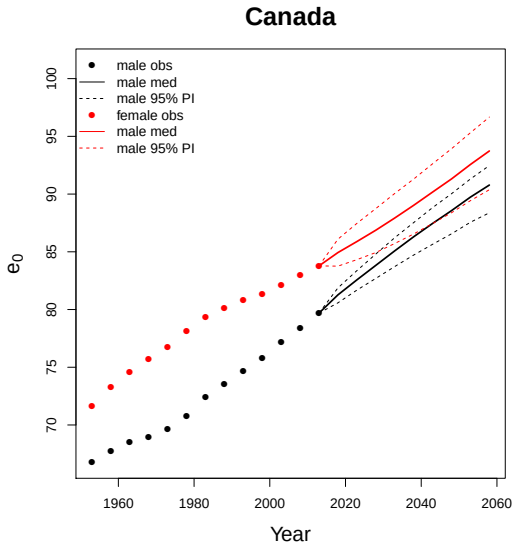


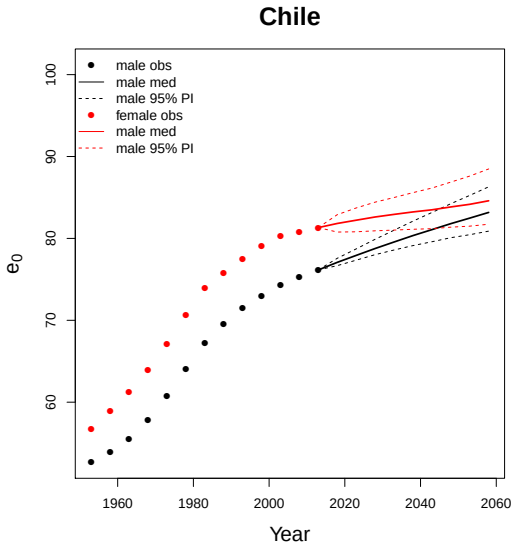

Bahrain
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

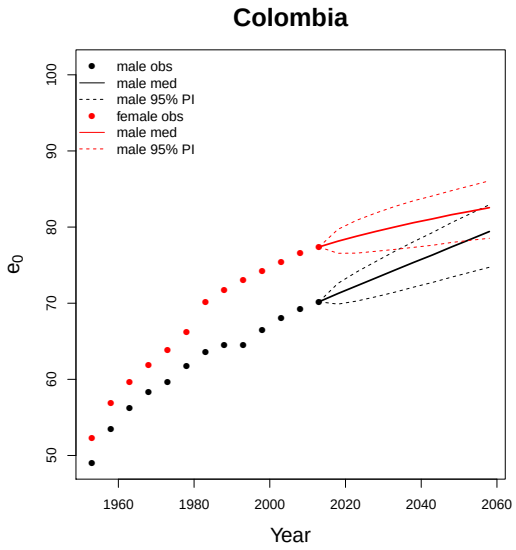

Belarus
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

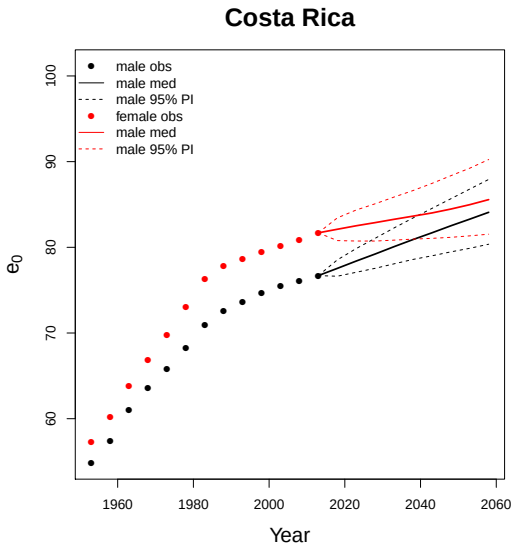

Belgium
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

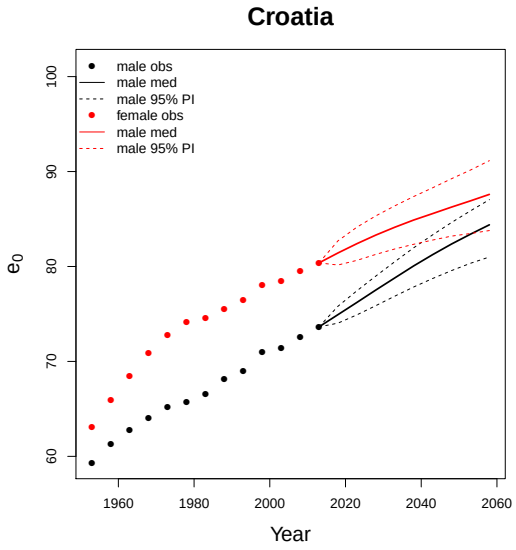

Brazil
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

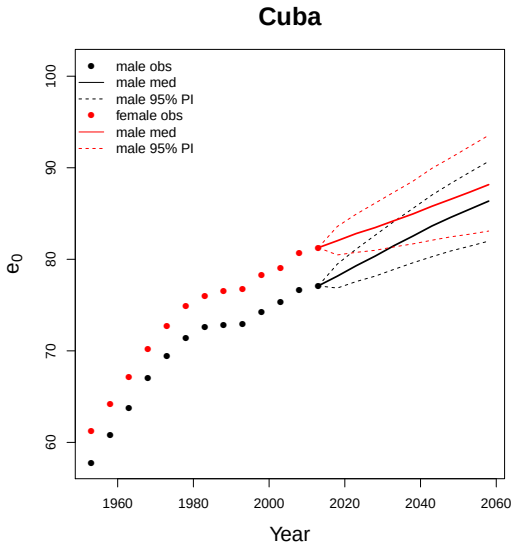

Canada
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

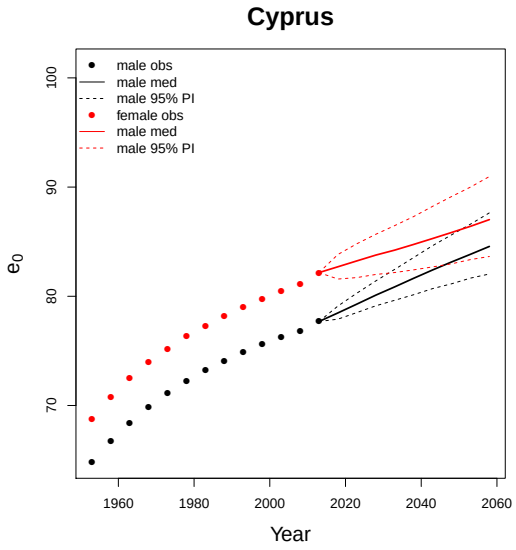

Chile
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

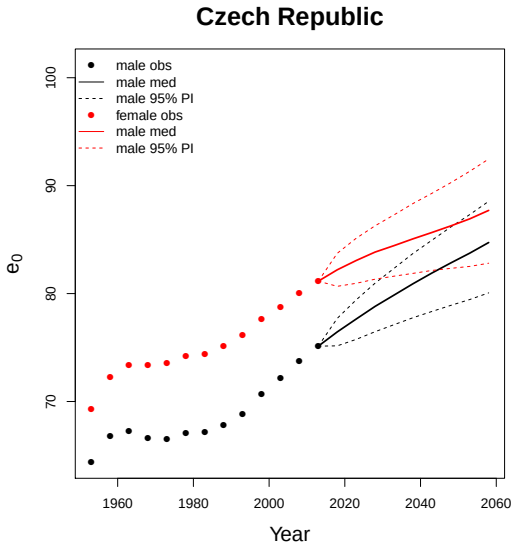

Colombia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

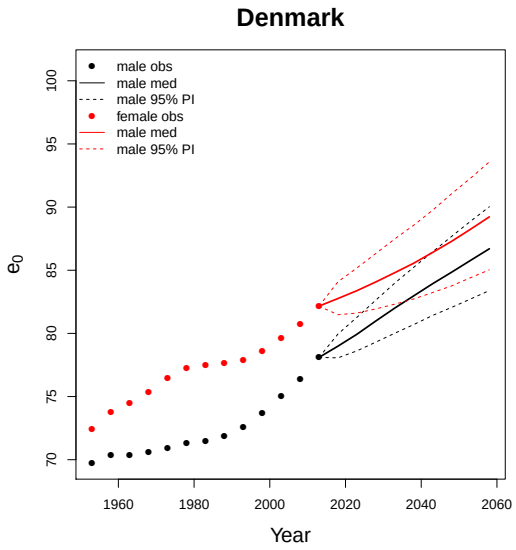

Costa Rica
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

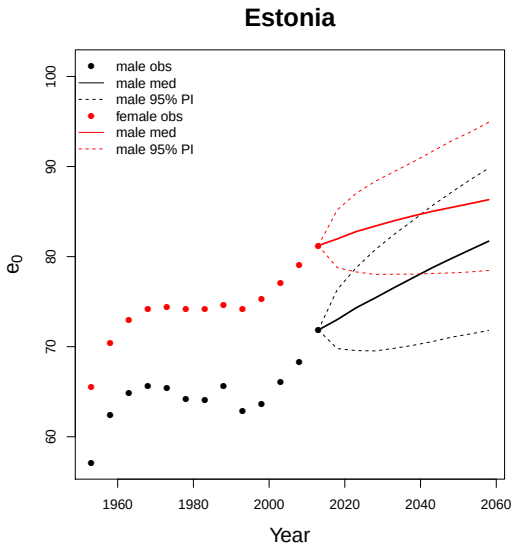

Croatia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

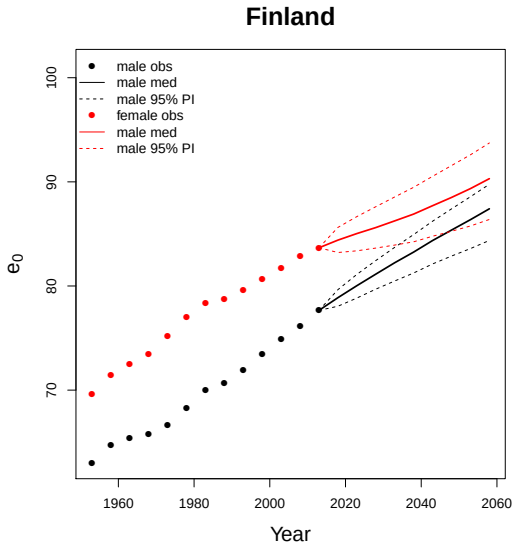

Cuba
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

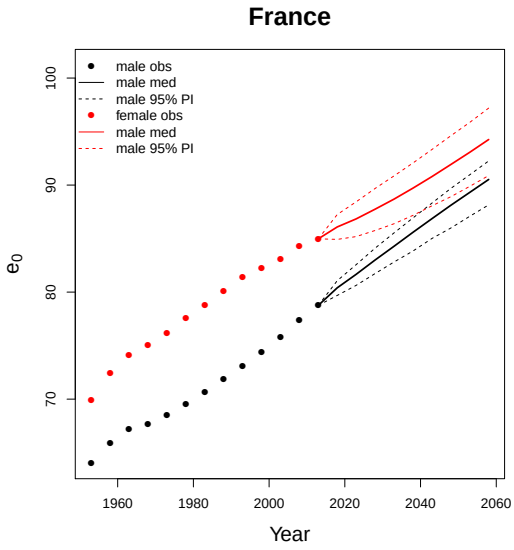

Cyprus
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

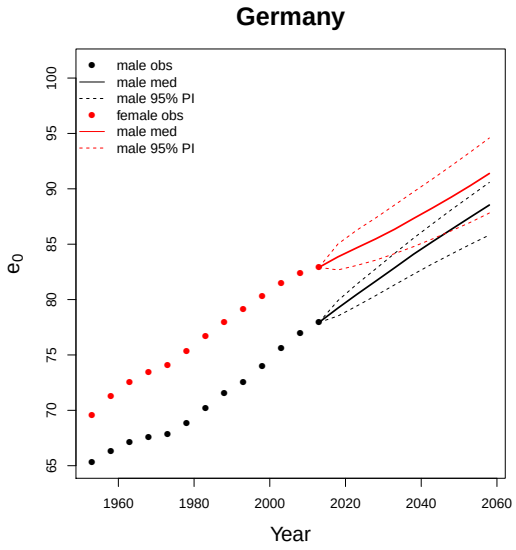

Czech Republic
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e_0
Year

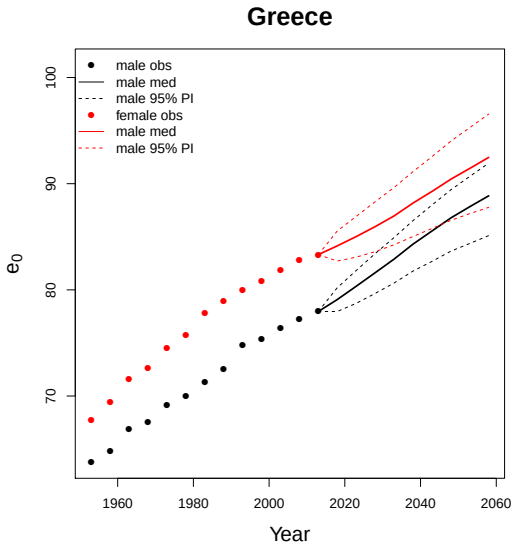

Denmark
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

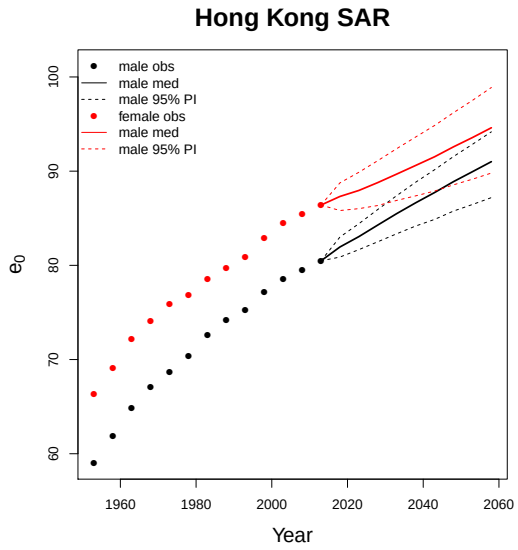

Estonia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

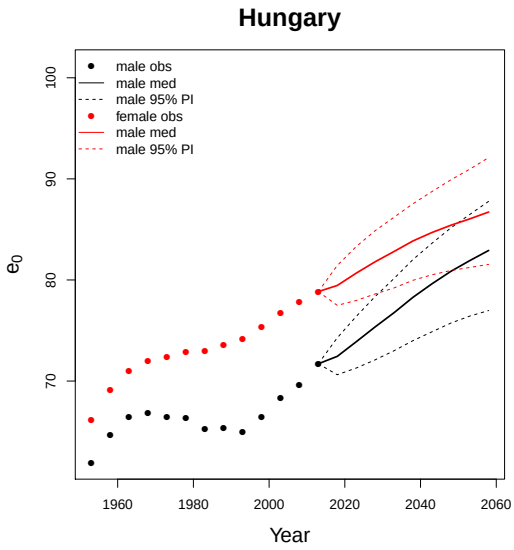

Finland
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

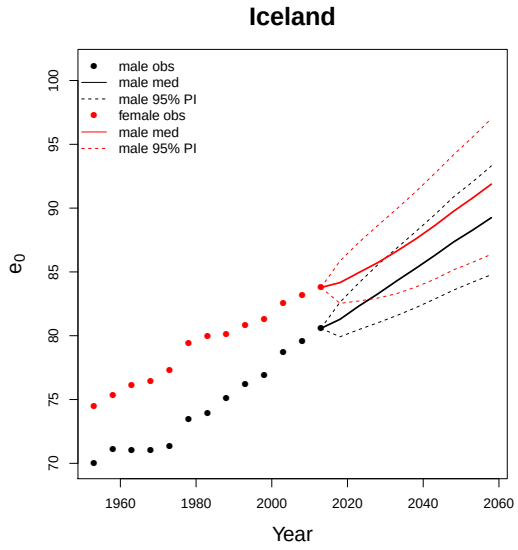

France
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

