"The interplay of lifecycle, history and generation – or age, period and cohort, in humbler terms – is one of the great puzzles in the study of individual and social change. This landmark collection is a major contribution on a topic that reaches from philosophy to social statistics. Don't be overtaken by time: start reading now!"
– Professor David Voas, University College London, UK

"Understanding age, period, and cohort effects is a mainstream concern in all the social sciences. But researchers have struggled with how exactly to approach 'APC' issues. This book provides invaluable, up-to-date methodological guidance."
– Professor Malcolm Fairbrother, Umeå University, Sweden

AGE, PERIOD AND COHORT EFFECTS

Statistical Analysis and the Identification Problem

Edited by Andrew Bell

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

8

BAYESIAN AGE–PERIOD–COHORT MODELS

Ethan Fosse

Introduction

Researchers in a wide range of fields have long sought to understand social and cultural change by identifying the unique contributions of age, period and cohort (APC) processes on various outcomes (Ryder 1965). The basic idea is that any temporal change can be attributed to three kinds of processes: (1) changes over the life course of individuals, or *age effects*;[1] (2) changes due to the events in particular years, or *period effects*; (3) changes due to the replacement of older cohorts of individuals with younger ones with different characteristics, or *cohort effects*. However, in what has been called the *APC identification problem* (Mason and Fienberg 1985a, 1985b), the linear effects of an APC model cannot be uniquely estimated due to the perfect linear dependency among the age, period and cohort variables. Intuitively, once we know a person's age and the year of measurement (or period), then we also know that person's birth year (or cohort). A variety of approaches have been proposed to deal with the APC identification problem, but the great majority of studies have taken a frequentist rather than Bayesian perspective.

As I discuss in this chapter, the main advantage of the Bayesian framework for APC analysis is that the analyst can, in principle, explicitly incorporate theoretical considerations, qualitative judgments and additional data into the inferential process. In the frequentist tradition, researchers have long noted the importance of integrating what the political scientist Philip Converse (1976) dubbed "side information" into an APC analysis, typically in the form of constraints on the parameters. However, the Bayesian paradigm, which places specification of the prior distribution front and center, underscores that it is more appropriate to talk about the importance of incorporating *primary information* into any APC analysis. The task ahead, then, is to develop a general Bayesian APC model that allows researchers to easily and explicitly incorporate such primary information into their analyses.

Doing so will require specifying what the statistician Paul Gustafson (2015: 15–18) calls a *transparent reparameterization* of an APC model, which renders clear the impact of nonidentifiability on one's conclusions.

The rest of this chapter is organized as follows. First, I introduce the basics of the Bayesian approach to inference. Second, I outline the classical APC (or C-APC) regression model in terms of the Bayesian perspective, comparing it with the classical (or frequentist) approach. Third, I discuss the identification problem from a Bayesian perspective using a transparent reparameterization of the C-APC model. Fourth, I delineate a general framework for Bayesian APC modeling. In doing so, I review previous Bayesian approaches, which have focused primarily on developing reliable forecasts and applying mechanical, one-size-fits-all, prior distributions. In addition, I discuss the specification of prior distributions in APC analysis, outlining a typology of four main kinds of priors. Next, using a transparently reparameterized model, I illustrate a Bayesian approach to APC analysis by examining the temporal effects of political party identification in the United States. Finally, I conclude with suggestions for future research, discussing how a Bayesian approach to the APC identification challenge places theoretical considerations to the fore.

Basics of the Bayesian approach

Before discussing Bayesian APC models, I first outline the basics of the Bayesian approach to inference.[2] Suppose we have an $n \times 1$ column vector of data $\mathbf{y} = (y_1, \ldots, y_n)^T$, where the superscript T denotes the transpose and an unknown parameter θ. For example, $\mathbf{y}$ could represent the values of a socioeconomic index in a sample of n individuals and θ could represent the average socioeconomic index in a population of individuals. Let $p(.)$ denote a probability distribution and $p(.|.)$ denote a conditional probability distribution. We can further define a data (or sampling) distribution $p(\mathbf{y}|\theta)$, which gives the probability distribution of $\mathbf{y}$ conditional on θ under an assumed parametric model. The typical goal is to obtain a reasonable estimate (or set of estimates) of the unknown parameter.

In the frequentist (or classical) tradition typically adopted by social scientists, the data $\mathbf{y}$ is treated as a random variable while the parameter θ is viewed as a single fixed, unknown quantity. As a result, in the frequentist perspective it makes little sense to talk about a probability distribution for the parameter given the data. Instead, to obtain an estimate of θ, we can define a likelihood function $p(\mathbf{y}|\theta) = \mathcal{L}(\theta|\mathbf{y})$, where $\mathcal{L}(\theta|\mathbf{y})$ produces the likelihood of the parameter given fixed values of the data (Aster et al. 2018).[3] For many of the possible values of θ, the likelihood function will produce output values very close to zero because these values of θ are unlikely to have generated the observed data $\mathbf{y}$. However, for other values of θ the likelihood function will produce considerably larger output values, indicating that the corresponding values of θ are much more likely to have generated the observed data. In fact, by maximizing the likelihood function we obtain the value of θ most likely to have produced the observed data, or what is known as the Maximum Likelihood Estimate (MLE) of θ. The MLE is usually calculated using numerical optimization methods, but in simpler cases it can be derived analytically.[4]

In contrast to the frequentist tradition, in the Bayesian perspective both the data $\mathbf{y}$ and parameter θ are viewed as random variables (Gill 2008).[5] Accordingly, our goal is to make an informed statement about $p(\theta \mid \mathbf{y})$, or the probability of the parameter given the data under a particular assumed parametric model. To obtain this distribution, we need to make use of Bayes' theorem, which is derived using basic probability theory. Because both $\mathbf{y}$ and θ are random variables, we can write out the joint probability of $\mathbf{y}$ and θ, or $p(\theta, \mathbf{y})$. This joint distribution can be factorized as

$$p(\theta, \mathbf{y}) = \sum_{\text{all } \theta} p(\theta)\, p(\mathbf{y} \mid \theta). \tag{8.1}$$

Moreover, because we can write $p(\theta, \mathbf{y}) = p(\mathbf{y})\, p(\theta \mid \mathbf{y})$, after substitution and rearranging terms we obtain Bayes' theorem:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta) p(\theta)}{p(\mathbf{y})}, \tag{8.2}$$

where $p(\mathbf{y} \mid \theta)$ is the likelihood, $p(\theta)$ is the prior distribution and $p(\theta \mid \mathbf{y})$ is the posterior distribution, which is a probability density function if θ is continuous and a probability mass function if θ is discrete. The denominator, $p(\mathbf{y})$, is simply the unconditional (or marginal) probability distribution of the data.[6] If θ is a discrete parameter, then we must sum over all possible values of θ to find the unconditional distribution of the data:

$$p(\mathbf{y}) = \sum_{\text{all } \theta} p(\theta)\, p(\mathbf{y} \mid \theta). \tag{8.3}$$

Alternatively, as is more commonly the case, if θ is continuous then we must use integration to find the unconditional distribution of the data:

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \theta) p(\theta) d\theta. \tag{8.4}$$

The quantity $p(\mathbf{y})$ can be viewed as a normalizing constant, the purpose of which is to ensure that the posterior distribution integrates (or sums) to 1 as required by the definition of a probability density (or mass) function. Because $p(\mathbf{y})$, which does not depend on θ, provides no information about which values of θ are more or less likely, the denominator is often omitted when displaying Bayes' theorem, which can be represented compactly as

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta) p(\theta) \tag{8.5}$$

or

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}, \tag{8.6}$$

where $\propto$ means "proportional to."[7] Equations 8.5 and 8.6 define the core machinery of Bayesian data analysis, showing how the likelihood $p(\mathbf{y} \mid \theta)$ can be "inverted" to produce the posterior $p(\theta \mid \mathbf{y})$. The general procedure is as follows.[8] Before we observe the data $\mathbf{y}$, we express our beliefs about the values of the parameter of a particular model using a prior distribution $p(\theta)$. The possible values of θ define what is known as the parameter space of our model. We then use the likelihood $p(\mathbf{y} \mid \theta)$ to update our prior beliefs, thereby producing a posterior distribution, or $p(\theta \mid \mathbf{y})$.

After obtaining the posterior, we can summarize its distribution to make meaningful conclusions about the probable values of the parameter θ. For example, a researcher can report the expected value, or $\mathrm{E}[\theta \mid \mathbf{y}]$, and the variance, or $\mathrm{Var}(\theta \mid \mathbf{y})$, of the posterior distribution. One can also select various quantiles. Most commonly, for example, researchers cut the posterior distribution at the 2.5% and 97.5% quantiles to construct a Bayesian credible interval (Gill 2008).[9] Alternatively, one might find the interval of minimum length that contains some specified probability level, or what is called the highest posterior density (HPD) interval (Gill 2008: 48–51). Also, because $p(\theta)$ is a distribution, one could present the findings graphically, displaying a range of parameter values and their corresponding probabilities. I now turn to a discussion of the basic APC model in terms of both frequentist and Bayesian perspectives.

The classical APC regression model

It is common in the APC literature to use index notation to keep track of the dimensions of a temporal data structure (Fienberg and Mason 1985: 67–71). We will let $i = 1,\ldots,I$ represent the age groups, $j = 1,\ldots,J$ the period groups, and $k = 1,\ldots,K$ the cohort groups with $k = j - i + I$ and $K = I + J - 1$.[10] Using this index notation, temporal effects in an age–period array can be represented using the classical APC (C-APC) model, also known as the multiple classification model (Mason et al. 1973: 243) or accounting model (Fienberg and Mason 1985: 46–47, 67), which has the following generic form (Fienberg and Mason, 1985: 67–68):[11]

$$Y_{ijk} = \mu + \alpha_i + \pi_j + \gamma_k + \varepsilon_{ijk} \tag{8.7}$$

where Y_{ijk} is the outcome variable to be explained, μ is the intercept, α_i represents the ith age effect, π_j represents the jth period effect, γ_k represents the kth cohort effect, and ε_{ijk} is the error term. The errors are assumed to be additive, independent and identically distributed (iid) according to a normal distribution with a mean zero and variance σ^2, such that $\varepsilon_{ijk} \sim N(0, \sigma^2)$. To avoid overparameterization, we apply the so-called "usual constraints" that the parameters sum to zero, or $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \pi_j = \sum_{k=1}^{K} \gamma_k = 0$.[12]

The parameterization shown in Equation 8.7 is very flexible, allowing the age, period and cohort effects to be highly nonlinear because there is one parameter for each age, period and cohort category (Mason et al. 1973: 246). This can be seen in Table 8.1, which shows how each cell is represented by a unique combination of parameters.

For ease of exposition it is convenient to represent the C-APC compactly using matrix notation:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{8.8}$$

where $\mathbf{X}$ is an $(I \times J) \times (I + J + K - 2)$ design matrix with a leading vector of 1s for the constant, β is an $(I \times J) \times 1$ vector of parameters to be estimated, $\mathbf{y}$ is an $(I \times J) \times 1$ vector of outcome values, and ϵ is an $(I \times J) \times 1$ vector of errors. As noted previously, the errors are assumed to be additive, independent and identically distributed (iid) according to a normal distribution with a mean zero and variance σ^2, such that $\epsilon \sim N(0, \sigma^2\mathbf{I})$, where 0 is an $(I \times J) \times 1$ vector of zeros, and $\mathbf{I}$ is an $(I \times J) \times (I \times J)$ diagonal matrix.