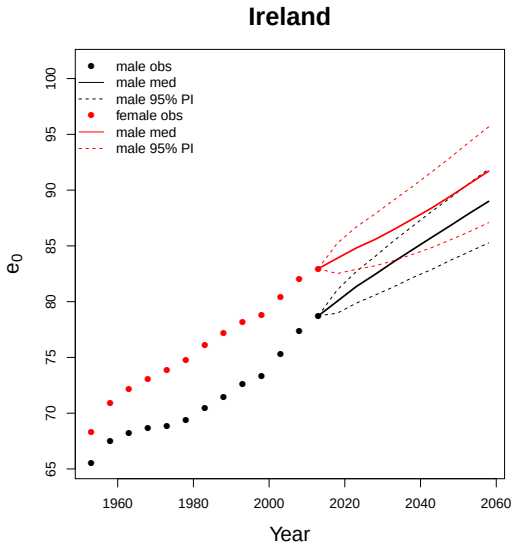

Germany
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

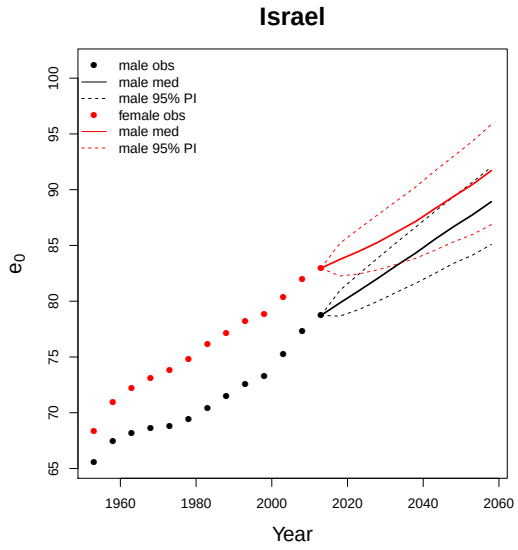

Greece
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

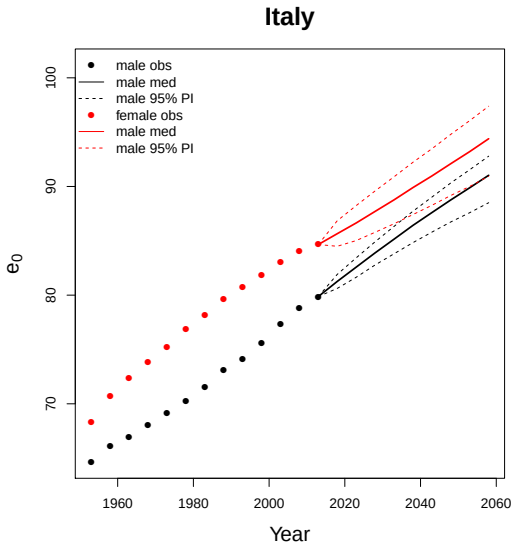

Hong Kong SAR
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

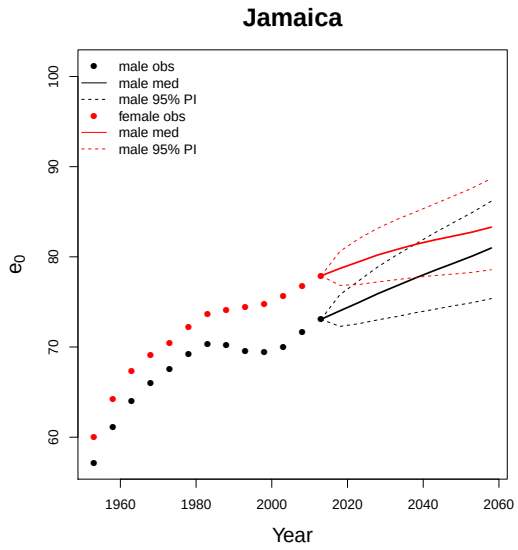

Hungary
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

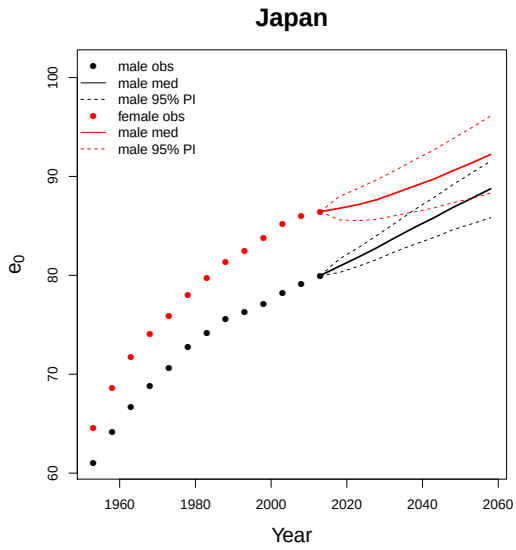

Iceland
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

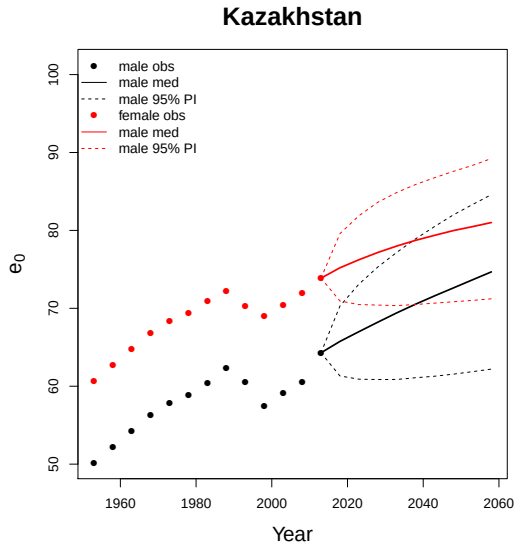

Ireland
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

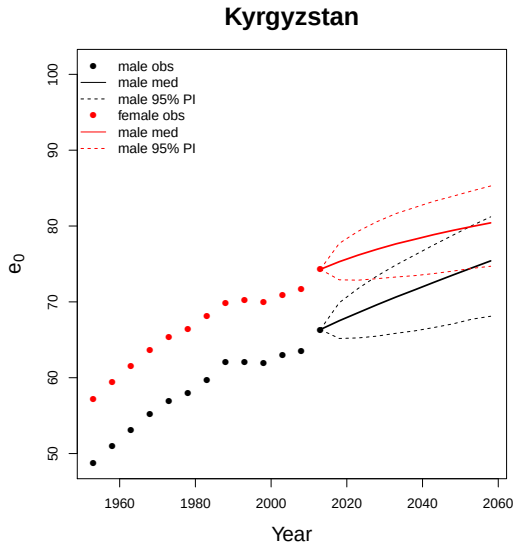

Israel
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

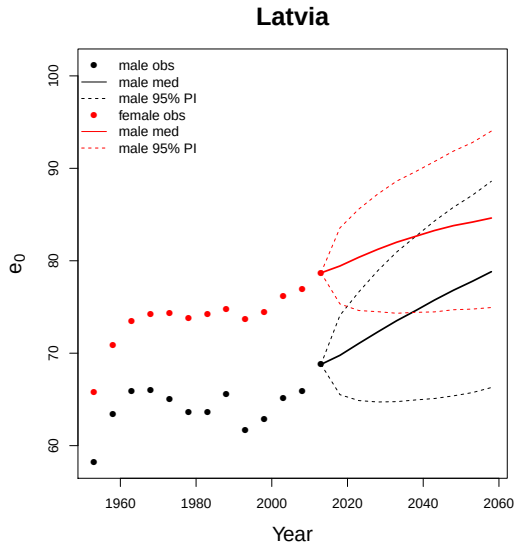

Italy
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

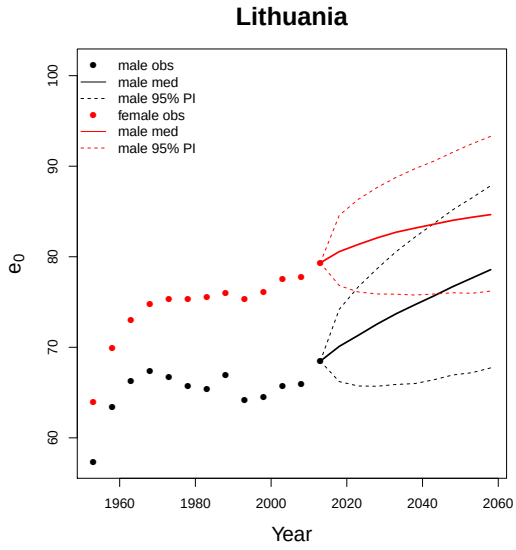

Jamaica
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

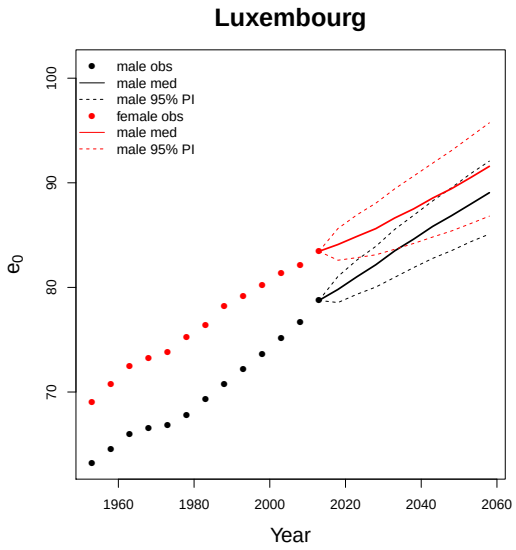

Japan
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

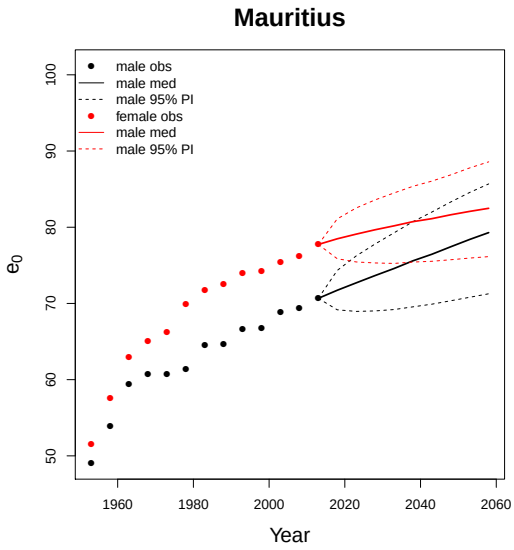

Kazakhstan
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

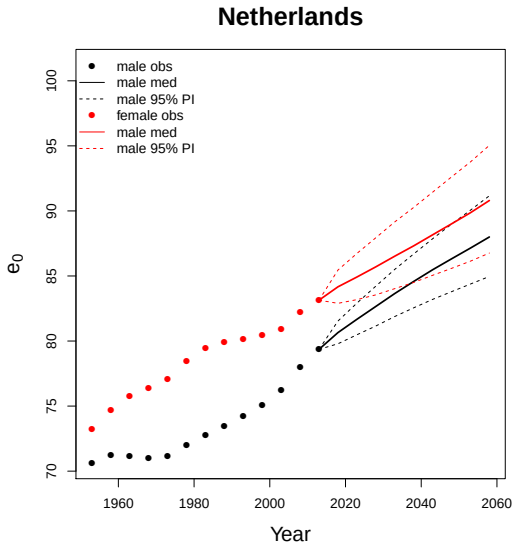

Kyrgyzstan
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

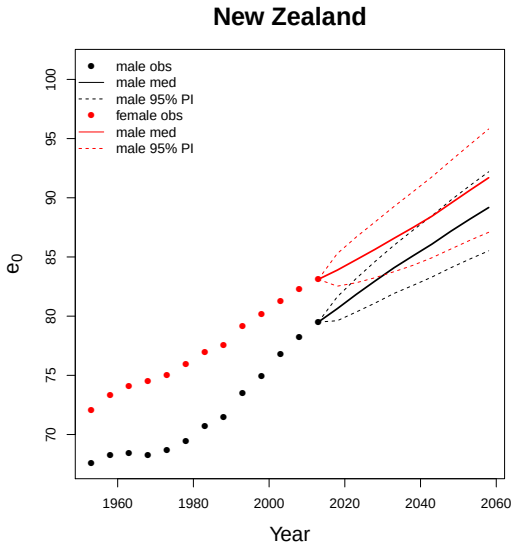

Latvia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

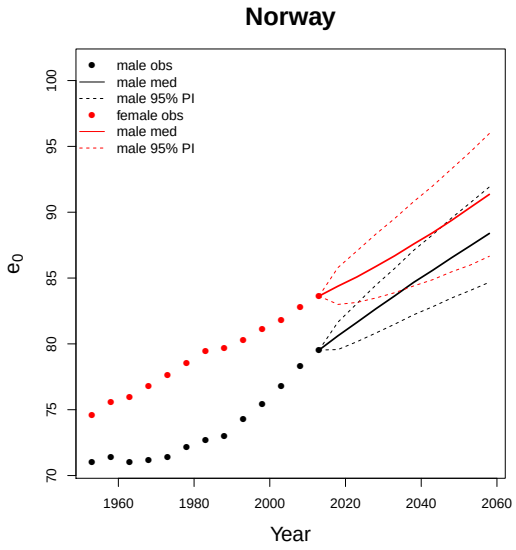

Lithuania
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

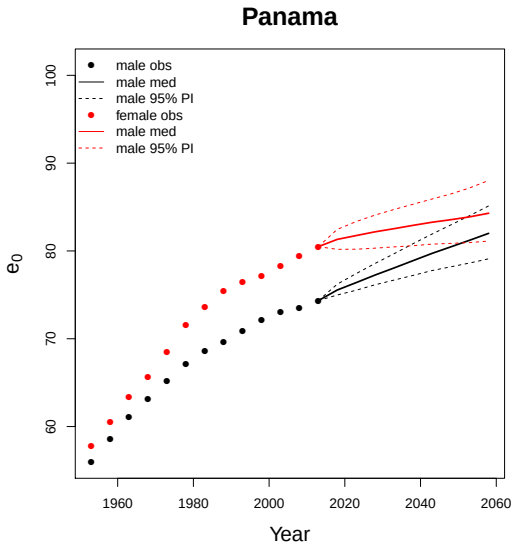

Luxembourg
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

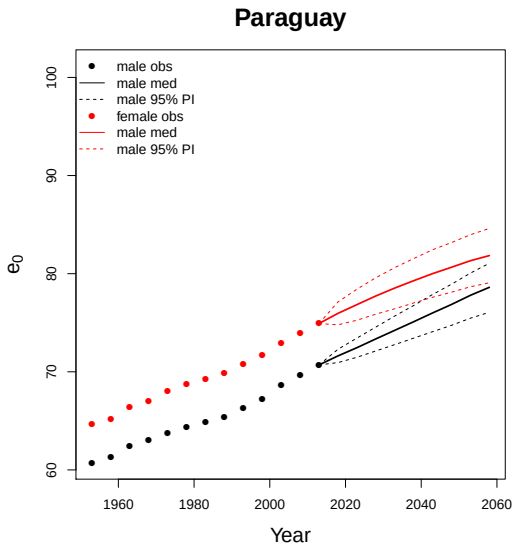

Mauritius
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

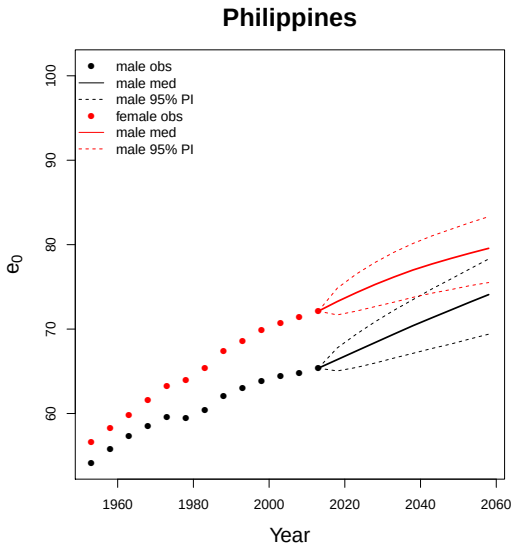

Netherlands
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

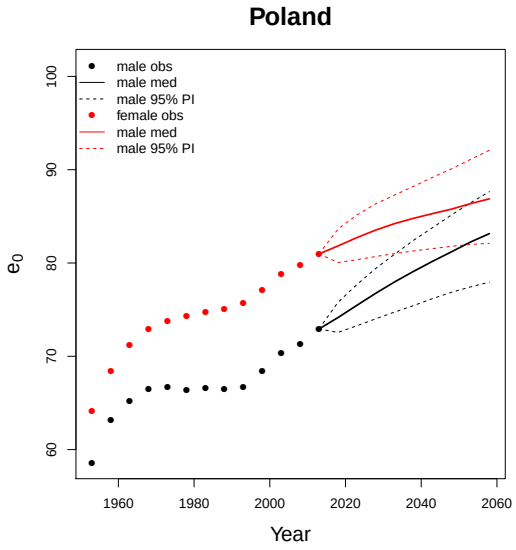

New Zealand
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

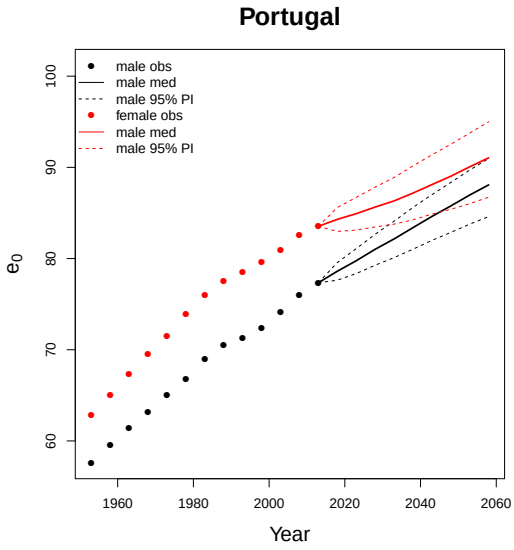

Norway
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

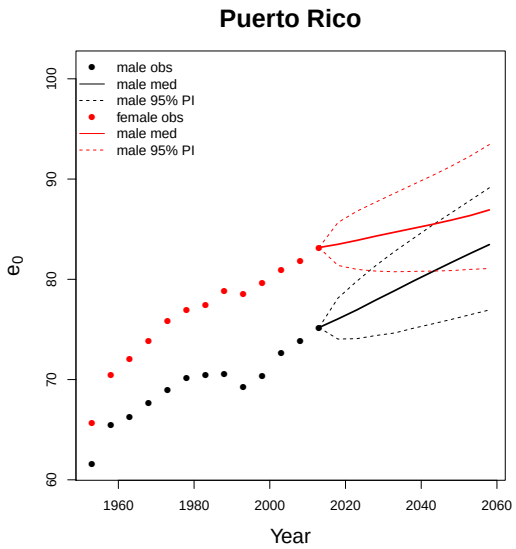

Panama
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

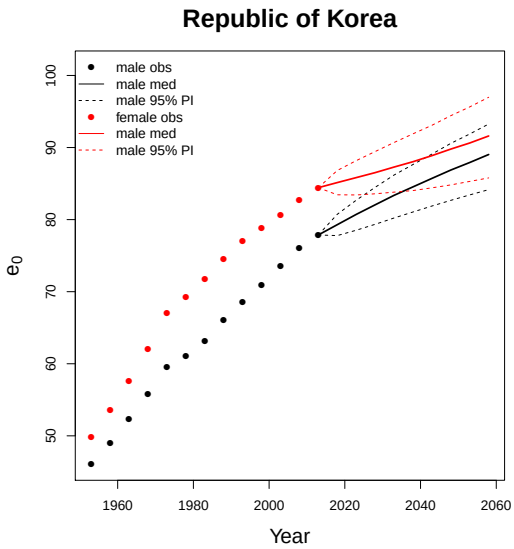

Paraguay
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

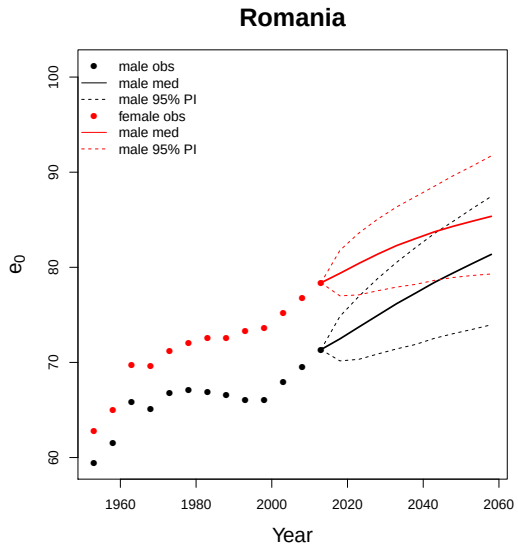

Philippines
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

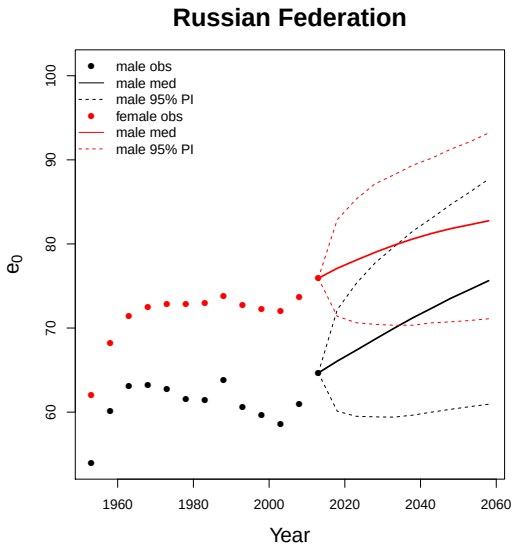

Poland
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

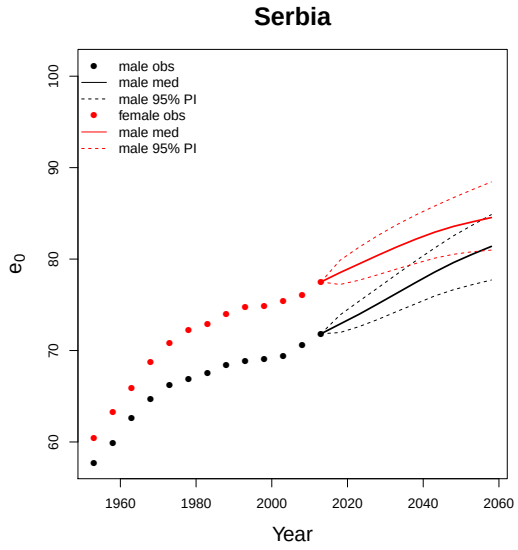

Portugal
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

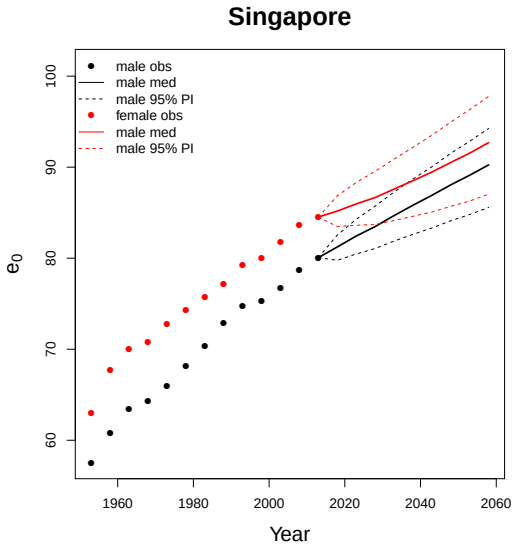

Puerto Rico
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

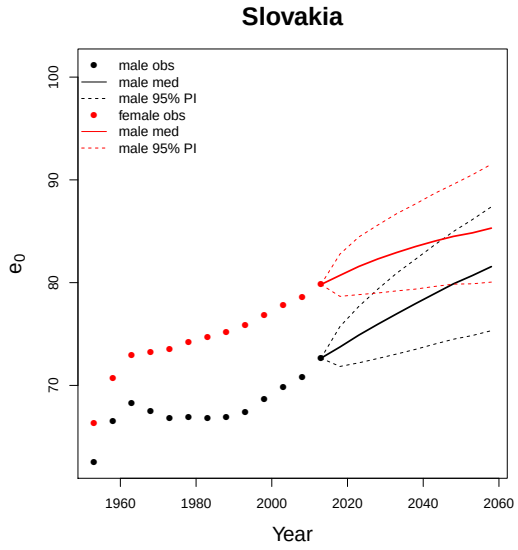

Republic of Korea
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

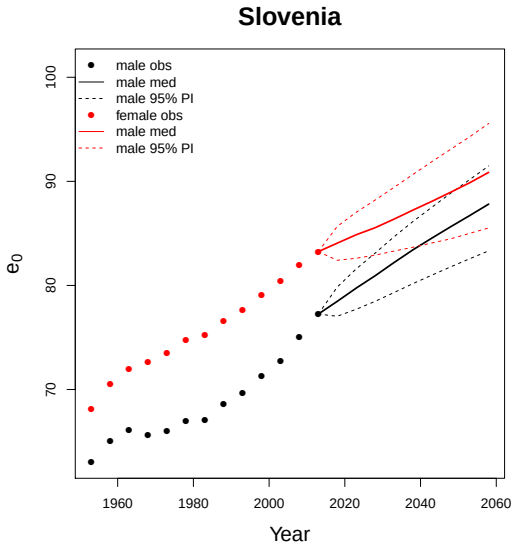

Romania
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

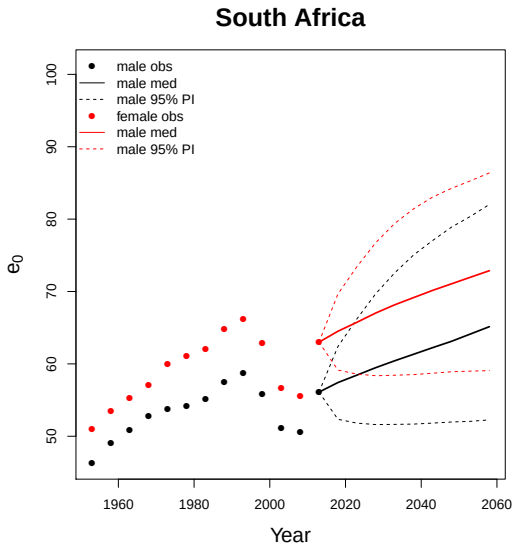

Russian Federation
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

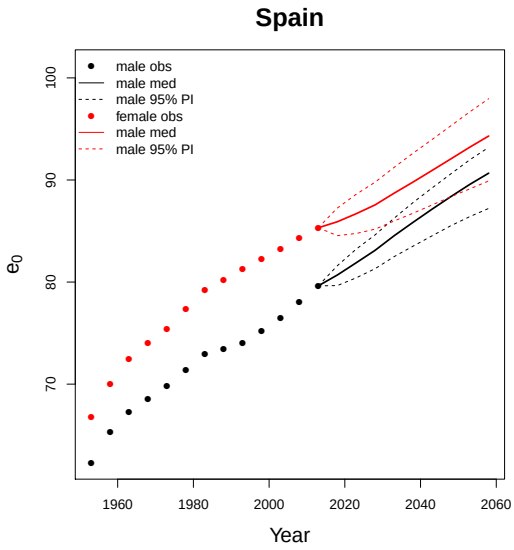

Serbia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

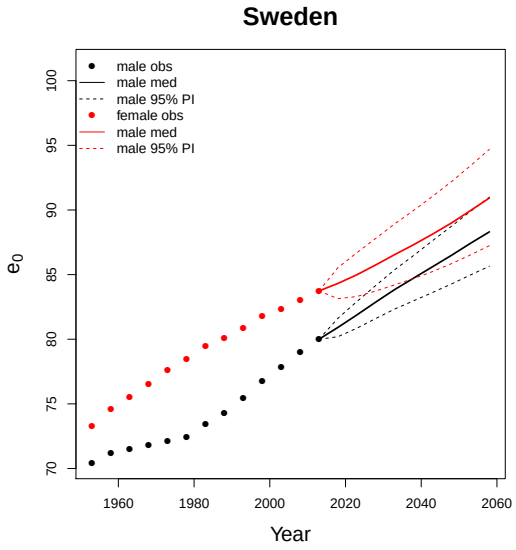

Singapore
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

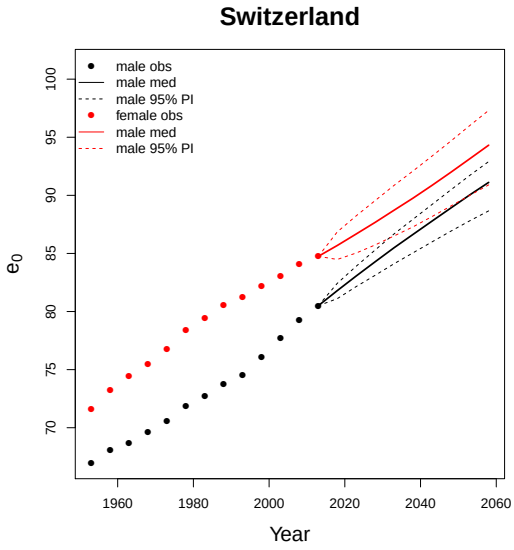

Slovakia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

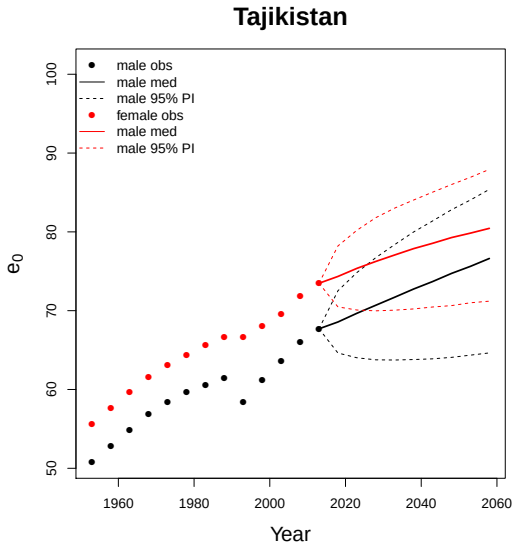

Slovenia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

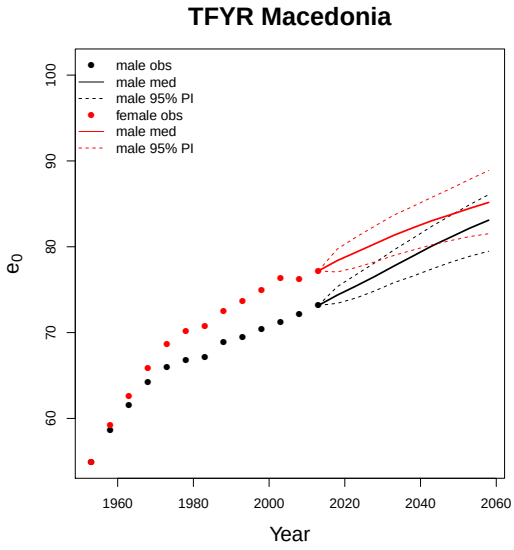