In the frequentist tradition, we can obtain estimates of the parameters by maximizing the likelihood function. For Equation 8.8, the likelihood function, or the probability of the data given the parameter vector β and scalar σ^2 as well as the input variables $\mathbf{X}$, is given by:

$$p(\mathbf{y} \mid \beta, \sigma^2, \mathbf{X}) = \mathcal{L}(\beta, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^{(I \times J)} \times e^{\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right]}, \tag{8.9}$$

where e and π are the conventional constants. Under the standard assumptions of the normal linear model, maximizing the likelihood function in Equation 8.9 results in estimates that coincide with those of ordinary least squares (OLS), such

TABLE 8.1 Classical APC model on an age–period array

Age groups	*Period groups*		
	j=1	*j=2*	*j=3=J*
i=1	$\mu+\alpha_1+\pi_1+\gamma_5+\epsilon_{1,1,5}$	$\mu+\alpha_1+\pi_2+\gamma_6+\epsilon_{1,2,6}$	$\mu+\alpha_1+\pi_3+\gamma_7+\epsilon_{1,3,7}$
i=2	$\mu+\alpha_2+\pi_1+\gamma_4+\epsilon_{2,1,4}$	$\mu+\alpha_2+\pi_2+\gamma_5+\epsilon_{2,2,5}$	$\mu+\alpha_2+\pi_3+\gamma_6+\epsilon_{2,3,6}$
i=3	$\mu+\alpha_3+\pi_1+\gamma_3+\epsilon_{3,1,3}$	$\mu+\alpha_3+\pi_2+\gamma_4+\epsilon_{3,2,4}$	$\mu+\alpha_3+\pi_3+\gamma_5+\epsilon_{3,3,5}$
i=4	$\mu+\alpha_4+\pi_1+\gamma_2+\epsilon_{4,1,2}$	$\mu+\alpha_4+\pi_2+\gamma_3+\epsilon_{4,2,3}$	$\mu+\alpha_4+\pi_3+\gamma_4+\epsilon_{4,3,4}$
i=5=*I*	$\mu+\alpha_5+\pi_1+\gamma_1+\epsilon_{5,1,1}$	$\mu+\alpha_5+\pi_2+\gamma_2+\epsilon_{5,2,2}$	$\mu+\alpha_5+\pi_3+\gamma_3+\epsilon_{5,3,3}$

that $\hat{\beta}_{\text{MLE}} = \hat{\beta}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where the superscript -1 denotes a regular inverse.[13]

The Bayesian C-APC model

From the Bayesian perspective our goal is to make a statement about the probability of the parameters given the data. In other words, we want to summarize the posterior distribution $p(\beta, \sigma^2 \mid \mathbf{y}, \mathbf{X})$. However, the design matrix $\mathbf{X}$ and outcome vector $\mathbf{y}$ are both commonly considered "the data". This implies a full Bayesian model includes not only the parameters β and σ^2 linked to the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$, but also another set of parameters ψ linked to the unconditional distribution of $\mathbf{X}$ (Gelman et al. 2014: 354). Using Bayes' theorem, this in turn suggests the following set-up (see Jackman 2009: 99–103; Trader 2014: 2):

$$p(\beta, \sigma^2, \psi \mid \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}, \mathbf{X} \mid \beta, \sigma^2, \psi)p(\beta, \sigma^2, \psi)}{p(\mathbf{y}, \mathbf{X})}, \tag{8.10}$$

where $p(\beta, \sigma^2, \psi \mid \mathbf{y}, \mathbf{X})$ is the posterior distribution, $p(\mathbf{y}, \mathbf{X} \mid \beta, \sigma^2, \psi)$ is the joint likelihood, $p(\beta, \sigma^2, \psi)$ is the prior distribution, and $p(\mathbf{y}, \mathbf{X})$ is a normalizing constant.

However, Equation 8.10 does not encode the typical model of interest among social scientists. Specifically, the Bayesian version of the C-APC model does not entail estimating a set of parameters for the joint distribution of $\mathbf{y}$ and $\mathbf{X}$; rather, as a regression model, it involves estimating parameters for the distribution of the outcome $\mathbf{y}$ conditional on the variables in $\mathbf{X}$. However, under standard regression assumptions, ψ provides no additional information about the parameters β and σ^2 after conditioning on $\mathbf{X}$. Accordingly, we can factorize Equation 8.10 into two distinct, independent parts:

$$\begin{aligned} p(\beta, \sigma^2, \psi \mid \mathbf{y}, \mathbf{X}) &= p(\beta, \sigma^2 \mid \mathbf{y}, \mathbf{X})p(\psi \mid \mathbf{X}) \\ &= \frac{p(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2)p(\beta, \sigma^2)}{p(\mathbf{y} \mid \mathbf{X})} \times \frac{p(\mathbf{X} \mid \nu)p(\psi)}{p(\mathbf{X})}. \end{aligned} \tag{8.11}$$

Equation 8.11 tells us that we can independently focus on one of two analyses. Either we can make inferences about β and σ^2 based on the conditional distribution of $\mathbf{y}$ given $\mathbf{X}$ or we can make inferences regarding ψ using the unconditional distribution of $\mathbf{X}$. Our interest lies in the former. Noting that $p(\psi \mid \mathbf{X}) = [p(\mathbf{X} \mid \psi)p(\psi)] / p(\mathbf{X})$, we can write Bayes' formula for the C-APC as:

$$p(\beta,\sigma^2 \mid \mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X},\beta,\sigma^2)p\left(\beta,\sigma^2\right)}{p(\mathbf{y} \mid \mathbf{X})}, \tag{8.12}$$

where $p(\beta,\sigma^2 \mid \mathbf{y},\mathbf{X})$ is the posterior distribution, $p(\mathbf{y} \mid \mathbf{X},\beta,\sigma^2)$ is the likelihood, $p\left(\beta,\sigma^2\right)$ is the prior distribution and $p(\mathbf{y} \mid \mathbf{X})$ is a normalizing constant. Because conditioning on the design matrix $\mathbf{X}$ is implicit in our model and the denominator is just a normalizing constant, Equation 8.12 is often simplified as:

$$p(\beta,\sigma^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \beta,\sigma^2)p\left(\beta,\sigma^2\right), \tag{8.13}$$

where as before $\propto$ means "proportional to" (cf. Equations 8.5 and 8.6). This is just another way of stating that our prior beliefs about the parameters, encoded in terms of a prior distribution, can be updated by the likelihood function, which arises from the statistical model and the data. The result of this updating is a posterior distribution, which represents our "post-data" beliefs about the parameters (Gustafson 2015: 6).

To obtain Bayesian estimates we must specify our beliefs about the parameters in the form of a prior distribution. Because there are many types of distributions, each of which can take any number of forms, there are many ways to set up our model. Using "stacked" notation (e.g., see McElreath 2018: 124), which explicitly reveals our distributional assumptions, one way to set up the Bayesian version of the C-APC model is as follows:

$$Y_{ijk} \sim \mathrm{N}\left(\hat{Y}_{ijk},\sigma\right) \tag{8.14}$$

$$\hat{Y}_{ijk} = \mu + \sum_{i}^{I-1}\alpha_i A_i + \sum_{j}^{J-1}\pi_j P_j + \sum_{k}^{K-1}\gamma_k C_k \tag{8.15}$$

$$\mu \sim \mathrm{N}\left(\mu_\mu,\sigma_\mu^2\right) \tag{8.16}$$

$$\begin{aligned}
\alpha_i &\sim \mathrm{N}\left(\mu_{\alpha_i},\sigma_{\alpha_i}^2\right) \quad \text{for} \quad i = 1,\ldots,I-1 \\
\pi_j &\sim \mathrm{N}\left(\mu_{\pi_j},\sigma_{\pi_j}^2\right) \quad \text{for} \quad j = 1,\ldots,J-1 \\
\gamma_k &\sim \mathrm{N}\left(\mu_{\gamma_k},\sigma_{\gamma_k}^2\right) \quad \text{for} \quad k = 1,\ldots,K-1
\end{aligned} \tag{8.17}$$

$$\sigma \sim \mathrm{Uniform}\left(a_\sigma,b_\sigma\right), \tag{8.18}$$

where A_i, P_j and C_k are sum-to-zero effect (or deviation) contrasts, with the last of each age, period and cohort category dropped. The corresponding parameters $\alpha_i,\ldots,\alpha_{I-1},\pi_j,\ldots,\pi_{J-1}$ and $\gamma_k,\ldots,\gamma_{K-1}$ give the age, period and cohort deviations, respectively, from the overall (or grand) mean, captured by the intercept μ. Equation 8.14 represents the likelihood and Equation 8.15 represents the model, while Equations 8.16–8.17 outline the prior distributions for the model parameters. In this specification of the C-APC model I use normal priors for the intercept and deviation parameters while I use a uniform prior for the variance.[14] The Bayesian version of the C-APC model outlined above is general, requiring the researcher to input center and spread values for the normal priors as well as lower and upper values for the uniform prior. Any number of input values are possible, reflecting the wide range of possible prior beliefs about the parameters. An initial approach might be to use what are known, somewhat incorrectly, as "noninformative" (i.e., diffuse or flat)[15] priors for the parameters, which would typically produce Bayesian estimates comparable to those from MLE or OLS in the frequentist tradition. However, the Bayesian estimates generated by the C-APC model are highly sensitive to the choice of priors, even those deemed to be "noninformative." As a consequence, researchers naively estimating the model outlined in Equations 8.14–8.18 may fail to realize that their findings are driven, in no small part, by their prior beliefs rather than the data and model. It is for this reason that I do not recommend that researchers use the C-APC model to estimate temporal effects, even in the context of a Bayesian analysis. To better understand why I advise against using the C-APC model, I now turn to the issue of model identification from a Bayesian perspective.

The identification problem

A model is unidentified[16] when multiple values of one or more of the parameters correspond to the same distribution of observed data (Gustafson 2015: 1–2). When a model is identified, there is a one-to-one mapping between the distribution of the observed data and the parameter space, or the set of possible values of the parameters of a model. Accordingly, altering the values of the parameters will in general change the distribution of the data. In contrast, when a model is unidentified, there is a one-to-many mapping between the data and the parameter space. As such, it is possible to alter the values of at least one of the parameters and yet keep the distribution of the data unchanged. Furthermore, in the absence of identification, this one-to-many mapping between the data and the parameter space remains even as the sample size goes to infinity.

As is well known, the C-APC model in Equation 8.7 suffers from a fundamental identification problem due to linear dependence in the columns of the design matrix (Fosse and Winship 2018; 2019a). Algebraically, at least one of the columns of $\mathbf{X}$ can be rewritten as a function of the other columns, such that the design matrix $\mathbf{X}$ is rank deficient one (i.e., singular).[17] Accordingly, a regular inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exist and the model lacks a unique set of parameter estimates. In other words, there is a one-to-many mapping between the data and the parameter space. More formally, the set of parameters of the C-APC model can be altered without affecting the likelihood. Viewing the likelihood as a "hill," the identification problem

corresponds to a "ridge" in the likelihood. This is a direction in the parameter space in which the likelihood is flat, extending from negative to positive infinity, no matter how large the sample (Gelman 2014: 89). In the frequentist tradition, researchers have typically dealt with the lack of identification by applying a constraint, such as setting two adjacent age groups equal to each other or dropping one of the temporal dimensions altogether. However, as a number of scholars have pointed out (e.g., Fosse and Winship 2019a; O'Brien 2015; Yang and Land 2013), such constraints can rely on quite strong, often untenable, theoretical assumptions.

In the Bayesian approach, it has been suggested that nonidentification is not inherently a problem because a prior must be specified (see Gelfand and Sahu 1999; Neath and Samaniego 1997; Poirier 1998). As the statistician Dennis Lindley (1972: 46) has noted, "unidentifiability causes no real difficulties in the Bayesian approach." That is, as long as the researcher specifies a legitimate probability distribution as the prior, then Bayes' theorem will produce the posterior in terms of a legitimate probability distribution. For example, using the Bayesian version of the C-APC model outlined in Equations 8.14–8.18, one can choose some set of values for the priors, "turn the Bayesian crank" and generate a set of Bayesian estimates, which can then be summarized in the conventional way (Gustafson 2015: 5–6). This is true even though the C-APC model is not identified. However, an unidentified Bayesian model is not without the potential for misuse, especially if the parameterization obscures the flow of information, as is the case with the C-APC model. As Gustafson (2015) has forcefully argued, although the choice of parameterization in an unidentified model is mathematically arbitrary, some parameterizations are more useful than others for clarifying the influence of the prior distribution on the posterior distribution (see also Gustafson 2005, 2009). In particular, he makes the case that nonidentified models should be expressed in terms of a transparent parameterization, which clearly separates those parameters directly informed by the observed data from those that are, so far as possible, indirectly informed by the data.

Transparent reparameterization of the classical APC model

In this section I outline a transparent reparameterization of the C-APC model that will clarify the underlying assumptions of the model. Let α, π and γ denote the true, unknown linear effects. Similarly, let α^*, π^* and γ^* denote some other set of values of the linear effects. We can write the identification problem algebraically as (Fosse and Winship 2018: 316):

$$\alpha^* = \alpha + v, \quad \pi^* = \pi - v, \quad \text{and} \quad \gamma^* = \gamma + v, \tag{8.19}$$

where v is some unknown scalar. Setting a range of values for v traces out what is known as the *canonical solution line*, which is the set of possible values of the linear effects consistent with the data (Fosse and Winship 2018: 313–319). This line corresponds with the maximum likelihood (or OLS) estimates in the frequentist tradition. The canonical solution line lies in a three-dimensional parameter space,

where the dimensions are the values of the age, period and cohort linear effects ranging from negative to positive infinity.[18]

The linearized APC model

Researchers typically fit some version of the C-APC model, which, under sum-to-zero constraints, is expressed in terms of deviations (or "effects") relative to a grand (or overall) mean. However, an equivalent representation entails decomposing each "effect" in the C-APC into their respective linear and nonlinear effects. Fosse and Winship (2019b) call this reparameterized model the linearized APC (or L-APC) model (see also Chapter 6).

The ith age effect in the C-APC can be represented in the L-APC model in terms of an overall linear age effect along with a unique parameter for the ith age nonlinearity: $\alpha_i = \left(i - i^*\right)\alpha + \tilde{\alpha}_i$, where α is the age linear effect, $\tilde{\alpha}_i$ is the ith age nonlinear effect, and the asterisk denotes a midpoint or referent index $i^* = \left(I+1\right)/2$ (which ensures that the sum-to-zero constraints are satisfied). In other words, each age effect α_i is split into the sum of a common parameter α representing the age slope for the entire array, with a value shifting across rows (or age categories) as a function of the age index i, and a unique parameter $\tilde{\alpha}_i$, which is a nonlinearity specific to each age category. One can similarly decompose each period effect as $\pi_j = \left(j - j^*\right)\pi + \tilde{\pi}_j$ with $j^* = \left(J+1\right)/2$ and each cohort effect as $\gamma_k = \left(k - k^*\right)\gamma + \tilde{\gamma}_k$ with $k^* = \left(K+1\right)/2$. The L-APC model thus has the following general form:

$$Y_{ijk} = \mu + \alpha\left(i - i^*\right) + \pi\left(j - j^*\right) + \gamma\left(k - k^*\right) + \tilde{\alpha}_i + \tilde{\pi}_j + \tilde{\gamma}_k + \epsilon_{ijk}. \tag{8.20}$$

The L-APC model is based on a design matrix in which the linear and nonlinear components are in some way orthogonal to each other. There is a variety of ways to set up this design matrix, but orthogonal polynomial contrasts give parameter results that are easiest to interpret (for details, see Fosse and Winship 2018).

It is important to underscore that, because each of the "effects" in the C-APC can be decomposed into their respective linear and nonlinear parts, the parameters from the L-APC and C-APC are fundamentally equivalent. That is, the L-APC parameter vector is simply a decomposed version of the C-APC, such that

$$\begin{aligned}\beta &= (\mu, \alpha^*, \pi^*, \gamma^*, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_{I-1}, \tilde{\pi}_j, \ldots, \tilde{\pi}_{J-1}, \tilde{\gamma}_k, \ldots, \tilde{\gamma}_{K-1})^T \\ &= (\mu, \alpha + v, \pi - v, \gamma + v, \tilde{\alpha}_i, \ldots, \tilde{\alpha}_{I-1}, \tilde{\pi}_j, \ldots, \tilde{\pi}_{J-1}, \tilde{\gamma}_k, \ldots, \tilde{\gamma}_{K-1})^T.\end{aligned} \tag{8.21}$$

I will make use of this identity in the next section when I separate the C-APC model into identified and unidentified components. Note that the presence of the

scalar v indicates some of the parameters are not identified. In particular, some set of slopes $\alpha^{\star}$, $\pi^{\star}$ and $\gamma^{\star}$ will correspond to the true parameter slopes α, π and γ only if v is exactly specified. This is equivalent to stating that the analyst needs to apply an exactly correct just-identifying constraint to recover the true, unknown values of the temporal effects. However, because the intercept and nonlinear effects do not include the scalar v, they are identified.

Transparent reparameterization of the classical model

To construct a transparent reparameterization of the C-APC, it is necessary to split the parameter vector into identified and unidentified components. To do so, I will take advantage of the L-APC parameterization discussed in the previous section. Formally let $\beta = (\xi, \lambda) = h(\beta)$, where β is the original C-APC parameterization (see Equation 8.7), ξ is some set of identified parameters, λ is an unidentified parameter, and $h(\beta)$ is the transparent reparameterization. There are several ways to define ξ and λ – and thus, by definition, $h(\beta)$. One approach is to drop the period linear component from the design matrix of the L-APC model, which results in the corresponding identified parameter vector:[19]

$$\xi = (\mu, \alpha^{\star}, \gamma^{\star}, \tilde{\alpha}_i, \ldots, \tilde{\alpha}_{I-1}, \tilde{\pi}_j, \ldots, \tilde{\pi}_{J-1}, \tilde{\gamma}_k, \ldots, \tilde{\gamma}_{K-1})^T. \tag{8.22}$$

Simple algebra can be used to show that ξ is, in fact, identified. To demonstrate this, note that dropping the period linear component from the design matrix is equivalent to stating that $\pi^{\star} = 0$ or, equivalently, $\pi - v = 0$. Accordingly, we know that, in the identified parameter vector above, $v = \pi$. Using the fact that $\alpha^{\star} = \alpha + v$ and $\gamma^{\star} = \gamma + v$, after plugging in $v = \pi$ we can thus express the identified parameter vector as

$$\xi = (\mu, \theta_1, \theta_2, \tilde{\alpha}_i, \ldots, \tilde{\alpha}_{I-1}, \tilde{\pi}_j, \ldots, \tilde{\pi}_{J-1}, \tilde{\gamma}_k, \ldots, \tilde{\gamma}_{K-1})^T, \tag{8.23}$$

where, in the terminology of Fosse and Winship (2019b), $\theta_1 = \alpha + \pi$ and $\theta_2 = \gamma + \pi$. Because dropping the linear component is equivalent to constraining $\pi^{\star} = 0$, we can express the unidentified parameter as $\lambda = \pi = v$, where v is an unknown scalar.

To summarize the foregoing, we can write the transparent reparameterization of the C-APC model as $h(\beta) = (\xi, \lambda)$, where ξ is defined in Equation 8.23 and $\lambda = v$. This transparent reparameterization of the C-APC model has two fundamental properties (Gustafson 2015). First, the distribution of the outcome given the design matrix depends only on the identified parameter vector ξ, not on the unidentified parameter λ. In other words, the likelihood with and without the unidentified parameter is the same, such that $p(\mathbf{y} \mid \mathbf{X}, \xi, \lambda, \sigma^2) = p(\mathbf{y} \mid \mathbf{X}, \xi, \sigma^2)$. Second, regular parametric asymptotic theory applies to the model represented by $p(\mathbf{y} \mid \mathbf{X}, \xi, \sigma^2)$. In other words, the estimate of ξ converges in probability to its true value as the sample size increases to infinity. With these two insights, we are now ready to examine the identification problem within a Bayesian framework.

Bayesian interpretation of the APC identification problem

Using the transparent reparameterization outlined previously, we can write out the identification problem in Bayesian terms. Recall that we can write the posterior distribution of the Bayesian version of the C-APC model as follows:

$$p(\beta, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2) p(\beta, \sigma^2)}{p(\mathbf{y} \mid \mathbf{X})}. \tag{8.24}$$

To make sense of the identification problem from a Bayesian perspective, it is useful to re-express the parameter vector β into identified and unidentified parts. The transparent reparameterization discussed previously informs us that we can substitute (ξ, λ) for β, where λ is an unidentified parameter and ξ is a vector of identified parameters. Two important relationships follow from this reparameterization. First, we can write the likelihood as:

$$p(\mathbf{y} \mid \mathbf{X}, \xi, \lambda, \sigma^2) = p(\mathbf{y} \mid \mathbf{X}, \xi, \sigma^2), \tag{8.25}$$

which is another way of stating that the unidentified parameter λ is not itself informative about the likelihood. Second, we can decompose the prior distribution as:

$$p(\xi, \lambda, \sigma^2) = p(\xi, \sigma^2) p(\lambda \mid \xi, \sigma^2), \tag{8.26}$$

where $p(\xi, \sigma^2)$ is the prior for the identifiable parameters and $p(\lambda \mid \xi, \sigma^2)$ is the prior for the unidentified parameter, conditional on the identified parameters. Using these identity relationships, we can thus split the model into two separate parts, one of which is identified and the other which is not (see Gustafson 2015; Nielsen and Nielsen 2014; Poirier 1998):

$$p(\xi, \sigma^2 \mid \mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \xi, \sigma^2) p(\xi, \sigma^2)}{p(\mathbf{y} \mid \mathbf{X})} \tag{8.27}$$

and

$$p(\lambda \mid \mathbf{y}, \mathbf{X}, \xi, \sigma^2) = p(\lambda \mid \xi, \sigma^2), \tag{8.28}$$

where $p(\xi, \sigma^2 \mid \mathbf{y}, \mathbf{X})$ is the posterior of the identified parameters and $p(\lambda \mid \mathbf{y}, \mathbf{X}, \xi, \sigma^2)$ is the posterior of the unidentified parameter, conditional on the identified parameters. Equation 8.27 tells us that the identified part acts as a standard Bayesian

model, such that the prior $p(\xi,\sigma^2)$ is updated by the likelihood $p(\mathbf{y} \mid \mathbf{X},\xi,\sigma^2)$ to produce the posterior $p(\xi,\sigma^2 \mid \mathbf{y},\mathbf{X})$. In contrast, Equation 8.28 reveals that the conditional prior of the unidentified parameter, or $p(\lambda \mid \xi,\sigma^2)$, is not updated by a likelihood function. In fact, regardless of the sample size, the conditional posterior distribution of the unidentified parameter equals its conditional prior distribution. It also follows that the predictive distribution does not depend on the conditional prior for λ (see Nielsen and Nielsen 2014: 8).

The implications of the Bayesian interpretation of the APC identification problem can be further clarified by restricting our attention to the linear effects (e.g., see Fosse and Winship 2018). Focusing just on the linear effects of the C-APC, this means that $\beta = (\xi,\lambda) = (\alpha^*,\pi^*,\gamma^*) = (\theta_1,\theta_2,\nu)$. The parameters θ_1 and θ_2, along with some value for ν, are sufficient to derive a set of values for α^*,π^* and γ^*. Crucially, the likelihood is equal to $p(\mathbf{y} \mid \mathbf{X},\beta) = p(\mathbf{y} \mid \mathbf{X},\xi)$ or $p(\mathbf{y} \mid \mathbf{X},\alpha^*,\pi^*,\gamma^*) = p(\mathbf{y} \mid \mathbf{X},\theta_1,\theta_2)$. In other words, the likelihood is a function of θ_1 and θ_2, not the unidentified parameter ν. The prior distribution can be written as $p(\beta) = p(\xi,\lambda) = p(\xi)p(\lambda \mid \xi)$ or $p(\alpha^*,\pi^*,\gamma^*) = p(\theta_1,\theta_2,\nu) = p(\theta_1,\theta_2)p(\nu \mid \theta_1,\theta_2)$. That is, we have a prior for the identifiable parameters, denoted by $p(\theta_1,\theta_2)$, and a prior for the unidentified parameter, conditional on the identified parameters, given by $p(\nu \mid \theta_1,\theta_2)$. Using Bayes' theorem, we can accordingly write the posterior distribution for the identified parameters as:

$$p(\xi \mid \mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X},\xi)p(\xi)}{p(\mathbf{y} \mid \mathbf{X})} \quad \text{or}$$

$$p(\theta_1,\theta_2 \mid \mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y} \mid \mathbf{X},\theta_1,\theta_2)p(\theta_1,\theta_2)}{p(\mathbf{y} \mid \mathbf{X})}, \tag{8.29}$$

which shows that the posterior distribution for θ_1 and θ_2 is a function of the likelihood $p(\mathbf{y} \mid \mathbf{X},\theta_1,\theta_2)$ as well as the prior distribution $p(\theta_1,\theta_2)$. In other words, our prior beliefs about θ_1 and θ_2 are, in fact, updated by the data. In contrast, focusing on the unidentified parameter, we can write:

$$p(\lambda \mid \mathbf{y},\mathbf{X},\xi) = p(\lambda \mid \xi) \quad \text{or} \quad p(\nu \mid \mathbf{y},\mathbf{X},\theta_1,\theta_2) = p(\nu \mid \theta_1,\theta_2), \tag{8.30}$$

which indicates that the conditional prior for the unidentified parameter ν is not updated by the likelihood. That is, given values of θ_1 and θ_2, the data do not modify our prior beliefs about the possible values of ν. As a consequence, the choice of the prior is absolutely critical in determining the estimates of the linear APC effects. To avoid arbitrary results, APC analysts should, whenever possible, use a transparent reparameterization of the underlying model, basing their priors on careful, theoretically informed decisions.