South Africa
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

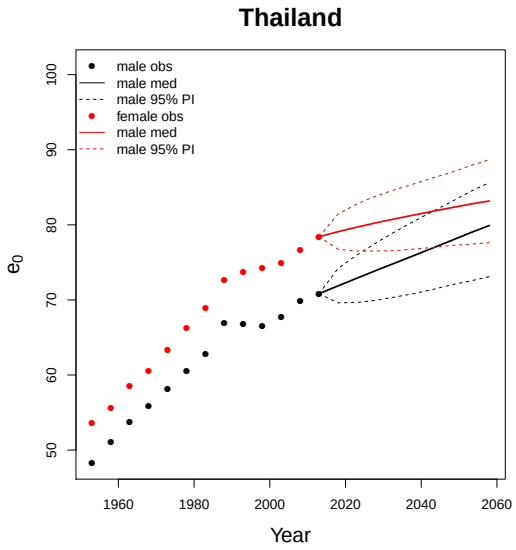

Spain
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

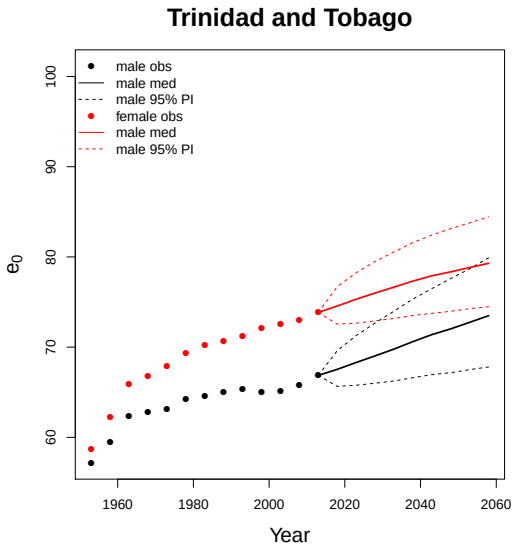

Sweden
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

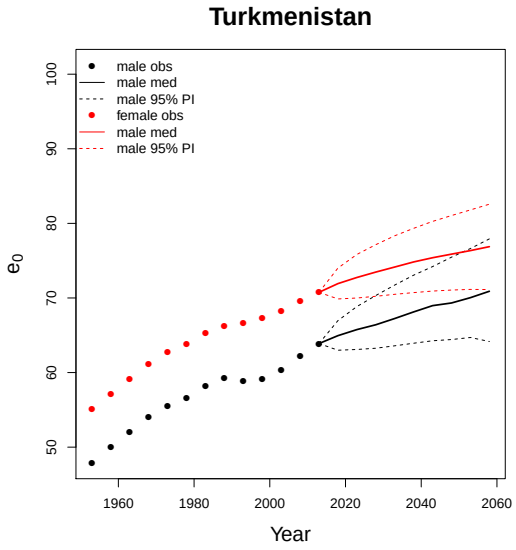

Switzerland
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

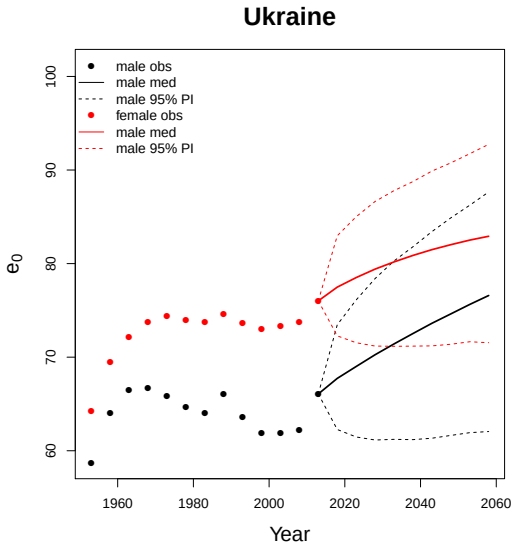

Tajikistan
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

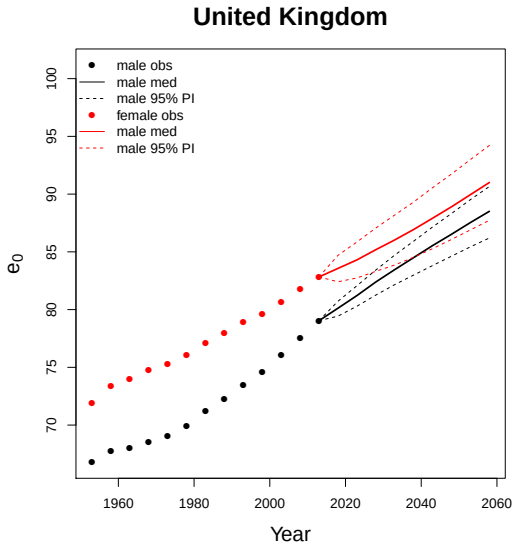

TFYR Macedonia
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

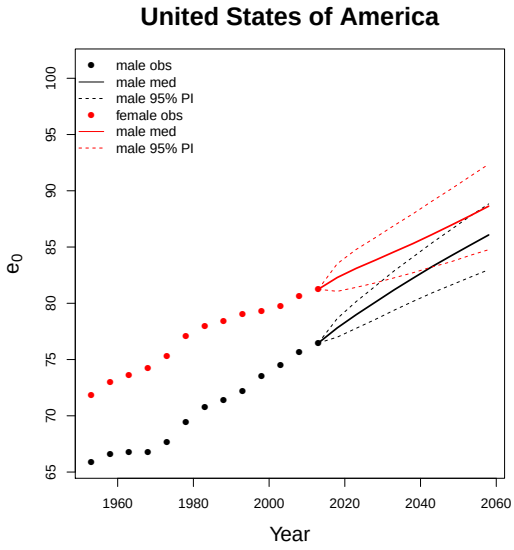

Thailand
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

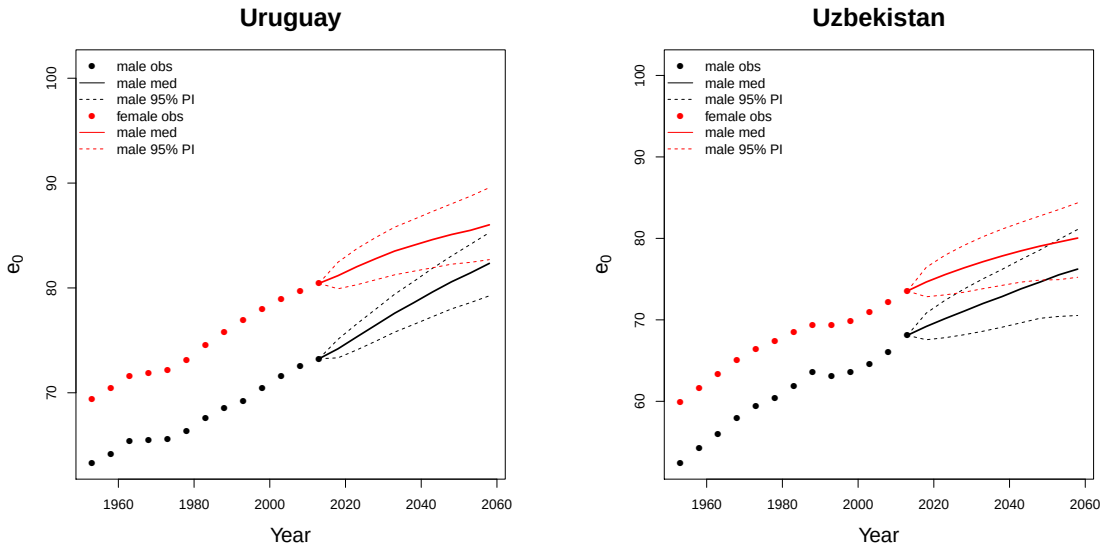

Trinidad and Tobago
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

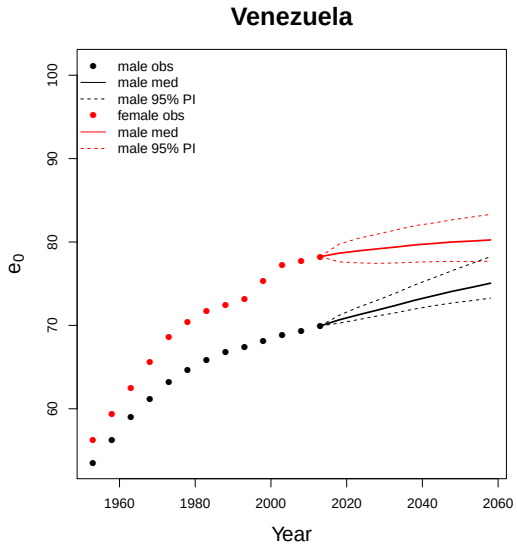

Turkmenistan
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year


Ukraine
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year


United Kingdom
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year


United States of America
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year


Uruguay
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year
Uzbekistan
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year


Venezuela
male obs
male med
male 95% PI
female obs
male med
male 95% PI
e0
Year

Appendix C

APPENDICES TO CHAPTER 4

In this appendix we present the proof of Theorem 10 in Chapter 4, which slightly extends the Bernstein-type inequality proven by Banna et al. (2016) in which the random matrix sequence is assumed to be β-mixing. The proof is largely identical to theirs, and we include it here mainly for completeness.

In the following, τ_k is abbreviate of $\tau(k)$ for $k \geq 1$. If a matrix $\mathbf{X}$ is positive semidefinite, denote it as $\mathbf{X} \succeq 0$. For any $x > 0$, we define $h(x) = x^{-2}(e^x - x - 1)$. Denote the floor, ceiling, and integer parts of a real number x by $\lfloor x \rfloor$, $\lceil x \rceil$, and $[x]$. For any two real numbers a, b, denote $a \vee b := \max\{a, b\}$. Denote the exponential of matrix $\mathbf{X}$ as $\exp(\mathbf{X}) = \mathbf{I}_p + \sum_{q=1}^{\infty} \mathbf{X}^q / q!$. Letting σ_1 and σ_2 be two sigma fields, denote $\sigma_1 \vee \sigma_2$ to be the smallest sigma field that contains σ_1 and σ_2 as sub-sigma fields.

A roadmap of this appendix is as follows. Section C.1 formally introduces the concept of τ-mixing coefficient. Section C.2 previews the proof of Theorem 10 and indicates some major differences from the proofs in Banna et al. (2016). Section C.3 contains the construction of Cantor-like set which is essential for decoupling dependent matrices. Section C.4 develops a major decoupling lemma for τ-mixing random matrices and will be used in Section C.6 to prove Lemma 21. Then Section C.5 finishes the proof of Theorem 10.

C.1 Introduction to τ-mixing random sequence

This section introduces the τ-mixing coefficient. Consider $(\Omega, \mathcal{F}, \mathbb{P})$ to be a probability space, X an L_1-integrable random variable taking value in a Polish space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, and $\mathcal{A}$ a sigma

algebra of $\mathcal{F}$. The τ-measure of dependence between X and $\mathcal{A}$ is defined to be

$$\tau(\mathcal{A}, X; \|\cdot\|_{\mathcal{X}}) = \Big\| \sup_{g \in \Lambda(\|\cdot\|_{\mathcal{X}})} \Big\{ \int g(x) \mathbb{P}_{X|\mathcal{A}}(\mathsf{d}x) - \int g(x) \mathbb{P}_X(\mathsf{d}x) \Big\} \Big\|_{L(1)},$$

where $\mathbb{P}_X$ is the distribution of X, $\mathbb{P}_{X|\mathcal{A}}$ is the conditional distribution of X given $\mathcal{A}$, and $\Lambda(\|\cdot\|_{\mathcal{X}})$ stands for the set of 1-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ with respect to the norm $\|\cdot\|_{\mathcal{X}}$.

The following two lemmas from Dedecker and Prieur (2004) and Dedecker et al. (2007) characterize the intrinsic "coupling property" of τ-measure of dependence, which will be heavily exploited in the derivation of our results.

Lemma 18 (Lemma 3 in Dedecker and Prieur (2004)). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X be an integrable random variable with values in a Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $\mathcal{A}$ a sigma algebra of $\mathcal{F}$. If Y is a random variable distributed as X and independent of $\mathcal{A}$, then*

$$\tau(\mathcal{A}, X; \|\cdot\|_{\mathcal{X}}) \leq \mathbb{E}\|X - Y\|_{\mathcal{X}}.$$

Lemma 19 (Lemma 5.3 in Dedecker et al. (2007)). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{A}$ be a sigma algebra of $\mathcal{F}$, and X be a random variable with values in a Polish space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$. Assume that $\int \|x - x_0\|_{\mathcal{X}} \mathbb{P}_X(\mathsf{d}x)$ is finite for any $x_0 \in \mathcal{X}$. Assume that there exists a random variable U uniformly distributed over $[0,1]$, independent of the sigma algebra generated by X and $\mathcal{A}$. Then there exists a random variable $\tilde{X}$, measurable with respect to $\mathcal{A} \vee \sigma(X) \vee \sigma(U)$, independent of $\mathcal{A}$ and distributed as X, such that*

$$\tau(\mathcal{A}, X; \|\cdot\|_{\mathcal{X}}) = \mathbb{E}\|X - \tilde{X}\|_{\mathcal{X}}.$$

Let $\{X_j\}_{j \in J}$ be a set of $\mathcal{X}$-valued random variables with index set J of finite cardinality. Then define

$$\tau(\mathcal{A}, \{X_j \in \mathcal{X}\}_{j \in J}; \|\cdot\|_{\mathcal{X}}) = \Big\| \sup_{g \in \Lambda(\|\cdot\|'_{\mathcal{X}})} \Big\{ \int g(x) \mathbb{P}_{\{X_j\}_{j \in J}|\mathcal{A}}(\mathsf{d}x) - \int g(x) \mathbb{P}_{\{X_j\}_{j \in J}}(\mathsf{d}x) \Big\} \Big\|_{L(1)},$$

where $\mathbb{P}_{\{X_j\}_{j \in J}}$ is the distribution of $\{X_j\}_{j \in J}$, $\mathbb{P}_{\{X_j\}_{j \in J}|\mathcal{A}}$ is the conditional distribution of $\{X_j\}_{j \in J}$ given $\mathcal{A}$, and $\Lambda(\|\cdot\|'_{\mathcal{X}})$ stands for the set of 1-Lipschitz functions from $\underbrace{\mathcal{X} \times \cdots \times \mathcal{X}}_{\mathrm{card}(J)}$ to

$\mathbb{R}$ with respect to the norm $\|x\|'_{\mathcal{X}} := \sum_{j\in J} \|x_j\|_{\mathcal{X}}$ induced by $\|\cdot\|_{\mathcal{X}}$ for any $x = (x_1, \ldots, x_J) \in \mathcal{X}^{\mathrm{card}(J)}$.

Using these concepts, for a sequence of temporally dependent data $\{X_t\}_{t\in\mathbb{Z}}$, we are ready to define measure of temporal correlation strength as follows,

$$\tau(k; \{X_t\}_{t\in\mathbb{Z}}, \|\cdot\|_{\mathcal{X}}) := \sup_{i>0} \max_{1\le\ell\le i} \frac{1}{\ell} \sup\{\tau\{\sigma(X_{-\infty}^{a}), \{X_{j_1}, \ldots, X_{j_\ell}\}; \|\cdot\|_{\mathcal{X}}\}, a+k \le j_1 < \cdots < j_\ell\},$$

where the inner supremum is taken over all $a \in \mathbb{Z}$ and all ℓ-tuples $(j_1, \ldots, j_\ell)$. $\{X_t\}_{t\in\mathbb{Z}}$ is said to be τ-mixing if $\tau(k; \{X_t\}_{t\in\mathbb{Z}}, \|\cdot\|_{\mathcal{X}})$ converges to zero as $k \to \infty$. In Dedecker et al. (2007) the authors gave numerous examples of random sequences that are τ-mixing.

C.2 Overview of proof of Theorem 10

The proof of Theorem 10 follows largely the proof of Theorem 1 in Banna et al. (2016). Section C.3 reviews the Cantor-set construction developed and used in Merlevède et al. (2009) and Banna et al. (2016). Lemma 20 is a slight extension of Lemma 8 in Banna et al. (2016). The major difference is that the 0-1 function used to quantify the distance between two random matrices under β-mixing by Berbee's decoupling lemma (Berbee, 1979) is replaced by an absolute distance function, which is used under τ-mixing by Lemma 18 (Dedecker and Prieur, 2004). Proofs of Lemma 21 and the rest of Theorem 10 follow largely the proofs of Proposition 7 and Theorem 1 in Banna et al. (2016) respectively, though with more algebras involved.

C.3 Construction of Cantor-like set

We follow Banna et al. (2016) to construct the Cantor-like set K_B for $\{1, \ldots, B\}$. Let $\delta = \frac{\log 2}{2\log B}$ and $\ell_B = \sup\{k \in \mathbb{Z}^+ : \frac{B\delta(1-\delta)^{k-1}}{2^k} \ge 2\}$. We abbreviate $\ell := \ell_B$. Let $n_0 = B$ and for $j \in \{1, \ldots, \ell\}$,

$$n_j = \left\lceil \frac{B(1-\delta)^j}{2^j} \right\rceil \quad \text{and} \quad d_{j-1} = n_{j-1} - 2n_j.$$

We start from the set $\{1,\ldots,B\}$ and divide the set into three disjoint subsets I_1^1, J_0^1, I_1^2 so that $\text{card}(I_1^1) = \text{card}(I_1^2) = n_1$ and $\text{card}(J_0^1) = d_0$. Specifically,

$$I_1^1 = \{1,\ldots,n_1\},\ J_0^1 = \{n_1+1,\ldots,n_1+d_0\},\ I_1^2 = \{n_1+d_0+1,\ldots,2n_1+d_0\},$$

where $B = 2n_1+d_0$. Then we divide I_1^1, I_1^2 with J_0^1 unchanged. I_1^1 is divided into three disjoint subsets I_2^1, J_1^1, I_2^2 in the same way as the previous step with $\text{card}(I_2^1) = \text{card}(I_2^2) = n_2$ and $\text{card}(J_1^1) = d_1$. We obtain

$$I_2^1 = \{1,\ldots,n_2\},\ J_1^1 = \{n_2+1,\ldots,n_2+d_1\},\ I_2^2 = \{n_2+d_1+1,\ldots,2n_2+d_1\},$$

where $n_1 = 2n_2+d_1$. Similarly, I_1^2 is divided into I_2^3, J_1^2, I_2^4 with $\text{card}(I_2^3) = \text{card}(I_2^4) = n_2$ and $\text{card}(J_1^2) = d_1$. We obtain

$$I_2^3 = \{2n_2+d_0+d_1+1,\ldots,3n_2+d_0+d_1\},\ J_1^2 = \{3n_2+d_0+d_1+1,\ldots,3n_2+d_0+2d_1\},$$
$$I_2^4 = \{3n_2+d_0+2d_1+1,\ldots,4n_2+d_0+2d_1\},$$

where $B = 4n_2+d_0+2d_1$.

Suppose we iterate this process for k times ($k \in \{1,\ldots,\ell\}$) with intervals $I_k^i, i \in \{1,\ldots,2^k\}$. For each I_k^i, we divide it into three disjoint subsets $I_{k+1}^{2i-1}, J_k^i, I_{k+1}^{2i}$ so that $\text{card}(I_{k+1}^{2i-1}) = \text{card}(I_{k+1}^{2i}) = n_{k+1}$ and $\text{card}(J_k^i) = d_k$. More specifically, if $I_k^i = \{a_k^i,\ldots,b_k^i\}$, then

$$I_{k+1}^{2i-1} = \{a_k^i,\ldots,a_k^i+n_{k+1}-1\},\ J_k^i = \{a_k^i+n_{k+1},\ldots,a_k^i+n_{k+1}+d_k-1\},$$
$$I_{k+1}^{2i} = \{a_k^i+n_{k+1}+d_k,\ldots,a_k^i+2n_{k+1}+d_k-1\}.$$

After ℓ steps, we obtain 2^ℓ disjoint subsets $I_\ell^i, i \in \{1,\ldots,2^\ell\}$ with $\text{card}(I_\ell^i) = n_\ell$. Then the Cantor-like set is defined as

$$K_B = \bigcup_{i=1}^{2^\ell} I_\ell^i,$$

and for each level $k \in \{0,\ldots,\ell\}$ and each $j \in \{1,\ldots,2^k\}$, define

$$K_k^j = \bigcup_{i=(j-1)2^{\ell-k}+1}^{j2^{\ell-k}} I_\ell^i.$$

Some properties derived from this construction are given by Banna et al. (2016):

1. $\delta \leq \frac{1}{2}$ and $\ell \leq \frac{\log B}{\log 2}$;

2. $d_j \geq \frac{B\delta(1-\delta)^j}{2^{j+1}}$ and $n_\ell \leq \frac{B(1-\delta)^\ell}{2^{\ell-1}}$;

3. Each $I_\ell^i, i \in \{1, \ldots, 2^\ell\}$ contains n_ℓ consecutive integers, and for any $i \in \{1, \ldots, 2^{\ell-1}\}$, I_ℓ^{2i-1} and I_ℓ^{2i} are spaced by $d_{\ell-1}$ integers;

4. $\text{card}(K_B) \geq \frac{B}{2}$;

5. For each $k \in \{0, \ldots, \ell\}$ and each $j \in \{1, \ldots, 2^k\}$, $\text{card}(K_k^j) = 2^{\ell-k} n_\ell$. For each $j \in \{1, \ldots, 2^{k-1}\}$, K_k^{2j-1} and K_k^{2j} are spaced by d_{k-1} integers;

6. $K_0^1 = K_B$ and $K_\ell^j = I_\ell^j$ for $j \in \{1, \ldots, 2^\ell\}$.