Even though one's prior beliefs are crucial in determining the results of the C-APC model, the data are still informative about the parameters, even those that are unidentified (Gustafson 2015: 18–22). Specifically, the transparent reparameterization

of the C-APC model reveals that the nonlinear effects are directly informed by the observed data, while the linear effects are indirectly informed by the data via the values of θ_1 and θ_2. Given a set of estimated values for θ_1 and θ_2, we will have a restricted set of possible estimates of the linear effects (Fosse and Winship 2018). That is, only a subset of possible values of the slopes are actually consistent with the observable data. Furthermore, in the limit of an infinite amount of data, the posterior distribution can be described as a *limiting posterior distribution*, with a point mass at some value of θ_1 and θ_2 combined with a conditional prior distribution for the unidentified parameter ν (Gustafson 2015: 16–39). In other words, as the sample size goes to infinity, the estimates of θ_1 and θ_2, which are directly informed by the data, become concentrated at a single point, providing maximal information on the set of possible combinations of the slopes. Thus, even if the only parameters of interest are the linear effects, in general a larger sample size will be preferable to a smaller sample size.

The transparent reparameterization outlined above clarifies the APC identification problem from a Bayesian perspective. Placing prior distributions over the parameters of the C-APC model has the potential to mislead researchers. The reason is that, in both the frequentist and Bayesian versions of the C-APC model, virtually the entire parameter vector is unidentified, because each estimated effect (i.e., each deviation from the grand mean) is composed of both linear and nonlinear effects. An unfortunate consequence is that the influence of the prior on the parameter estimates is unclear, because the prior is only partly updated by the likelihood.[20] In contrast, the reparameterized model splits the parameter vector into a set of identifiable parameters and an unidentified parameter, elucidating the central role of the prior in Bayesian APC models.

Introducing the linearized Bayesian APC model

The previous sections discussed how one can reparameterize the C-APC model into the L-APC model, which can then be used to illuminate the identification problem from a Bayesian perspective. Besides clarifying the identification problem, we can also use the L-APC model's parameterization to set up a linearized Bayesian APC (L-BAPC) model. As with the conventional (i.e., non-Bayesian) L-APC model, the design matrix is one that simply separates the linear from the nonlinear components. Prior distributions are then placed over the parameters, with particular attention to priors placed over the linear effects in light of their particular sensitivity to the specified priors. There are a wide range of possibilities for setting up the L-APC model. For example, one way to specify the L-BAPC model is as follows (again using "stacked" notation):

$$Y_{ijk} \sim \mathrm{N}\left(\hat{Y}_{ijk},\sigma\right) \tag{8.31}$$

$$\hat{Y}_{ijk} = \mu + \alpha A_L + \pi P_L + \gamma C_L + \sum_{2}^{I-1}\tilde{\alpha}_i A_i + \sum_{2}^{J-1}\tilde{\pi}_j P_j + \sum_{2}^{K-1}\tilde{\gamma}_k C_k \tag{8.32}$$

$$\mu \sim \mathrm{N}\left(\mu_\mu, \sigma_\mu^2\right) \tag{8.33}$$

$$\alpha \sim \mathrm{Uniform}\left(a_\alpha, b_\alpha\right)$$

$$\pi \sim \mathrm{Uniform}\left(a_\pi, b_\pi\right) \tag{8.34}$$

$$\gamma \sim \mathrm{Uniform}\left(a_\gamma, b_\gamma\right)$$

$$\tilde{\alpha}_i \sim \mathrm{N}\left(\mu_{\tilde{\alpha}_i}, \sigma_{\tilde{\alpha}_i}^2\right) \quad \text{if} \quad i = 2,\ldots,l-1$$

$$\tilde{\pi}_j \sim \mathrm{N}\left(\mu_{\tilde{\pi}_j}, \sigma_{\tilde{\pi}_j}^2\right) \quad \text{if} \quad j = 2,\ldots,m-1 \tag{8.35}$$

$$\tilde{\gamma}_k \sim \mathrm{N}\left(\mu_{\tilde{\gamma}_k}, \sigma_{\tilde{\gamma}_k}^2\right) \quad \text{if} \quad k = 2,\ldots,n-1$$

$$\tilde{\alpha}_i \sim \mathrm{Laplace}\left(\mu_{\tilde{\alpha}_i}, \sigma_{\tilde{\alpha}_i}^2\right) \quad \text{if} \quad i = l,\ldots,I-1$$

$$\tilde{\pi}_j \sim \mathrm{Laplace}\left(\mu_{\tilde{\pi}_j}, \sigma_{\tilde{\pi}_j}^2\right) \quad \text{if} \quad j = m,\ldots,J-1 \tag{8.36}$$

$$\tilde{\gamma}_k \sim \mathrm{Laplace}\left(\mu_{\tilde{\gamma}_k}, \sigma_{\tilde{\gamma}_k}^2\right) \quad \text{if} \quad k = n,\ldots,K-1$$

$$\sigma \sim \mathrm{Uniform}\left(a_\sigma, b_\sigma\right), \tag{8.37}$$

where A_L, P_L and C_L represent the linear components and A_i, P_i and C_i represent the nonlinear components. The linear and nonlinear components are represented in terms of sum-to-zero orthogonal polynomial contrasts. For example, A_2 denotes a quadratic orthogonal polynomial age contrast, A_3 a cubic orthogonal polynomial age contrast, A_4 a quartic orthogonal polynomial age contrast, and so on. As with the C-APC model, the overall (or grand) mean is captured by the intercept μ. Equation 8.31 denotes the likelihood and Equation 8.32 denotes the model, while Equations 8.33–8.37 outline the prior distributions for the model parameters. In this particular version of the L-BAPC model I use uniform priors for the linear effects and variance. However, as I discuss below, many other options are available, which is an especially important consideration for the linear effects. When using orthogonal polynomials it is desirable to set more restrictive priors for the higher-order polynomials, which arguably just capture noise (or relatively minor fluctuations in the data). I define higher-order (versus lower-order) polynomials by some cut-off levels l, m and n for age, period and cohort, respectively. For the lower-order polynomials I use normal priors, while for the higher-order polynomials

I use Laplace distributions, also known as double-exponential distributions. The Laplace distribution can be specified so that there is a spiked concentration near zero, acting to "shrink" the coefficients.

Typology of priors

The priors outlined in the L-BAPC model above are just one set of distributions that could be placed over the parameters. Depending on the choice of prior distributions the L-BAPC model can be used to encode a wide range of explicit theoretical assumptions. In general, there are four main kinds of priors one can place over the linear and nonlinear effects: (1) proxy variables; (2) variable selection; (3) smoothing; and (4) bounding. This typology not only can help guide APC model-building, but also elucidates how a flexible Bayesian framework maps onto existing APC models coming out of the frequentist tradition. The distinctions between these priors are not hard-and-fast, however, and in practice multiple kinds of priors can be used in the same L-BAPC model.[21] For example, one could easily incorporate all four main kinds of priors by simply altering the input values of the priors in the L-BAPC model outlined in Equations 8.31–8.37. I review each of these types of priors in turn.

First, there are priors that represent information about proxy variables (or mechanisms). The conventional proxy variables approach involves replacing age, period or cohort with another variable thought to represent an underlying mechanism, such as relative cohort size in lieu of cohort (O'Brien, 1989). A more sophisticated version of the proxy variables technique is the mechanism-based approach advocated by Fosse and Winship (2019a), which entails specifying multiple mechanisms between one or more of the temporal variables and the outcome (see also Winship and Harding 2008). Using the L-BAPC model, the proxy variables approach implies placing informative priors on the linear and nonlinear parameters based on knowledge of specific mechanisms, possibly using estimates from previous studies or other datasets.

Second, variable selection priors can be used to encode beliefs that the parameter of interest is at or near zero. This is not an uncommon assumption. In fact, as noted by Fosse and Winship (2019a: 475), the great majority of studies examining social change simply drop one of the temporal variables altogether without explicitly specifying a full APC model. The variable selection approach using the L-BAPC would entail setting extremely strong priors towards zero on the linear and nonlinear effects of age, period or cohort. For example, instead of dropping the period variable, one could specify highly concentrated Laplace distributions at or near zero for the period linear effect and nonlinear effects.

Third, some priors can be used to smooth the temporal effects. In the frequentist tradition, a number of researchers have used age and age-squared instead of a full set of age parameters in an APC model, effectively smoothing the age nonlinearities. In terms of the L-BAPC, a comparable approach would entail, for instance, using highly concentrated Laplace distributions centered around zero for the higher-order

age polynomials. Another way of applying a smoothing restriction in the frequentist tradition is the equality constraints approach, which typically involves grouping some pair of adjacent categories in the C-APC model. This constraint can be interpreted as a kind of smoothing technique in that the overall effects for the two adjacent categories are forced to be the same. An equivalent approach using the L-BAPC would involve, for example, specifying a prior for one of the linear effects with a distribution concentrated at that particular slope value implied by the desired equality constraint (see Fosse and Winship 2019a: 475–476).

Finally, there are priors that reflect bounding analyses. Using theoretically informed sign, size and shape constraints, Fosse and Winship (2019b) demonstrate how researchers can "zero out" parts of the parameter space to set upper and lower limits on one or more of the temporal effects. An equivalent approach using the L-BAPC would be the specification of uniform priors based on various beliefs about the sign, size or shape of the temporal effects. However, a wide range of other distributions are possible and in some cases slightly more exotic distributions might more closely represent existing theoretical knowledge on a topic. For example, instead of using a uniform prior for the age linear effect to encode the belief that the slope ranges from zero to positive infinity with equal probability, one could use a gamma distribution, reflecting the belief that the slope ranges from zero to positive infinity with some decreasing probability (Fosse and Winship 2019b: 2001).

Previous Bayesian APC models

The L-BAPC model presents a transparent parameterization, clarifying which parts of the model are identified and which parts are not, with the goal of using theoretical considerations to place informative prior distributions over some of the parameters. Previous Bayesian APC models have taken a different approach, focusing on either developing a more-or-less general estimator or using Bayesian models for forecasting. By and large, Bayesian APC models have not been widely used in sociology and related fields. Nonetheless, two Bayesian approaches have received significant attention in the APC literature: the Nakamura model and RW-1/RW-2 models (for discussions, see Glenn 2005; Smith and Wakefield 2016).[22] Both of these approaches begin with the C-APC model as the baseline parameterization.

The Bayesian model proposed by the statistician Takashi Nakamura (1986) has been touted as a "mechanical solution" that is applicable in a wide range of applied contexts (Sasaki and Suzuki 1989: 761). As Sasaki and Suzuki (1987: 1063) have claimed: "The Bayesian procedure in Nakamura's new method can provide a satisfactory explanation for the data almost automatically, without the identification specification that has occurred in previous cohort analysis and resulted in misleading findings". The basic idea of Nakamura's approach is that the temporal effects (i.e., deviations from the overall mean) of the C-APC model change relatively gradually, such that first-order differences in the successive effects are "close to zero" (Fukuda 2006; Nakamura 1986).

Specifically, let us define first-order differences $\alpha_i - \alpha_{i+1}$ for age groups $i = 1,\ldots,I-1$, $\pi_j - \pi_{j+1}$ for period groups $j = 1,\ldots,J-1$, and $\gamma_k - \gamma_{k+1}$ for cohort groups $k = 1,\ldots,K-1$. Nakamura's method entails minimizing a weighted sum of squares of the first-order differences of the effects, or the following:

$$\frac{1}{\sigma_\alpha^2}\sum_{i=1}^{I-1}(\alpha_i - \alpha_{i+1})^2 + \frac{1}{\sigma_\pi^2}\sum_{j=1}^{J-1}(\pi_j - \pi_{j+1})^2 + \frac{1}{\sigma_\gamma^2}\sum_{k=1}^{K-1}(\gamma_k - \gamma_{k+1})^2, \tag{8.38}$$

where σ_α^2, σ_π^2 and σ_γ^2 are hyperparameters (Fukuda 2006, 2007, Miller and Nakamura 1996, 1997). Given values for these hyperparameters, the parameter vector β of the C-APC model can be estimated by the mode of the posterior distribution proportional to

$$p(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2)p(\beta^* \mid \sigma_\alpha^2, \sigma_\pi^2, \sigma_\gamma^2, \sigma^2), \tag{8.39}$$

where β^* is the parameter vector β excluding the intercept. To select values of the hyperparameters σ_α^2, σ_π^2 and σ_γ^2, Nakamura uses a fit statistic known as the Akaike Bayesian Information Criterion (ABIC), which in this case is defined as

$$\text{ABIC} = -2\ln\int p(\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2)p(\beta^* \mid \sigma_\alpha^2, \sigma_\pi^2, \sigma_\gamma^2, \sigma^2)d\beta^* + 2h, \tag{8.40}$$

where h is the number of hyperparameters and again β^* is the parameter vector β excluding the intercept (see Akaike 1998).

As with any APC approach that attempts to separate out all three time effects, Nakamura's technique is only as valid as its theoretical assumptions. There are two caveats regarding the Nakamura method (see also the criticisms by Glenn 1989, 2005). First, it is assumed to be a one-size-fits-all, mechanical estimator by at least some of its proponents (e.g., Sasaki and Suzuki 1987). However, only theoretical knowledge or additional data can justify the assumptions implied by the model. In particular, the claim that the parameters change gradually might be valid for some aging processes, but wholly contrary to basic assumptions for period and cohort-related processes, where abrupt discontinuities may be expected. Second, the technique is based on the C-APC parameterization and, as such, the flow of information is obscured. In light of the fact that only the intercept and nonlinear effects are identified, the prior distribution is imposing potentially strong assumptions on the linear effects.

A related set of Bayesian models has been developed based on random walk smoothing priors (Besag et al. 1995; Berzuini and Clayton 1993; Havulinna 2014; Knorr-Held and Rainer 2001; Schmid and Held 2004; Smith and Wakefield 2016). Typically these models have been used for extracting more reliable forecasts from APC models, essentially sidestepping the identification issue (e.g., see Bray et al.

2001; Havulinna 2014; Riebler and Held 2017; Schmid and Held 2007). The first-order random walk (RW1) prior penalizes deviations from a constant, stochastically shrinking the first-order differences towards zero, while the second-order random walk (RW2) prior penalizes deviations from a linear effect, stochastically restricting the second-order differences towards zero (Havulinna 2014: 847). The RW1 prior corresponds to the following for, say, the age effects:

$$\alpha_i \mid \alpha_i,\ldots,\alpha_1 \sim \mathrm{N}\left(\alpha_{i-1},\sigma_\alpha^2\right), \quad \text{for} \quad i=2,\ldots,I, \tag{8.41}$$

with a uniform prior for the first age effect, α_1. In contrast, the RW2 prior corresponds to:

$$\alpha_i \mid \alpha_i,\ldots,\alpha_1 \sim \mathrm{N}\left(2\alpha_{i-1}-\alpha_{i-2},\sigma_\alpha^2\right), \quad \text{for} \quad i=3,\ldots,I, \tag{8.42}$$

with independent uniform priors for the first and second age effects, α_1 and α_2. More generally, the random walk smoothing prior has the following form:

$$p(\alpha \mid \kappa_\alpha) \propto \kappa_\alpha^{\left(\frac{I-1}{2}\right)} e^{\left(-\frac{1}{2}\alpha^T \kappa_\alpha \mathbf{R}\alpha\right)}, \tag{8.43}$$

where $\alpha=(\alpha_i,\ldots,\alpha_I)^T$ is a column vector of age effects; κ_α is the precision, or inverse of the variance for age (i.e., $1/\sigma_\alpha^2$); e is the well-known constant; and $\mathbf{R}$ is a so-called "structure matrix" of dimension $I \times I$ that reflects some specified neighborhood structure depending on whether a RW1 or RW2 prior is desired (for examples, see Rue and Held 2005).[23] The precision κ_α is an estimated parameter that determines the degree of smoothing: the higher the precision (i.e., the lower the variance), the smoother the corresponding set of estimated temporal effects.

There are two main advantages to using models with random walk smoothing priors. First, smoothing the temporal effects is desirable because the extreme categories of age and cohort tend to have relatively few observations. As a result, in absence of smoothing, the estimated effects can fluctuate wildly. Second, smoothing is desirable for the purposes of social forecasting. In general, researchers have found that a model with RW2 priors gives more reliable forecasts than one with RW1 priors (e.g., see Smith and Wakefield 2016). The reason is that the estimated effects from the RW2 model tend to be smoother than those from the RW1 model, thereby generating projections that are less dependent on local variation in the data. The primary limitation of the APC literature on RW1 and RW2 models is that there have been few attempts to directly incorporate informative priors for the linear effects, which constitute the crux of the identification problem. Moreover, most APC studies using random walk smoothing priors have focused on the C-APC model rather than a reparameterized version that clearly differentiates those parameters that are identified from those that are not (for an exception, see Smith and Wakefield 2016: 603–608).

Example: political party strength

For the purposes of illustrating the utility of the L-BAPC with informative priors, I now turn to an examination of APC effects on political party strength (Converse 1976). The data consists of $n = 51{,}956$ respondents from the U.S. General Social Survey (GSS). Age and period are grouped into five-year intervals. The outcome variable captures political party strength, with higher values indicating greater strength, and lower values less strength.[24] Specifically, the variable is calculated by assigning a numerical value to one of four groups: 0 = independent, 1 = lean independent, 2 = weak party affiliation, 3 = strong party affiliation. This is identical to the coding used by the political scientist Philip Converse (1976: 166). For simplicity of exposition, and to parallel Converse's analysis, I assume the outcome is continuous.[25]

The joint estimated effects of age, period and cohort on party strength are shown in Figures 8.1 and 8.2. The number in each cell of Figure 8.1 and the surface height of Figure 8.2 indicate the predicted average party strength from a C-APC model with an arbitrary equality constraint. Note that the predicted means are identified, such that the C-APC model will generate the same set of predicted values regardless of the just-identifying constraint. The pattern of averages in Figures 8.1 and 8.2 suggests that all three temporal effects are operating, with age playing a particularly dominant role. However, these conclusions are tentative at best. Extreme care should be taken when interpreting the pattern of effects in Figures 8.1 and 8.2. Due to the linear dependency of the three time scales one cannot, from these visualizations alone, determine the unique contributions of age, period and cohort on party strength.

I next estimated the L-BAPC in Equations 8.31–8.37 using Markov Chain Monte Carlo (MCMC) techniques (for a detailed discussion, see Gelman et al. 2014: 275–292). For all models I used three chains, which is conventional in Bayesian modeling (McElreath 2018: 356–357). For all results reported here, standard diagnostic measures and visual output indicated convergence was achieved across chains. For instance, Gelman and Rubin's potential scale reduction factor, or $\hat{R}$, was near 1 for all parameters (Gelman et al. 2014: 285). Similarly, traceplots showed random scatter around an average value, indicating that the chains were "mixing" well. I placed noninformative prior distributions over the intercept and variance. For the quadratic, cubic, quartic and quintic polynomial parameters I used noninformative normal priors, while for the remaining higher-order polynomials I used Laplace priors concentrated around zero. As discussed previously, this set-up for the polynomials helps to reduce noise in the tails of age and cohort, which tend to be quite sparse. To derive informative priors for the linear effects, I considered three main sources of information: previously published results, theoretical claims

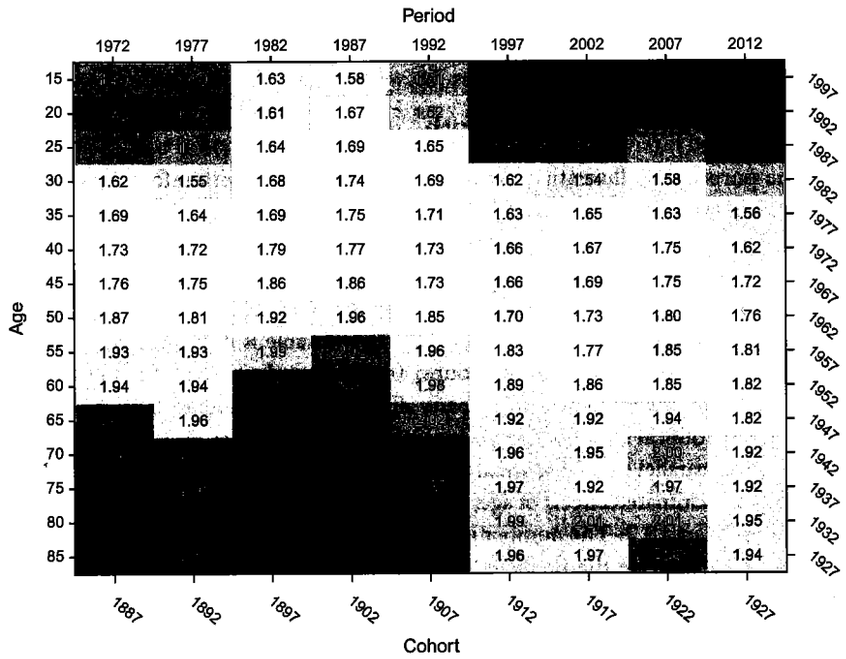

FIGURE 8.1 Joint estimated effects of age, period and cohort on party strength

in the literature on party strength, and qualitative judgments elicited from one or more subject matter experts.

First, regarding previously published results, Dassonneville (2017) examines the relationship between aging and party strength. The author fits a hierarchical age–period–cohort (HAPC) model, which tends to fix the cohort linear effect near zero (Fosse and Winship 2019a: 477–479).[26] An equivalent approach using the L-BAPC model entails placing a strong prior near zero for the cohort slope and noninformative priors for the age and period slopes. Accordingly, I used a highly concentrated Laplace distribution centered on zero for the cohort slope and diffuse normal priors for the age and period slopes. The results are shown in Figures 8.3 and 8.4. Figure 8.3 displays the posterior distributions for the linear and quadratic effects, while Figure 8.4 shows the overall estimated APC effects. For each distribution in Figure 8.3, a thick vertical line denotes the mean, and the shaded area the 95% credible interval. These results, mirroring the assumptions embedded in Dassonneville's model, reveal that party strength increases dramatically across the life course. However, these findings rely on the extremely strong assumption that the cohort effects exhibit trendless fluctuation, which may be called into question on a priori grounds.