C.4 A decoupling lemma for τ-mixing random matrices

This section introduces the key tool to decouple τ-mixing random matrices using Cantor-like set constructed in Section C.3. With some abuse of notation, within this section let's use $\{\mathbf{X}_j\}_{j\in\{1,\ldots,n\}}$ to denote a generic sequence of $p \times p$ symmetric random matrices. Assume $\mathbb{E}(\mathbf{X}_j) = \mathbf{0}$ and $\|\mathbf{X}_j\| \leq M$ for some positive constant M and for all $j \geq 1$. For a collection of index sets H_1^k, $k \in \{1, \ldots, d\}$, we assume that their cardinalities are equal and even. Denote $\{\mathbf{X}_j\}_{j\in H_1^k}$ to be the set of matrices whose indices are in H_1^k. Assume $\{\mathbf{X}_j\}_{j\in H_1^1}, \ldots, \{\mathbf{X}_j\}_{j\in H_1^d}$ are mutually independent, while within each block H_1^k the matrices are possibly dependent. For each k, decompose H_1^k into two disjoint sets H_2^{2k-1} and H_2^{2k} with equal size, containing the first and second half of H_1^k respectively. In addition, we denote $\tau_0 := \tau\{\sigma(\{\mathbf{X}_j\}_{j\in H_2^{2k-1}}), \{\mathbf{X}_j\}_{j\in H_2^{2k}}; \|\cdot\|\}$ for some constant $\tau_0 \geq 0$ and for all $k \in \{1, \ldots, d\}$. For a given $\epsilon > 0$, we achieve the following decoupling lemma.

Lemma 20. *We obtain for any $\epsilon > 0$,*

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big) \le \sum_{i=0}^{d}\binom{d}{i}(1+L_1+L_2)^{d-i}(L_1)^i\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^i t\Big(\sum_{k=1}^{2d}\sum_{j\in H_2^k}\tilde{\mathbf{X}}_j\Big)\Big\},$$

$$\mathbb{E}\,\mathrm{tr}\exp\Big(-t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big) \le \sum_{i=0}^{d}\binom{d}{i}(1+L_1+L_2)^{d-i}(L_1)^i\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^{i+1} t\Big(\sum_{k=1}^{2d}\sum_{j\in H_2^k}\tilde{\mathbf{X}}_j\Big)\Big\},$$

where

$$L_1 := pt\epsilon\exp(t\epsilon), \quad L_2 := \exp\{card(H_1^1)tM\}\tau_0/\epsilon,$$

and $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^k}$, $k\in\{1,\dots,2d\}$, are mutually independent and have the same distributions as $\{\mathbf{X}_j\}_{j\in H_2^k}$, $k\in\{1,\dots,2d\}$.

Proof. We prove this lemma by induction. For any $k\in\{1,\dots,d\}$, we have $H_1^k = H_2^{2k-1}\cup H_2^{2k}$ and hence $\sum_{j\in H_1^k}\mathbf{X}_j = \sum_{j\in H_2^{2k-1}}\mathbf{X}_j + \sum_{j\in H_2^{2k}}\mathbf{X}_j$.

By Lemma 19, for each $k\in\{1,\dots,d\}$, we could find a sequence of random matrices $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^{2k}}$ and an independent uniformly distributed random variable U_k on $[0,1]$ such that

1. $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^{2k}}$ is measurable with respect to the sigma field $\sigma(\{\mathbf{X}_j\}_{j\in H_2^{2k-1}})\vee\sigma(\{\mathbf{X}_j\}_{j\in H_2^{2k}})\vee\sigma(U_k)$;

2. $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^{2k}}$ is independent of $\sigma(\{\mathbf{X}_j\}_{j\in H_2^{2k-1}})$;

3. $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^{2k}}$ has the same distribution as $\{\mathbf{X}_j\}_{j\in H_2^{2k}}$;

4. $\mathbb{P}(\|\sum_{j\in H_2^{2k}}\mathbf{X}_j - \sum_{j\in H_2^{2k}}\tilde{\mathbf{X}}_j\| > \epsilon_k) \le \mathbb{E}(\|\sum_{j\in H_2^{2k}}\mathbf{X}_j - \sum_{j\in H_2^{2k}}\tilde{\mathbf{X}}_j\|)/\epsilon_k \le \tau_0/\epsilon_k$ by Markov's inequality and the fact that $\tau_0 = \sum_{j\in H_2^{2k}}\mathbb{E}(\|\mathbf{X}_j - \tilde{\mathbf{X}}_j\|)$.

To make notation easier to follow, we set equal value to ϵ_k for $k\in\{1,\dots,d\}$ and denote it as ϵ. Moreover, we denote the event $\Gamma_k = \{\|\sum_{j\in H_2^{2k}}\tilde{\mathbf{X}}_j - \sum_{j\in H_2^{2k}}\mathbf{X}_j\| \le \epsilon\}$ for $k\in\{1,\dots,d\}$.

For the base case $k=1$.

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big) = \underbrace{\mathbb{E}\Big\{\mathbb{1}_{\Gamma_1}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}}_{I} + \underbrace{\mathbb{E}\Big\{\mathbb{1}_{(\Gamma_1)^c}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}}_{II}.$$

Notice the definitions of terms I and II therein.

We have

$$
\begin{aligned}
I &= \mathbb{E}\Big[\mathbb{1}_{\Gamma_1} \operatorname{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\mathbf{X}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}\Big] \\
&\le \mathbb{E}\operatorname{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\} \\
&+ \mathbb{E}\Big(\mathbb{1}_{\Gamma_1}\Big[\operatorname{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\mathbf{X}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\} - \operatorname{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}\Big]\Big)
\end{aligned}
$$

By linearity of expectation and the facts that $\operatorname{tr}(\mathbf{X}) \le p\|\mathbf{X}\|$ and $\|\exp(\mathbf{X}) - \exp(\mathbf{Y})\| \le \|\mathbf{X}-\mathbf{Y}\|\exp(\|\mathbf{X}-\mathbf{Y}\|)\exp(\|\mathbf{Y}\|)$, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big(\mathbb{1}_{\Gamma_1}\Big[\operatorname{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\mathbf{X}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\} - \operatorname{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}\Big]\Big) \\
\le&\mathbb{E}\Big[\mathbb{1}_{\Gamma_1}p\Big\|\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\mathbf{X}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\} - \exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}\Big\|\Big] \\
\le&\mathbb{E}\Big[\mathbb{1}_{\Gamma_1}p\Big\|t\sum_{j\in H_2^2}(\mathbf{X}_j - \tilde{\mathbf{X}}_j)\Big\|\exp\Big\{\Big\|t\sum_{j\in H_2^2}(\mathbf{X}_j - \tilde{\mathbf{X}}_j)\Big\|\Big\}\exp\Big\{\Big\|t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\|\Big\}\Big].
\end{aligned}
$$

By spectral mapping theorem, for a symmetric matrix $\mathbf{X}$ with $\|\mathbf{X}\| \le M$, we have $\exp(\|\mathbf{X}\|) \le \|\exp(\mathbf{X})\| \vee \|\exp(-\mathbf{X})\| \le \|\exp(\mathbf{X})\| + \|\exp(-\mathbf{X})\|$. Moreover, since $\exp(\mathbf{X})$ is always positive definite for any matrix $\mathbf{X}$ and $\|\mathbf{X}\| \le \operatorname{tr}(\mathbf{X})$ for any positive definite symmetric matrix $\mathbf{X}$, we obtain $\|\exp(\mathbf{X})\| \le \operatorname{tr}\exp(\mathbf{X})$ and $\|\exp(-\mathbf{X})\| \le \operatorname{tr}\exp(-\mathbf{X})$. In addition, since we have $\|\sum_{j\in H_2^2}(\mathbf{X}_j - \tilde{\mathbf{X}}_j)\| \le \epsilon$ on Γ_1, we could further bound the inequality

above by

$$\mathbb{E}\Big[\mathbb{1}_{\Gamma_1} pt\epsilon \exp(t\epsilon)\Big\| \exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}\Big\|\Big]$$

$$\leq pt\epsilon \exp(t\epsilon)\Big[\mathbb{E}\,\mathrm{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}$$

$$+\mathbb{E}\,\mathrm{tr}\exp\Big\{-t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}\Big].$$

Putting together, we reach

$$I \leq \{1 + pt\epsilon \exp(t\epsilon)\}\mathbb{E}\,\mathrm{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}$$

$$+ pt\epsilon \exp(t\epsilon)\mathbb{E}\,\mathrm{tr}\exp\Big\{-t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}. \tag{C.1}$$

We then aim at II. For this, the proof largely follows the same argument as in Banna et al. (2016). Omitting the details, we obtain

$$II \leq \exp\{\mathrm{card}(H_1^1)tM\}(\tau_0/\epsilon)\mathbb{E}\,\mathrm{tr}\exp\Big\{t\Big(\sum_{j\in H_2^2}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_k^1}\mathbf{X}_j\Big)\Big\}. \tag{C.2}$$

Denote $L_1 := pt\epsilon \exp(t\epsilon)$ and $L_2 := \exp\{\mathrm{card}(H_1^1)tM\}\tau_0/\epsilon$. Combining (C.1) and (C.2) yields

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)$$

$$\leq (1 + L_1 + L_2)\mathbb{E}\,\mathrm{tr}\exp\Big\{t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}$$

$$+ L_1\mathbb{E}\,\mathrm{tr}\exp\Big\{-t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}$$

$$= \sum_{i=0}^{1}\binom{1}{i}(1 + L_1 + L_2)^{1-i}(L_1)^i\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^i t\Big(\sum_{j\in H_2^1}\mathbf{X}_j + \sum_{j\in H_2^2}\tilde{\mathbf{X}}_j + \sum_{k=2}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\Big\}.$$

This finishes the base case.

The induction steps are followed similarly and we omit the details. By iterating d times, we arrive at the following inequality:

$$\begin{aligned}&\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\\&\leq\sum_{i=0}^{d}\binom{d}{i}(1+L_1+L_2)^{d-i}(L_1)^i\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^i t\Big(\sum_{k=1}^{d}\sum_{j\in H_2^{2k-1}}\mathbf{X}_j+\sum_{k=1}^{d}\sum_{j\in H_2^{2k}}\tilde{\mathbf{X}}_j\Big)\Big\},\end{aligned}\tag{C.3}$$

where $\{\mathbf{X}_j\}_{j\in H_2^{2k-1}},\ k\in\{1,\ldots,d\}$ and $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^{2k}},\ k\in\{1,\ldots,d\}$ are mutually independent. In addition, they have the same distributions as $\{\mathbf{X}_j\}_{j\in H_2^{2k-1}},\ k\in\{1,\ldots,d\}$ and $\{\mathbf{X}_j\}_{j\in H_2^{2k}},\ k\in\{1,\ldots,d\}$, respectively. For the sake of simplicity and clarity, we add an upper tilde to the matrices with indices in $H_2^{2k-1},\ k\in\{1,\ldots,d\}$, i.e., $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^{2k-1}}$ is identically distributed as $\{\mathbf{X}_j\}_{j\in H_2^{2k-1}}$ for $k\in\{1,\ldots,d\}$. Hence (C.3) could be rewritten as

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\leq\sum_{i=0}^{d}\binom{d}{i}(1+L_1+L_2)^{d-i}(L_1)^i\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^i t\Big(\sum_{k=1}^{2d}\sum_{j\in H_2^k}\tilde{\mathbf{X}}_j\Big)\Big\},$$

where $\{\tilde{\mathbf{X}}_j\}_{j\in H_2^k},\ k\in\{1,\ldots,2d\}$ are mutually independent and their distributions are the same as $\{\mathbf{X}_j\}_{j\in H_2^k},\ k\in\{1,\ldots,2d\}$.

By changing $\mathbf{X}$ to $-\mathbf{X}$, we immediately get the following bound:

$$\mathbb{E}\,\mathrm{tr}\exp\Big(-t\sum_{k=1}^{d}\sum_{j\in H_1^k}\mathbf{X}_j\Big)\leq\sum_{i=0}^{d}\binom{d}{i}(1+L_1+L_2)^{d-i}(L_1)^i\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^{i+1}t\Big(\sum_{k=1}^{2d}\sum_{j\in H_2^k}\tilde{\mathbf{X}}_j\Big)\Big\}.$$

This completes the proof of Lemma 20. □

C.5 Proof of Theorem 10

Proof. Without loss of generality, let $\psi_1=\tilde{\psi}_1$.

Case I. First of all, we consider $M=1$.

Step I (Summation decomposition). Let $B_0=n$ and $\mathbf{U}_j^{(0)}=\mathbf{X}_j$ for $j\in\{1,\ldots,n\}$. Let K_{B_0} be the Cantor-like set from $\{1,\ldots,B_0\}$ by construction of Section C.3, $K_{B_0}^c=$

$\{1,\ldots,B_0\}\setminus K_{B_0}$, and $B_1=\mathrm{card}(K_{B_0}^c)$. Then define

$$\mathbf{U}_j^{(1)}=\mathbf{X}_{i_j},\ \text{where } i_j\in K_{B_0}^c=\{i_1,\ldots,i_{B_1}\}.$$

For each $i\geq 1$, let K_{B_i} be constructed from $\{1,\ldots,B_i\}$ by the same Cantor-like set construction. Denote $K_{B_i}^c=\{1,\ldots,B_i\}\setminus K_{B_i}$ and $B_{i+1}=\mathrm{card}(K_{B_i}^c)$. Then

$$\mathbf{U}_j^{(i+1)}=\mathbf{U}_{k_j}^{(i)},\ \text{where } k_j\in K_{B_i}^c=\{k_1,\ldots,k_{B_{i+1}}\}.$$

We stop the process when there is a smallest L such that $B_L\leq 2$. Then we have for $i\leq L-1$, $B_i\leq n2^{-i}$ because each Cantor-like set $K_{B_{i+1}}$ has cardinality greater than $B_i/2$. Also notice that $L\leq[\log n/\log 2]$.