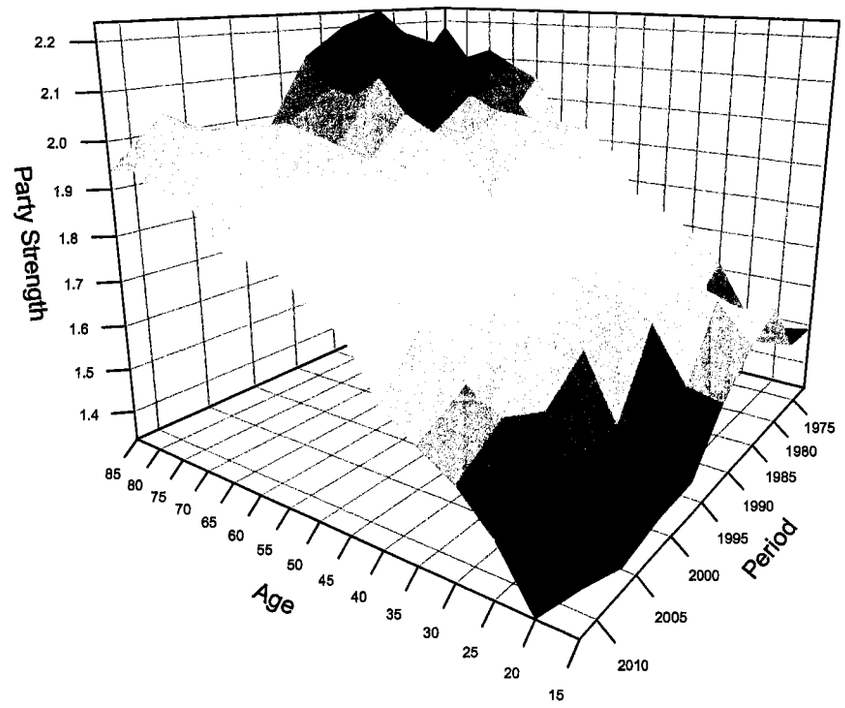

FIGURE 8.2 Joint estimated effects of age, period and cohort on party strength

A second source for constructing informative priors is sociological or political theory. Converse (1976) outlines a cognitive-behavioral argument for why partisanship is likely to increase with age (see also Converse 1969). In essence, he argues that partisan strength increases monotonically with age because people accumulate particular patterns of voting. The act of voting for one party more than another, even if initially by chance, will develop over time into a sustained preference for one party over another. Based on Converse's theory, the aging effect is approximated by a minimum monotonically increasing quadratic curve (see Figure 8.2 in Converse 1976: 44). To derive a prior for the age slope, I first ran the L-BAPC model with arbitrary priors on the slopes and noninformative priors on the remaining parameters. I estimated the age nonlinear effects using only the quadratic term, reflecting the smoothness of Converse's hypothetical age curve. Next, I used a monotonicity constraint to find the age slope corresponding to the minimum monotonically increasing set of age effects, where the effects are based only on the linear and quadratic terms (for details, see Fosse and Winship 2019b: 1989–1994). I then used the resultant age slope (0.234) as the mean in a normal prior for age in the full L-BAPC model, with the variance set to an arbitrarily small value.[27] The findings are shown

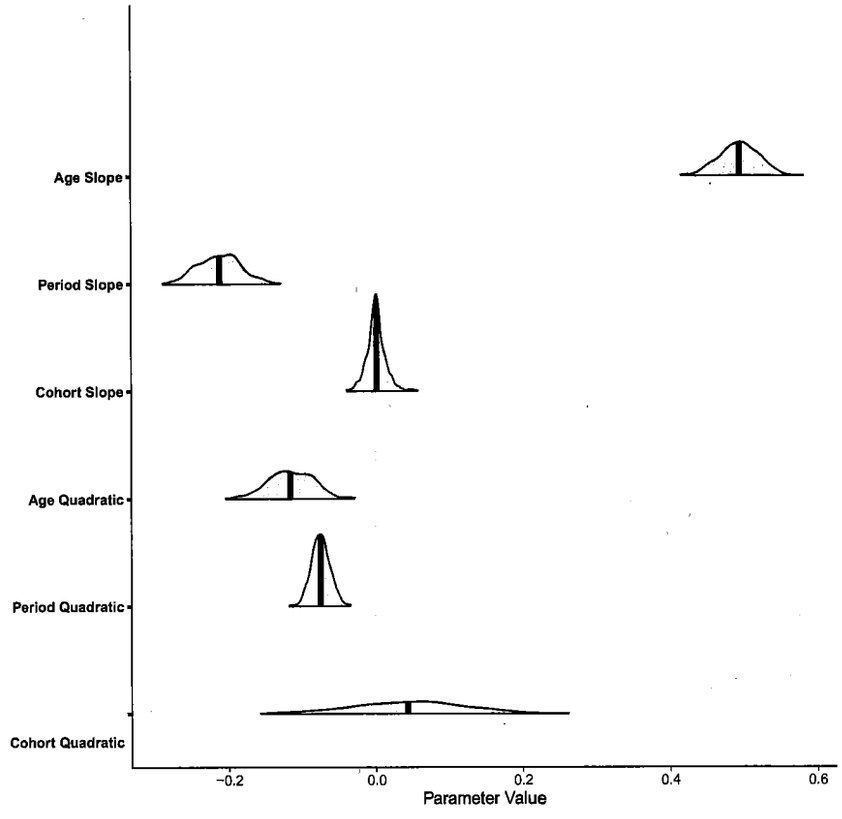

FIGURE 8.3 Posterior distributions for linear and quadratic effects, from L-BAPC model with a strong zero prior for the cohort slope

in Figures 8.5 and 8.6. As with the previous example, Figure 8.5 displays the posterior distributions for the linear and quadratic effects, while Figure 8.6 shows the corresponding overall estimated APC effects. Most strikingly, these findings reveal that, given Converse's cognitive-behavioral theory, there has been a steep decline in partisan affiliation across cohorts.

Finally, I elicited expert knowledge to extract a range of values for the age linear effect (Gill and Walker 2005; Kadane and Wolfson 1998; Meyer and Booker 2001). I used a graphical approach to elicit the requisite information, which has been shown to be superior to numerical-based methods of elicitation (Casement and Kahle 2018; Jones and Johnson 2014). Specifically, I recruited a subject matter expert who was knowledgeable on the aging process and life course theory. Next, I showed this expert a set of estimated age effects with the age slope fixed to zero. I then varied the slope parameter, asking the expert to report the most likely age

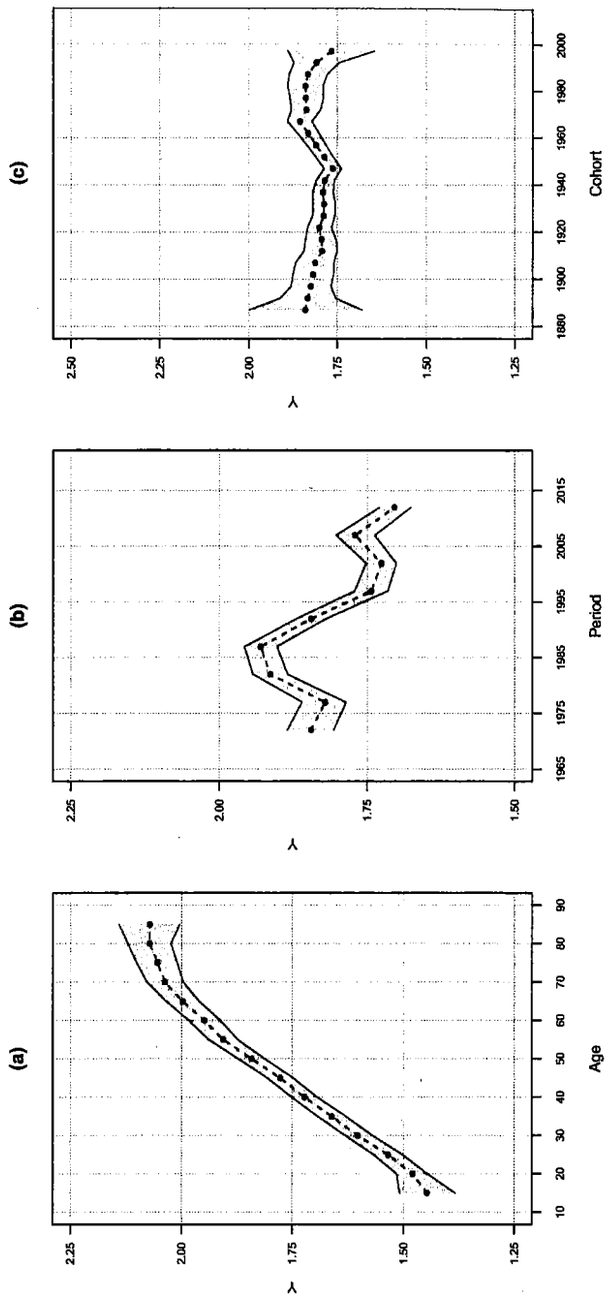

FIGURE 8.4 Predicted APC effects, from L-BAPC model with a strong zero prior for the cohort slope

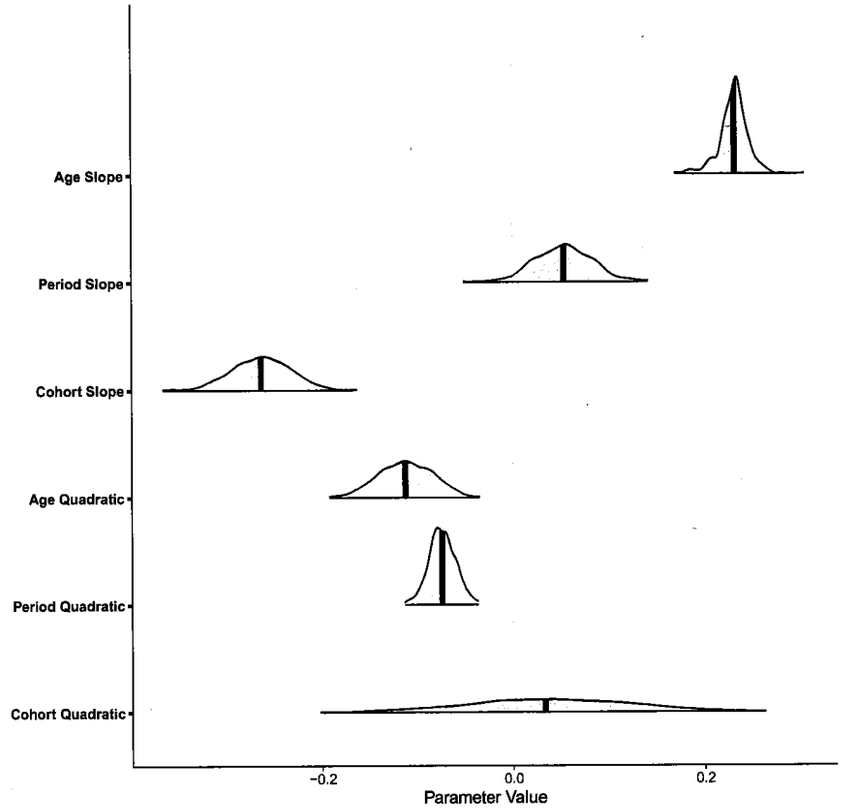

FIGURE 8.5 Posterior distributions for linear and quadratic effects of APC, from L-BAPC model with age assumed to monotonically increase

curve. For example, Figure 8.7 shows a set of different age curves, where the only difference is that the age slope is fixed to a different value. The expert was not shown graphs of the period or cohort effects during elicitation. Once the expert selected a particular curve as the most likely one, I set upper and lower limits of increasing size around it. For each set of limits I asked whether or not the interval contained 95% of the theoretically possible age slopes. From this graphical elicitation I obtained an implied prior on the linear age effect. The estimated APC effects are shown in Figure 8.8. These findings reveal that party strength has declined primarily due to cohort replacement, although there is a somewhat smaller negative period effect as well. Analyses also indicate that as people age they become more partisan, consistent with Converse's claim that younger people in general have not yet formed strong partisan attachments.

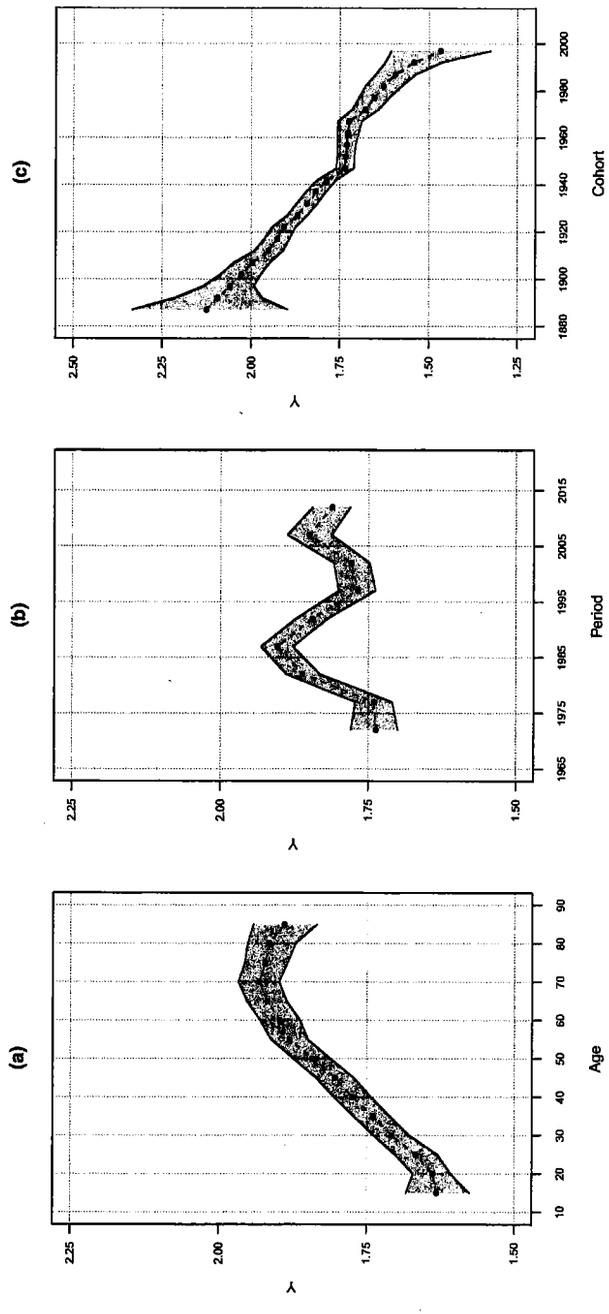

FIGURE 8.6 Predicted APC effects, from L-BAPC model with age constrained to monotonically increase

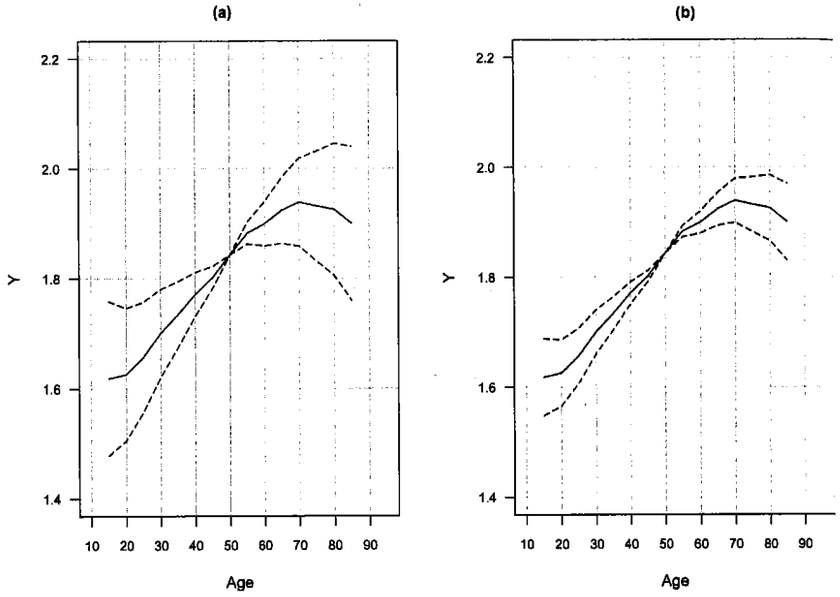

FIGURE 8.7 Age curves under different slope constraints, as shown to a subject matter expert

Conclusion

This chapter outlined a Bayesian perspective on APC modeling, illustrating how a transparent reparameterization can clarify the underlying, sometimes implicit, assumptions of many temporal models. The Bayesian framework can be viewed as a generalization of the constraint-based approach commonly used by APC analysts coming out of the frequentist tradition. In many cases, a Bayesian model using concentrated prior distributions will provide virtually the same point and interval estimates as constraint-based methods. Notwithstanding this, the advantage of the Bayesian framework is that one can arguably use a much more diverse set of prior distributions than those implied by the frequentist approach.

From a modeling perspective, a Bayesian analysis is often not that different from a frequentist approach, but there are some important issues that have prevented Bayesian methods from gaining wider use. For example, to estimate the normalized posterior distribution one needs the unconditional distribution of the data, which typically requires evaluating a high-dimensional integral. In all but the simplest cases, the unconditional distribution of the data has no tractable closed-form solution.[28] As a result, Bayesian inference focuses extensively on the appropriate use of computational procedures, most commonly MCMC methods. This can entail a fairly high upfront cost for the researcher in terms of time and effort. Besides computational issues, additional

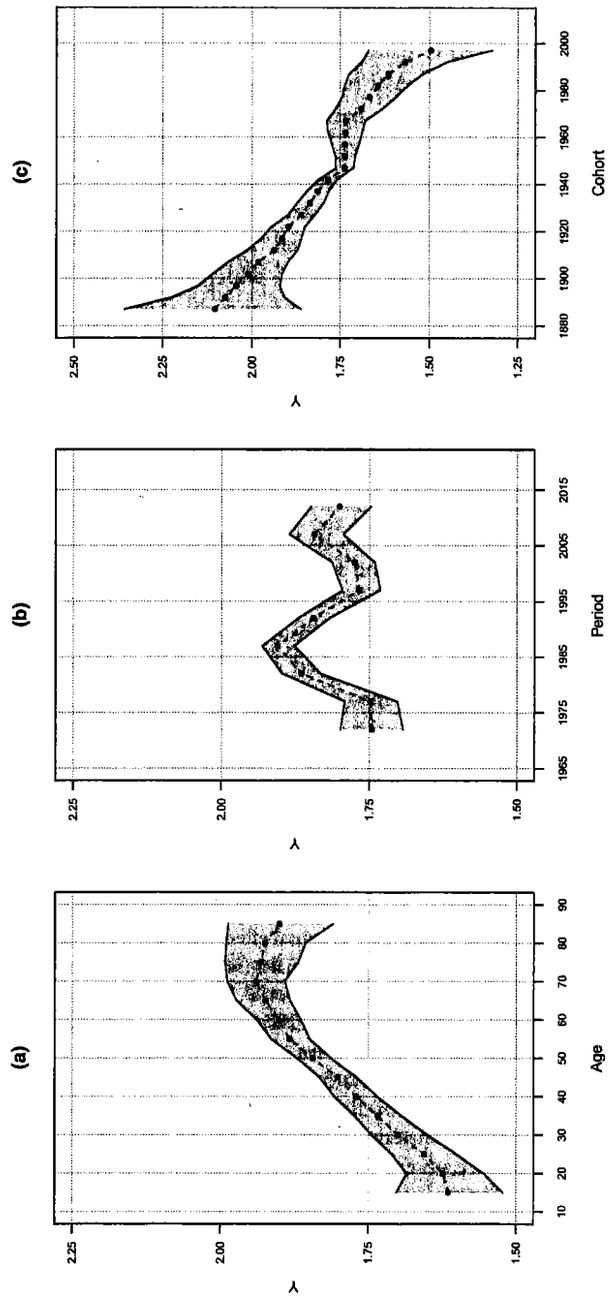

FIGURE 8.8 Predicted APC effects, from model with prior on age slope elicited from a subject matter expert

considerations in the Bayesian framework include, for example, eliciting relevant information for the prior distribution from subject matter experts, choosing an appropriate likelihood function, and succinctly summarizing the posterior distribution.

However, in the case of APC data, additional care is required because of the underlying identification problem. As I have demonstrated, the choice of the prior distribution in APC models typically has a very strong influence over the posterior distribution, because only one part of the data, in fact, induces variation in the likelihood function. Moreover, the influence of the prior on the unidentified parameters does not diminish as the sample size increases. This issue is complicated by the fact that the influence of the prior is often "hidden" due to the way in which APC models are conventionally parameterized, which fails to separate the identified from the unidentified components. Consequently, there is a real danger of researchers using a Bayesian APC model mechanically without fully understanding the extent to which the results rely critically on potentially strong assumptions encoded in the prior distribution.

The recovery of the true, unknown APC effects is only as reliable as the social, biological or cultural theories on which one's assumptions are based. Theories of underlying processes may be fundamentally flawed, thereby leading to mistaken conclusions about APC effects. There is, in this sense, no ultimate resolution of the APC identification problem. Yet, the Bayesian framework is a powerful engine for incorporating additional information, or what I have called *primary information*, into an APC analysis. Future work should consider more carefully the various ways in which such primary information, encoded as prior distributions, can be more effectively elicited and incorporated into APC models.

Online supplementary material relating to this book can be found at www.routledge.com/9780367174439.

Notes

1 Following the convention in the APC literature, I use the shorthand of "effects" when referring to age, period and cohort processes (e.g., Fienberg et al. 1979; Glenn 1981; Mason et al. 1973;). These "effects" need not refer to causal effects in the sense of parameters with well-defined potential outcomes or (counterfactuals) (see Morgan and Winship 2014).
2 For excellent technical overviews, see Gelman et al. (2014: 3–28); Gill (2008: 1–71); Jackman (2009: 3–48); Lynch (2007: 47–76); and Wang et al. (2018: 3–18).
3 It is important to remember that $\mathcal{L}(\theta \mid \mathbf{y})$ is *not* a probability distribution for the parameter θ given the data $\mathbf{y}$. Rather, it is a function that expresses how probable a given set of observations is for different values of the parameter, with the uncertainty deriving not from the fixed (albeit unknown) quantity θ but from the random variable $\mathbf{y}$.
4 For mathematical convenience, instead of a likelihood function, researchers will often use a log-likelihood function, denoted by $\ell(\theta \mid \mathbf{y}) = \log(\mathcal{L}(\theta \mid \mathbf{y}))$. However, conceptually the log-likelihood function presents no additional complications.
5 Even though the parameters are treated as random variables in the Bayesian approach, by convention they are still denoted using Greek letters because they are unknown quantities.
6 Sometimes the denominator is also referred to as the *marginal likelihood* or the *prior predictive distribution* (Gill, 2008: 44).

7 The product $p(\mathbf{y} \mid \theta)p(\theta)$ can be interpreted as the unnormalized posterior distribution. Often a researcher can focus on estimating an unnormalized posterior distribution because, in many but not all cases, the posterior can be renormalized in the final step of the analysis (Gill 2008: 43–44).
8 So far I have focused on inference for a single parameter, but the discussion extends easily to a vector of parameters. Define $\theta = (\theta_1, \ldots, \theta_k)^T$, where k is the number of parameters. With multiple parameters, we can simply refer to a joint prior distribution $p(\theta)$, joint likelihood $p(\mathbf{y} \mid \theta)$ and joint posterior distribution $p(\theta \mid \mathbf{y})$, with $p(\theta \mid \mathbf{y}) \propto p(\theta)p(\mathbf{y} \mid \theta)$.
9 With diffuse or flat priors, the Bayesian credible interval is similar to a 95% confidence interval in the frequentist perspective.
10 Note that I is added to $j - i$ so that the cohort index begins at $k = 1$. This ensures that, for example, $i = j = k = 1$ refers to the first group for all three temporal measures. One could just as easily index the cohorts using $k = j - i$, but this identity would be lost.
11 This model assumes we have aggregated data in a Lexis table. If we have individual-level data, we may also want to index the individuals in the data using $n = 1, \ldots, q$, where q is the sample size. This would lead to a model specified as $Y_{ijkn} = \mu + \alpha_i + \pi_j + \gamma_k + \epsilon_{ijkn}$. For simplicity in the rest of this chapter I assume we are using aggregated data, such that the data has $I \times J$ rows.
12 For the rest of this chapter, I will assume that sum-to-zero constraints are specified, with the last category of each temporal variable dropped. However, a number of other constraints are possible. For example, one could fix the parameters at one of the levels to zero (e.g., $\alpha_{i=1} = \pi_{j=1} = \gamma_{k=1} = 0$ or $\alpha_{i=I} = \pi_{j=J} = \gamma_{k=K} = 0$).
13 Both $\hat{\beta}_{\text{MLE}}$ and $\hat{\beta}_{\text{OLS}}$ give unbiased estimates of the true parameter vector β. However, unlike OLS, maximum likelihood estimation will generate a biased estimate of σ^2, thus requiring a bias correction (Wang et al. 2018: 40).
14 Before the widespread availability of powerful computing, it was especially important to specify what is called a *conjugate* prior (Gill 2008: 54, 111–116). A conjugate prior refers to a prior in which the posterior has the same probability distribution family. Accordingly, there is an analytical solution – an explicit formula – for the posterior distribution expressed in terms of the prior parameters and the data. The main limitation of using conjugate priors is that analytical solutions are typically only feasible for quite simple models (Gelman et al. 2014: 35–36). However, with modern computing techniques there is considerably greater flexibility in modeling choices.
15 All priors are informative in the sense that the analyst is introducing some kind of information into the model.
16 I use the terms "unidentified" and "nonidentified" interchangeably. Gustafson (2015: 4) prefers the phrase "partially identified" to underscore that even when a model is not identified there is often some information to be gleaned from the unidentified parameters.
17 The linear dependence is reflected in the null vector (i.e., the eigenvector with a zero eigenvalue) of the design matrix, which has non-zero elements (e.g., Kupper 1985: 829).
18 Fosse and Winship (2019b) show how the canonical solution line can be visualized using what they call a *2D-APC graph* (1984–1987).
19 Note that the identified parameter vector length is one less than that of the full parameter vector, reflecting the fact that the full design matrix of $\mathbf{X}$ is rank deficient one after applying sum-to-zero constraints.
20 It is common in the APC literature to use a non-transparent parameterization, which can easily lead analysts astray. Two researchers analyzing the same APC data may inadvertently

impose different conditional prior distributions on the unidentified parameter, thereby generating divergent findings.