For $i\in\{0,\ldots,L-1\}$, denote

$$\mathbf{S}_i=\sum_{j\in K_{B_i}}\mathbf{U}_j^{(i)}\ \text{and}\ \mathbf{S}_L=\sum_{j\in K_{B_{L-1}^c}}\mathbf{U}_j^{(L)}.$$

Then we observe

$$\sum_{j=1}^{n}\mathbf{X}_j=\sum_{i=0}^{L}\mathbf{S}_i.$$

Step II (Bounding Laplacian transform). This step hinges on the following lemma, which provides an upper bound for the Laplace transform of sum of a sequence of random matrices which are τ-mixing with geometric decay, i.e., $\tau(k)\leq\psi_1\exp\{-\psi_2(k-1)\}$ for all $k\geq 1$ for some constants $\psi_1,\psi_2>0$.

Lemma 21 (Proof in Section C.6)**.** *For a sequence of $p\times p$ matrices $\{\mathbf{X}_i\}$, $i\in\{1,\ldots,B\}$ satisfying conditions in Theorem 10 with $M=1$ and $\psi_1\geq p^{-1}$, there exists a subset $K_B\subseteq\{1,\ldots,B\}$ such that for $0<t\leq\min\{1,\frac{\psi_2}{8\log(\psi_1 B^6 p)}\}$,*

$$\log\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{j\in K_B}\mathbf{X}_j\Big)\leq\log p+4h(4)Bt^2\nu^2+151\Big[1+\exp\Big\{\frac{1}{\sqrt{p}}\exp\Big(-\frac{\psi_2}{64t}\Big)\Big\}\Big]\frac{t^2}{\psi_2}\exp\Big(-\frac{\psi_2}{64t}\Big).$$

For each $\mathbf{S}_i, i\in\{0,\ldots,L-1\}$, by applying Lemma 21 with $B=B_i$, we have for any positive t satisfying $0<t\leq\min\{1,\frac{\psi_2}{8\log\{\psi_1(n2^{-i})^6 p\}}\}$,

$$\log\mathbb{E}\,\mathrm{tr}\exp(t\mathbf{S}_i)\leq\log p+t^2(C_1 2^{-i}n+C_{2,i})$$

where $C_1 := 4h(4)\nu^2, C_{2,i} := 302 \cdot 2^{\frac{6i}{8}}/\psi_2 n^{\frac{6}{8}}$.

Denote

$$\tilde{f}(\psi_1, \psi_2, i) := \min\Big\{1, \frac{\psi_2}{8\log\{\psi_1(n2^{-i})^6 p\}}\Big\}.$$

For any $0 < t \le \tilde{f}(\psi_1, \psi_2, i)$, we obtain

$$\log \mathbb{E}\,\mathrm{tr}\exp(t\mathbf{S}_i) \le \log p + \frac{t^2(C_1 2^{-i} n + C_{2,i})}{1 - t/\tilde{f}(\psi_1, \psi_2, i)} \le \log p + \frac{t^2\{C_1^{\frac{1}{2}}(2^{-i}n)^{\frac{1}{2}} + C_{2,i}^{\frac{1}{2}}\}^2}{1 - t/\tilde{f}(\psi_1, \psi_2, i)}.$$

For $\mathbf{S}_L$, since $B_L \le 2$, for $0 < t \le 1$,

$$\log \mathbb{E}\,\mathrm{tr}\exp(t\mathbf{S}_L) \le \log p + t^2 h(2t)\lambda_{\max}\{\mathbb{E}(\mathbf{S}_L^2)\} \le \log p + \frac{2t^2\nu^2}{1-t}.$$

Denote $\sigma_i := C_1^{\frac{1}{2}}(2^{-i}n)^{\frac{1}{2}} + C_{2,i}^{\frac{1}{2}}$, $\sigma_L := \sqrt{2}\nu$, $\kappa_i := 1/\tilde{f}(\psi_1, \psi_2, i)$, and $\kappa_L := 1$.

Summing up, we have

$$\sum_{i=0}^{L} \sigma_i = \sum_{i=0}^{L-1}\{C_1^{\frac{1}{2}}(2^{-i}n)^{\frac{1}{2}} + C_{2,i}^{\frac{1}{2}}\} + \sqrt{2}\nu \le 15\sqrt{n}\nu + 60\sqrt{1/\psi_2},$$
$$\sum_{i=0}^{L} \kappa_i \le \frac{\log n}{\log 2}\max\Big\{1, \frac{8\log(\psi_1 n^6 p)}{\psi_2}\Big\} := \tilde{\psi}(\psi_1, \psi_2, n, p).$$

Hence by Lemma 3 in Merlevède et al. (2009), for $0 < t \le \{\tilde{\psi}(\psi_1, \psi_2, n, p)\}^{-1}$, we have

$$\log \mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{j=1}^{n}\mathbf{X}_j\Big) \le \log p + \frac{t^2\Big(15\sqrt{n}\nu + 60\sqrt{1/\psi_2}\Big)^2}{1 - t\tilde{\psi}(\psi_1, \psi_2, n, p)}.$$

Step III (Matrix Chernoff bound). Lastly by matrix Chernoff bound, we obtain

$$\mathbb{P}\Big\{\lambda_{\max}\Big(\sum_{j=1}^{n}\mathbf{X}_j\Big) \ge x\Big\} \le p\exp\Big\{-\frac{x^2}{8(15^2 n\nu^2 + 60^2/\psi_2) + 2x\tilde{\psi}(\psi_1, \psi_2, n, p)}\Big\}.$$

Case II. We consider general $M > 0$. It is obvious that if $\{\mathbf{X}_t\}_{t\in\mathbb{Z}}$ is a sequence of τ-mixing random matrices such that $\tau(k; \{\mathbf{X}_t\}_{t\in\mathbb{Z}}, \|\cdot\|) \le M\psi_1\exp\{-\psi_2(k-1)\}$, then $\{\mathbf{X}_i/M\}_{i\in\mathbb{Z}}$ is also a sequence of τ-mixing random matrices such that $\tau(k; \{\mathbf{X}_t/M\}_{t\in\mathbb{Z}}, \|\cdot\|) \le$

$\psi_1 \exp\{-\psi_2(k-1)\}$ and $\|\mathbf{X}_t/M\| \le 1$. Then applying the result of Case I to $\{\mathbf{X}_i/M\}_{i\in\mathbb{Z}}$, we obtain

$$\mathbb{P}\Big\{\lambda_{\max}\Big(\sum_{j=1}^{n}\mathbf{X}_j/M\Big) \ge x\Big\} \le p\exp\Big\{-\frac{x^2}{8(15^2 n\nu_M^2 + 60^2/\psi_2) + 2x\tilde{\psi}(\psi_1,\psi_2,n,p)}\Big\},$$

where $\nu_M^2 := \sup_{K\subseteq\{1,\dots,n\}} \frac{1}{\mathrm{card}(K)}\lambda_{\max}\Big\{\mathbb{E}\Big(\sum_{i\in K}\mathbf{X}_i/M\Big)^2\Big\} = \nu^2/M^2$ for ν^2 defined in Theorem 10. Thus

$$\mathbb{P}\Big\{\lambda_{\max}\Big(\sum_{j=1}^{n}\mathbf{X}_j\Big) \ge x\Big\} \le p\exp\Big\{-\frac{x^2}{8(15^2 n\nu^2 + 60^2 M^2/\psi_2) + 2xM\tilde{\psi}(\psi_1,\psi_2,n,p)}\Big\}.$$

This completes the proof of Theorem 10. □

C.6 The proof of Lemma 21

Proof. Let K_B be constructed as in Section C.3 for any arbitrary $B \ge 2$ and $M = 1$.

Case I. If $0 < t \le 4/B$, by Lemma 4 in Banna et al. (2016), we have

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{i\in K_B}\mathbf{X}_i\Big) \le p\exp\Big[t^2 h\Big\{t\lambda_{\max}\Big(\sum_{i\in K_B}\mathbf{X}_i\Big)\Big\}\lambda_{\max}\Big\{\mathbb{E}\Big(\sum_{i\in K_B}\mathbf{X}_i\Big)^2\Big\}\Big].$$

By Weyl's inequality, $\lambda_{\max}(\sum_{i\in K_B}\mathbf{X}_i) \le B$ since $\mathrm{card}(K_B) \le B$, and by definition of ν^2 in Theorem 10, we have $\lambda_{\max}\{\mathbb{E}(\sum_{i\in K_B}\mathbf{X}_i)^2\} \le B\nu^2$. Therefore, we obtain $h\{t\lambda_{\max}(\sum_{i\in K_B}\mathbf{X}_i)\} \le h(tB) \le h(4)$ and

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{i\in K_B}\mathbf{X}_i\Big) \le p\exp\{t^2 h(4) B\nu^2\}. \tag{C.4}$$

Case II. Now we consider the case where $4/B < t \le \min\{1, \frac{\psi_2}{8\log(\psi_1 B^6 p)}\}$.

Step I. Let J be a chosen integer from $\{0,\dots,\ell_B\}$ whose actual value will be determined later. We will use the same notation to denote Cantor-like sets as in Section C.3. By Lemma 20 and similar induction argument as in Banna et al. (2016), we obtain

$$\mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{j\in K_0^1}\mathbf{X}_j\Big) \le \sum_{i_1=0}^{2^0}\cdots\sum_{i_J=0}^{2^{J-1}}\Big[\Big(\prod_{k=1}^{J}A_{k,i_k}\Big)\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^{\sum_{k=1}^{J}i_k}t\Big(\sum_{i'=1}^{2^J}\sum_{j\in K_J^{i'}}\tilde{\mathbf{X}}_j\Big)\Big\}\Big], \tag{C.5}$$

where $\{\tilde{\mathbf{X}}_j\}_{j\in K_J^{i'}}$ for $i' \in \{1, \cdots, 2^J\}$ are mutually independent and have the same distributions as $\{\mathbf{X}_j\}_{j\in K_J^{i'}}$ for $i' \in \{1, \cdots, 2^J\}$,and

$$\begin{aligned}
A_{k,i_k} :=& \binom{2^{k-1}}{i_k}(1 + L_{k,1} + L_{k,2})^{2^{k-1}-i_k}(L_{k,1})^{i_k},\\
\epsilon_k :=& (2pt)^{-\frac{1}{2}}\{2^{\ell-k}n_\ell \exp(t2^{\ell-k+1}n_\ell)\tau_{d_{k-1}+1}\}^{\frac{1}{2}},\\
L_{k,1} :=& (pt/2)^{\frac{1}{2}} \exp(t\epsilon_k)\{2^{\ell-k}n_\ell \exp(t2^{\ell-k+1}n_\ell)\tau_{d_{k-1}+1}\}^{\frac{1}{2}},\\
L_{k,2} :=& (2pt)^{\frac{1}{2}} \exp(t\epsilon_k)\{2^{\ell-k}n_\ell \exp(t2^{\ell-k+1}n_\ell)\tau_{d_{k-1}+1}\}^{\frac{1}{2}},
\end{aligned}$$

Step II: Now we choose J as follows:

$$J = \inf\Big\{k \in \{0, \ldots, \ell\} : \frac{B(1-\delta)^k}{2^k} \le \min\Big\{\frac{\psi_2}{8t^2}, B\Big\}\Big\}.$$

We first bound $\mathbb{E}\operatorname{tr}\exp\{t(\sum_{i'=1}^{2^J}\sum_{j\in K_J^{i'}}\tilde{\mathbf{X}}_j)\}$ and $\mathbb{E}\operatorname{tr}\exp\{-t(\sum_{i'=1}^{2^J}\sum_{j\in K_J^{i'}}\tilde{\mathbf{X}}_j)\}$. From (C.5) we obtain 2^J sets of $\{\tilde{\mathbf{X}}_j\}$ that are mutually independent. To make notation less cluttered, we will remove the upper tilde from $\tilde{\mathbf{X}}_j$ for all j. Denote the number of matrices in each set K_J^i to be $q := 2^{\ell-J}n_\ell$. For each set $K_J^i,\ i \in \{1, \ldots, 2^J\}$, we divide it into consecutive sets with cardinality $\tilde{q}$ and potentially a residual term if q is not divisible by $\tilde{q}$. More specifically, we have $2\tilde{q} \le q$ and $m_{q,\tilde{q}} := [q/2\tilde{q}]$. The value $\tilde{q}$ will be determined later.

Then each set K_J^i contains $2m_{q,\tilde{q}}$ numbers of sets with cardinality $\tilde{q}$ and one set with cardinality less than $2\tilde{q}$. For each $K_J^i,\ i \in \{1, \ldots, 2^J\}$, denote these consecutive sets described above by $Q_k^i,\ k \in \{1, \ldots, 2m_{q,\tilde{q}} + 1\}$. Given these notation, we could rewrite the bound as the following:

$$\begin{aligned}
&\mathbb{E}\operatorname{tr}\exp\Big(t\sum_{i=1}^{2^J}\sum_{j\in K_J^i}\mathbf{X}_j\Big)\\
=&\mathbb{E}\operatorname{tr}\exp\Big(t\sum_{i=1}^{2^J}\sum_{k=1}^{2m_{q,\tilde{q}}+1}\sum_{j\in Q_k^i}\mathbf{X}_j\Big) = \mathbb{E}\operatorname{tr}\exp\Big(t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}}\sum_{j\in Q_{2k}^i}\mathbf{X}_j + t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\Big).
\end{aligned}$$

Since $\operatorname{tr}\exp(\cdot)$ is convex (cf. Proposition 2 in Petz (1994)), by Jensen's inequality, we have

$$\mathbb{E}\operatorname{tr}\exp\Big(t\sum_{i=1}^{2^J}\sum_{j\in K_J^i}\mathbf{X}_j\Big) \le \frac{1}{2}\mathbb{E}\operatorname{tr}\exp\Big(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}}\sum_{j\in Q_{2k}^i}\mathbf{X}_j\Big) + \frac{1}{2}\mathbb{E}\operatorname{tr}\exp\Big(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\Big).$$

Since the number of odd index sets is always equal to or one more than that of the even index sets, the upper bound of $\frac{1}{2}\mathbb{E}\,\mathrm{tr}\exp\left(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}}\sum_{j\in Q_{2k}^i}\mathbf{X}_j\right)$ will always be less than or equal to that of $\frac{1}{2}\mathbb{E}\,\mathrm{tr}\exp\left(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\right)$. Hence we only need to provide an upper bound for $\mathbb{E}\,\mathrm{tr}\exp\left(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\right)$. Our goal is then to replace all $\{\mathbf{X}_j\}_{j\in Q_{2k-1}^i}$ in the last inequality by mutually independent copies $\{\tilde{\mathbf{X}}_j\}_{j\in Q_{2k-1}^i}$ with same distributions for $k\in\{1,\ldots,2m_{q,\tilde{q}}+1\}$, $i\in\{1,\ldots,2^J\}$. Again we will proceed by induction. We first show

$$\begin{aligned}&\mathbb{E}\,\mathrm{tr}\exp\left(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\right)\\ \le&\sum_{i_1=0}^{1}\tilde{A}_{i_1}\mathbb{E}\,\mathrm{tr}\exp\left\{(-1)^{i_1}2t\Big(\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^1}\tilde{\mathbf{X}}_j+\sum_{i=2}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\Big)\right\},\end{aligned}$$

where the constants $\tilde{A}_{i_1}$ will be specified later. For each $\{\mathbf{X}_j\}_{j\in Q_{2k-1}^1}$, $k\in\{1,\ldots,m_{q,\tilde{q}}+1\}$, we could find a sequence of $\{\tilde{\mathbf{X}}_j\}_{j\in Q_{2k-1}^1}$, $k\in\{1,\ldots,m_{q,\tilde{q}}+1\}$ that are mutually independent with each other. More specifically, let $\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1}=\{\mathbf{X}_j\}_{j\in Q_1^1}$. By applying Lemma 19 on $\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1}$ and $\{\mathbf{X}_j\}_{j\in Q_3^1}$ with a chosen $\tilde{\epsilon}>0$, we may find a sequence of random matrices $\{\tilde{\mathbf{X}}_j\}_{j\in Q_3^1}$ such that for each $j_0\in Q_3^1$, we have

1. $\tilde{\mathbf{X}}_{j_0}$ is measurable with respect to $\sigma(\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1})\vee\sigma(\mathbf{X}_{j_0})\vee\sigma(\tilde{U}_{j_0}^1)$;
2. $\tilde{\mathbf{X}}_{j_0}$ is independent of $\sigma(\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1})$;
3. $\tilde{\mathbf{X}}_{j_0}$ has the same distribution as $\mathbf{X}_{j_0}$;
4. $\mathbb{P}(\|\tilde{\mathbf{X}}_{j_0}-\mathbf{X}_{j_0}\|\ge\tilde{\epsilon})\le\mathbb{E}(\|\tilde{\mathbf{X}}_{j_0}-\mathbf{X}_{j_0}\|)/\tilde{\epsilon}\le\tau_{\tilde{q}+1}/\tilde{\epsilon}$ by Markov's inequality.

For each $j_0\in Q_3^1$, $\tilde{U}_{j_0}^1$ is independent with $\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1}$ and $\mathbf{X}_{j_0}$. In addition, since there are at least $\tilde{q}$ number of matrices between $\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1}$ and $\mathbf{X}_{j_0}$ by our construction, we have $\tau\{\sigma(\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1}),\mathbf{X}_{j_0};\|\cdot\|\}\le\tau_{\tilde{q}+1}$. Note that $\{\tilde{\mathbf{X}}_j\}_{j\in Q_3^1}$ is independent with $\{\tilde{\mathbf{X}}_j\}_{j\in Q_1^1}$ but not mutually independent within the set Q_3^1.

Following the induction steps similar to the previous step and without redundancy, we obtain

$$\begin{aligned}
&\mathbb{E}\,\mathrm{tr}\exp\Big(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\Big)\\
&\le\sum_{i_1=0}^{1}\tilde{A}_{i_1}\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^{i_1}2t\Big(\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^1}\tilde{\mathbf{X}}_j+\sum_{i=2}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\Big)\Big\},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\epsilon}&:=(4pt)^{-\frac{1}{2}}\{\exp(2tq)\tau_{\tilde{q}+1}\}^{\frac{1}{2}},\\
\tilde{L}_1&:=\frac{1}{2}(4pt)^{\frac{1}{2}}q\exp(2tq\tilde{\epsilon})\{\exp(2tq)\tau_{\tilde{q}+1}\}^{\frac{1}{2}},\\
\tilde{L}_2&:=(4pt)^{\frac{1}{2}}q\{\exp(2tq)\tau_{\tilde{q}+1}\}^{\frac{1}{2}},\\
\tilde{A}_{i_1}&:=\binom{1}{i_1}(1+\tilde{L}_1+\tilde{L}_2)^{1-i_1}(\tilde{L}_1)^{i_1},
\end{aligned}$$

This completes the base case.

Iterating the above calculation, we arrive at the following bound:

$$\begin{aligned}
&\mathbb{E}\,\mathrm{tr}\exp\Big(2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\mathbf{X}_j\Big)\\
&\le\sum_{i_1=0}^{1}\cdots\sum_{i_{2^J}=0}^{1}\Big(\prod_{r=1}^{2^J}\tilde{A}_{i_r}\Big)\mathbb{E}\,\mathrm{tr}\exp\Big\{(-1)^{\sum_{r=1}^{2^J}i_r}2t\sum_{i=1}^{2^J}\sum_{k=1}^{m_{q,\tilde{q}}+1}\sum_{j\in Q_{2k-1}^i}\tilde{\mathbf{X}}_j\Big\},
\end{aligned}\tag{C.6}$$

where $\{\tilde{\mathbf{X}}_j\}_{j\in Q_{2k-1}^i}$ for $(i,k)\in\{1,\dots,2^J\}\times\{1,\dots,m_{q,\tilde{q}}+1\}$ are mutually independent and identically distributed as $\{\mathbf{X}_j\}_{j\in Q_{2k-1}^i}$ for $(i,k)\in\{1,\dots,2^J\}\times\{1,\dots,m_{q,\tilde{q}}+1\}$, and

$$\begin{aligned}
\tilde{\epsilon}&:=(4pt)^{-\frac{1}{2}}\{\exp(2tq)\tau_{\tilde{q}+1}\}^{\frac{1}{2}},\\
\tilde{L}_1&:=\frac{1}{2}(4pt)^{\frac{1}{2}}q\exp(2tq\tilde{\epsilon})\{\exp(2tq)\tau_{\tilde{q}+1}\}^{\frac{1}{2}},\\
\tilde{L}_2&:=(4pt)^{\frac{1}{2}}q\{\exp(2tq)\tau_{\tilde{q}+1}\}^{\frac{1}{2}},\\
\tilde{A}_{i_r}&:=\binom{1}{i_r}(1+\tilde{L}_1+\tilde{L}_2)^{1-i_r}(\tilde{L}_1)^{i_r}.
\end{aligned}$$

Let $\tilde{q} := [2/t] \wedge [q/2]$. $\{\tilde{\mathbf{X}}_j\}_{j \in Q^i_{2k-1}}$ for $(i,k) \in \{1,\dots,2^J\} \times \{1,\dots,m_{q,\tilde{q}}+1\}$ are mutually independent with mean $\mathbf{0}$ and $2^J \sum_{k=1}^{m_{\tilde{q},q}+1} \mathrm{card}(Q^i_{2k-1}) \le B$. Moreover by Weyl's inequality, for $(i,k) \in \{1,\dots,2^J\} \times \{1,\dots,m_{q,\tilde{q}}+1\}$, we have

$$2\lambda_{\max}\Big(\sum_{j \in Q^i_{2k-1}} \tilde{\mathbf{X}}_j \Big) \le 2\tilde{q} \le \frac{4}{t}.$$

By Lemma 4 in Banna et al. (2016), we obtain

$$\mathbb{E}\,\mathrm{tr}\exp\Big(2t \sum_{i=1}^{2^J} \sum_{k=1}^{m_{q,\tilde{q}}+1} \sum_{j \in Q^i_{2k-1}} \tilde{\mathbf{X}}_j \Big) \le p \exp\{4h(4)Bt^2\nu^2\}, \tag{C.7}$$

$$\mathbb{E}\,\mathrm{tr}\exp\Big(-2t \sum_{i=1}^{2^J} \sum_{k=1}^{m_{q,\tilde{q}}+1} \sum_{j \in Q^i_{2k-1}} \tilde{\mathbf{X}}_j \Big) \le p \exp\{4h(4)Bt^2\nu^2\}. \tag{C.8}$$

Plugging (C.7) and (C.8) into (C.6) and using the fact that $\sum_{i_r=0}^{1} \tilde{A}_{i_r} = 1 + 2\tilde{L}_1 + \tilde{L}_2$, we obtain

$$\mathbb{E}\,\mathrm{tr}\exp\Big(2t \sum_{i=1}^{2^J} \sum_{k=1}^{m_{q,\tilde{q}}+1} \sum_{j \in Q^i_{2k-1}} \mathbf{X}_j \Big) \le (1 + 2\tilde{L}_1 + \tilde{L}_2)^{2^J} p \exp\{4h(4)Bt^2\nu^2\}. \tag{C.9}$$

By replacing $\mathbf{X}$ by $-\mathbf{X}$, we obtain

$$\mathbb{E}\,\mathrm{tr}\exp\Big(-2t \sum_{i=1}^{2^J} \sum_{k=1}^{m_{q,\tilde{q}}+1} \sum_{j \in Q^i_{2k-1}} \mathbf{X}_j \Big) \le (1 + 2\tilde{L}_1 + \tilde{L}_2)^{2^J} p \exp\{4h(4)Bt^2\nu^2\}. \tag{C.10}$$

Combining (C.5) with (C.9) and (C.10), we get

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}\exp\Big(t \sum_{j \in K_B} \mathbf{X}_j\Big) &\le \sum_{i_1=0}^{2^0} \cdots \sum_{i_J=0}^{2^{J-1}} \Big[\Big(\prod_{k=1}^{J} A_{k,i_k}\Big)(1 + 2\tilde{L}_1 + \tilde{L}_2)^{2^J} p \exp\{4h(4)Bt^2\nu^2\}\Big] \\ &= \Big\{\prod_{k=1}^{J}(1 + 2L_{k,1} + L_{k,2})^{2^{k-1}}\Big\}(1 + 2\tilde{L}_1 + \tilde{L}_2)^{2^J} p \exp\{4h(4)Bt^2\nu^2\},\end{aligned} \tag{C.11}$$

where the last equality follows by $\sum_{i_k=1}^{2^{k-1}} A_{k,i_k} = (1 + 2L_{k,1} + L_{k,2})^{2^{k-1}}$.

By using $\log(1+x) \le x$ for $x \ge 0$, we have

$$\log \mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{j\in K_B}\mathbf{X}_j\Big) \le \sum_{k=1}^{J} 2^{k-1}(2L_{k,1}+L_{k,2}) + 2^J(2\tilde{L}_1+\tilde{L}_2) + \log[p\exp\{4h(4)Bt^2\nu^2\}]. \tag{C.12}$$

For simplicity, we denote $I = \sum_{k=1}^{J} 2^{k-1}(2L_{k,1}+L_{k,2})$, $II = 2^J(2\tilde{L}_1+\tilde{L}_2)$ in (C.12).

Step III: Following calculations similar to Banna et al. (2016), we obtain

$$I \le \frac{32\sqrt{2}}{\log 2}\Big[1+\exp\Big\{\frac{1}{\sqrt{2p}}\exp\Big(-\frac{\psi_2}{16t}\Big)\Big\}\Big]\frac{t^2}{\psi_2}\exp\Big(-\frac{\psi_2}{32t}\Big). \tag{C.13}$$

and

$$II \le 128\Big[1+\exp\Big\{\frac{1}{\sqrt{p}}\exp\Big(-\frac{\psi_2}{32t}\Big)\Big\}\Big]\frac{t^2}{\psi_2}\exp\Big(-\frac{\psi_2}{64t}\Big). \tag{C.14}$$

Hence by combining (C.4), (C.12), (C.13) and (C.14), we obtain for $0 < t \le \min\{1, \frac{\psi_2}{8\log(\psi_1 B^6 p)}\}$,

$$\begin{aligned}&\log \mathbb{E}\,\mathrm{tr}\exp\Big(t\sum_{j\in K_B}\mathbf{X}_j\Big)\\ &\le \log p + 4h(4)Bt^2\nu^2 + 151\Big[1+\exp\Big\{\frac{1}{\sqrt{p}}\exp\Big(-\frac{\psi_2}{64t}\Big)\Big\}\Big]\frac{t^2}{\psi_2}\exp\Big(-\frac{\psi_2}{64t}\Big).\end{aligned}$$

This completes the proof of Lemma 21. □