21 For an excellent overview of the three main kinds of priors used in Bayesian analysis (conjugate, noninformative and informative), see Gill (2008: 135–189).

22 Due to space limitations I do not cover here the hierarchical age–period–cohort (HAPC) model, which has an implicit Bayesian interpretation due to the hierarchical structure of the model. For an explicitly Bayesian implementation of the HAPC, see Yang (2006). For an overview of the assumptions of the HAPC, see Fosse and Winship (2019a: 477–479).

23 Note that the structure matrix is of rank $I-1$ for the RW1 prior and rank $I-2$ for the RW2 prior. The RW1 and RW2 priors are both prominent examples of intrinsic Gaussian Markov random fields (GMRF) (for details, see Rue and Held 2005).

24 Converse (1976: 10–11) distinguishes between the direction of party choice (e.g., Democratic vs. Republican) and strength of party identification regardless of party choice (e.g., "strong" vs. "weak").

25 I obtain similar results treating the outcome as an ordered categorical variable.

26 See Table 7.2 and Figure 7.4 in Dassonneville (2017: 153–154), which indicate a near-zero linear effect for cohort.

27 Unfortunately, Converse's theory does not suggest a spread for the age slope prior distribution.

28 A special case occurs when the prior distribution is a conjugate prior, which enables a relatively simple analytical solution. With a conjugate prior, the prior distribution is chosen so that the likelihood and prior combine to generate a posterior distribution in the same family as the prior distribution.

References

Akaike, Hirotugu (1998). "Likelihood and the Bayes Procedure". In *Selected Papers of Hirotugu Akaike*. Edited by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. New York, NY: Springer, pp. 309–332.

Aster, Richard C., Brian Borchers, and Clifford H. Thurber (2018). *Parameter Estimation and Inverse Problems*. Cambridge, MA: Elsevier.

Berzuini, Carlo (1993). "Bayesian Inference on the Lexis Diagram". *Bulletin of the International Statistical Institute* 50, pp. 149–164.

Besag, Julian et al. (1995). "Bayesian Computation and Stochastic Systems". *Statistical Science* 10.1, pp. 3–41.

Bray, Isabelle, Paul Brennan, and Paolo Bo (2001). "Recent Trends and Future Projections of Lymphoid Neoplasms: A Bayesian Age–Period–Cohort Analysis". Cancer Causes & Control, 12, 813–820.

Casement, Christopher J. and David J. Kahle (2018). "Graphical Prior Elicitation in Univariate Models". *Communications in Statistics – Simulation and Computation* 47.10, pp. 2906–2924.

Converse, Philip E. (1969). "Of Time and Partisan Stability". *Comparative Political Studies* 2.2, pp. 139–171.

——— (1976). *The Dynamics of Party Support: Cohort-Analyzing Party Identification*. Beverly Hills: Sage.

Dassonneville, Ruth (2017). "Age and Voting". In Arzheimer, Kai, Jocelyn Evans, and Michael S. Lewis-Beck, eds. *The SAGE Handbook of Electoral Behaviour*. London: Sage, pp. 137–158.

Fienberg, Stephen E. (1979). "Identification and Estimation of Age–Period–Cohort Models in the Analysis of Discrete Archival Data". *Sociological Methodology* 10, pp. 1–67.

Fienberg, Stephen E. and William M. Mason (1985) "Specification and implementation of age, period and cohort models". In Mason, William M. and Stephen E. Fienberg, eds. *Cohort Analysis in Social Research*. New York: Springer, pp. 45–88.

Fosse, Ethan and Christopher Winship (2018). "Moore–Penrose Estimators of Age–Period–Cohort Effects: Their Interrelationship and Properties". *Sociological Science* 5.14, pp. 304–334.

——— (2019a). "Analyzing Age–Period–Cohort Data: A Review and Critique". *Annual Review of Sociology* 45, pp. 467–492.

——— (2019b). "Bounding Analyses of Age–Period–Cohort Effects". *Demography* 56.5, pp. 1975–2004.

Fukuda, Kosei (2006). "A Cohort Analysis of Female Labor Participation Rates in the U.S. and Japan". *Review of Economics of the Household* 4.4, pp. 379–393.

——— (2007). "An Empirical Analysis of US and Japanese Health Insurance Using Age–Period–Cohort Decomposition". *Health Economics* 16.5, pp. 475–489.

Gelfand, Alan E. and Sujit K. Sahu (1999). "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models". *Journal of the American Statistical Association* 94.445, pp. 247–253.

Gelman, Andrew et al. (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton: CRC Press.

Gill, Jeff (2008). *Bayesian Methods: A Social and Behavioral Sciences Approach*. 3rd ed. Boca Raton: CRC Press.

Gill, Jeff and Lee D. Walker (2005). "Elicited Priors for Bayesian Model Specifications in Political Science Research". *The Journal of Politics* 67.3, pp. 841–872.

Glenn, Norval D. (1981). "The Utility and Logic of Cohort Analysis". *The Journal of Applied Behavioral Science* 17.2, pp. 247–257.

——— (1989). "A Caution About Mechanical Solutions to the Identification Problem in Cohort Analysis: Comment on Sasaki and Suzuki". *American Journal of Sociology* 95.3, pp. 754–761.

——— (2005). *Cohort Analysis*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Gustafson, Paul (2005). "On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables". *Statistical Science* 20.2, pp. 111–140.

——— (2009). "What Are the Limits of Posterior Distributions Arising From Nonidentified Models, and Why Should We Care?" *Journal of the American Statistical Association* 104.488, pp. 1682–1695.

——— (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*. Boca Raton: CRC Press.

Havulinna, Aki S. (2014). "Bayesian Age–Period–Cohort Models with Versatile Interactions and Long-Term Predictions: Mortality and Population in Finland 1878–2050". *Statistics in Medicine* 33.5, pp. 845–856.

Jackman, Simon (2009). *Bayesian Data Analysis for the Social Sciences*. Chichester: Wiley.

Jones, Geoffrey and Wesley O. Johnson (2014). "Prior Elicitation: Interactive Spreadsheet Graphics With Sliders Can Be Fun, and Informative". *The American Statistician* 68.1, pp. 42–51.

Kadane, Joseph B. and Lara J. Wolfson (1998). "Experiences in Elicitation". *Journal of the Royal Statistical Society. Series D (The Statistician)* 47.1, pp. 3–19.

Knorr-Held, Leonhard and Evi Rainer (2001). "Projections of Lung Cancer Mortality in West Germany: A Case Study in Bayesian Prediction". *Biostatistics* 2.1, pp. 109–129.

Kupper, Lawrence L. et al. (1985). "Statistical Age–Period–Cohort Analysis: A Review and Critique". *Journal of Chronic Diseases* 38.10, pp. 811–830.

Lindley, D. V. (1972). *Bayesian Statistics: A Review*. Philadelphia: Society for Industrial and Applied Mathematics.

Lynch, Scott M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer.

Mason, Karen Oppenheim, William M. Mason, et al. (1973). "Some Methodological Issues in Cohort Analysis of Archival Data". *American Sociological Review* 38.2, p. 242.

Mason, William M. and Stephen E. Fienberg, eds. (1985a). *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York, NY: Springer.

——— (1985b). "Introduction: Beyond the Identification Problem". In Mason, William M. and Stephen E. Fienberg, eds. *Cohort Analysis in Social Research*. New York: Springer, pp. 1–8.

McElreath, Richard (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 1st ed. Boca Raton: Chapman and Hall/CRC.

Meyer, Mary A. and Jane M. Booker (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics and American Statistical Association.

Miller, Alan S. and Takashi Nakamura (1996). "On the Stability of Church Attendance Patterns during a Time of Demographic Change: 1965–1988". *Journal for the Scientific Study of Religion* 35.3, p. 275.

——— (1997). "Trends in American Public Opinion: A Cohort Analysis of Shifting Attitudes from 1972–1990". *Behaviormetrika* 24.2, pp. 179–191.

Morgan, Stephen L. and Christopher Winship (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. New York: Cambridge University Press.

Nakamura, Takashi (1986). "Bayesian Cohort Models for General Cohort Table Analyses". *Annals of the Institute of Statistical Mathematics* 38.1, pp. 353–370.

Neath, Andrew A. and Francisco J. Samaniego (1997). "On the Efficacy of Bayesian Inference for Nonidentifiable Models". *The American Statistician* 51.3, pp. 225–232.

Nielsen, Bent and Jens P. Nielsen (2014). "Identification and Forecasting in Mortality Models". *The Scientific World Journal* 2014, pp. 1–24.

O'Brien, Robert (1989). "Relative Cohort Size and Age-Specific Crime Rates: An Age-Period-Relative-Cohort-Size Model". *Criminology* 27, pp. 57–78.

O'Brien, Robert (2015). *Age–Period–Cohort Models: Approaches and Analyses with Aggregate Data*. Boca Raton: CRC Press.

Poirier, Dale J. (1998). "Revising Beliefs in Nonidentified Models". *Econometric Theory* 14.4, pp. 483–509.

Riebler, Andrea and Leonhard Held (2017). "Projecting the Future Burden of Cancer: Bayesian Age–Period–Cohort Analysis with Integrated Nested Laplace Approximations: Projecting the Future Burden of Cancer". *Biometrical Journal* 59.3, pp. 531–549.

Rue, Håvard and Leonhard Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman & Hall/CRC.

Ryder, Norman B. (1965). "The Cohort as a Concept in the Study of Social Change". *American Sociological Review* 30.6, pp. 843–861.

Sasaki, Masamichi and Tatsuzo Suzuki (1987). "Changes in Religious Commitment in the United States, Holland, and Japan". *American Journal of Sociology* 92.5, pp. 1055– 1076.

——— (1989). "A Caution About the Data to Be Used for Cohort Analysis: Reply to Glenn". *American Journal of Sociology* 95.3, pp. 761–765.

Schmid, Volker and Leonhard Held (2004). "Bayesian Extrapolation of Space-Time Trends in Cancer Registry Data". *Biometrics* 60.4, pp. 1034–1042.

——— (2007). "Bayesian Age–Period–Cohort Modeling and Prediction – BAMP". *Journal of Statistical Software* 21.8, pp. 1–15.

Smith, Theresa R. and Jon Wakefield (2016). "A Review and Comparison of Age–Period–Cohort Models for Cancer Incidence". *Statistical Science* 31.4, pp. 591–610.

Trader, Ramona L. (2014). "Regression, Bayesian". *Wiley StatsRef: Statistics Reference Online*. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat00246.

Wang, Xiaofeng, Yuryan Yue, and Julian J. Faraway (2018). *Bayesian Regression Modeling with Inla*. 1st ed. Boca Raton: Chapman and Hall/CRC.

Winship, Christopher and David J. Harding (2008). "A Mechanism-Based Approach to the Identification of Age–Period–Cohort Models". *Sociological Methods & Research* 36.3, pp. 362–401.

Yang, Yang. (2006). "Bayesian Inference for Hierarchical Age–Period–Cohort Models of Repeated Cross-Section Survey Data". *Sociological Methodology* 36, pp. 39–74.

Yang, Yang and Kenneth C. Land (2013). *Age–Period–Cohort Analysis: New Models, Methods, and Empirical Applications*. Boca Raton: Chapman and Hall/CRC.