5 Priors over Grouped Continuous Variables

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

5.1 Definition and Analysis of Prior Indifference

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

A huge literature in statistics attempts to formalize what information, or lack of it, is represented by a prior, and especially how we can write priors that are minimally informative. Deep philosophical issues arise, primarily around how to represent complete ignorance in the form of a specific probability density—a philosophical stance sometimes known as "logical Bayesianism." Numerous creative ideas have been suggested to try to achieve this goal in some part, such as making the prior invariant to reparameterization (Jeffreys, [1939] 1961), but, as is recognized, the task is ultimately impossible: here, as everywhere, counting on scientific progress about purely philosophical issues would not be prudent.

As Dawid (1983, p. 235) writes, "The formalization of ignorance thus remains the central object of a continuing quest by the knights of the Bayesian round table: inspiring them to imaginative feats of daring, while remaining, perhaps, forever unattainable."

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

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

The main problem we tackle here is that we often know some things and not others about the same set of parameters and wish to write informative priors only for the things we know. For the quantities we do not know, we cannot write a proper prior, and so we use a flat, constant (improper) prior. We see no need to justify the constant prior by appeal to the "principle of insufficient reason" (Laplace, [1820] 1951) or other such concepts (that themselves are based on insufficient reason!). Instead, we view this choice as a simple combination of likelihood and Bayesian inference: when we have information, we use it and the likelihood; when we have no such information, we use only the likelihood.

Our approach has much in common with the spirit of "robust Bayesian analysis" (Berger, 1994; King and Zeng, 2002), although the technology is very different. More relevant to our particular technical approach is the pioneering work of Julian Besag and the literature on spatial smoothing (Besag, 1974, 1975; Besag and Kooperberg, 1995), as well as the work of Speckman and Sun (2001) and the literature on nonparametric regression, in particular the seminal work of Wahba (1978). In the following, we define what we call *prior indifference*, or the indifference of a prior density to a specific set of patterns or values of a set of parameters. We borrow freely from the authors cited here, and others, and combine and extend strands of literature from a diverse set of scholarly fields in order to present a simple but coherent approach. (This chapter requires only some linear algebra and basic mathematical concepts. For readers not familiar with the mathematical concepts we use, such as vector spaces, subspaces, orthogonality, and null spaces, we provide a self-contained review in appendix B, and a glossary of our notation in appendix A.)

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

5.1.1 A Simple Special Case

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

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

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

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

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

5.1.2 General Expressions for Prior Indifference

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

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

Its r-dimensional orthogonal complement—that is, the set of vectors in $\mathbb{R}^d$ that are orthogonal to all the elements of $\mathbb{S}_\circ$ (See appendix B.1.11, page 225)—is then:

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

Clearly any vector $\mu \in \mathbb{R}^d$ can be uniquely decomposed into the sum of two vectors, one in $\mathbb{S}_\circ$, which we denote by $\mu_\circ$, and one in $\mathbb{S}_\perp$, which we denote by $\mu_\perp$. The vectors $\mu_\circ$ and $\mu_\perp$ can be obtained as linear transformations of the vector μ, that is $\mu_\circ = P_\circ\mu$ and $\mu_\perp = P_\perp\mu$, where the matrices $P_\circ$ and $P_\perp$ are called the *projectors* onto $\mathbb{S}_\circ$ and $\mathbb{S}_\perp$, respectively. The projector onto a subspace is uniquely determined by the subspace; that is, for any given subspace, we can easily derive the corresponding projector (as described in appendix B.1.13, page 226). Thus, in the present case, it is easy to see that

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

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

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

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

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

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

5.1.3 Interpretation

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

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

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

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

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

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

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

Thus, the preceding equation is equivalent to

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

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

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

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

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

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

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

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

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

where we have defined the matrix $W \equiv P_\perp B P_\perp$ (remember that $P_\perp$ is symmetric). The only difference between this prior and a regular normal prior is that here, because $P_\perp$ is singular, the matrix W is singular and admits a nontrivial null space $\mathfrak{N}(W)$. Recall that the null space of a matrix W is the set of vectors μ such that $W\mu = 0$ (see appendix B.2.1, page 228, for more detail). In this case the null space $\mathfrak{N}(W)$ coincides with $\mathbb{S}_\circ$, from equation 5.5. Because W is singular, the prior is improper, as expected. The improperness comes only from the existence of the null space, which is the set of vectors "invisible" to W; that is, $W\mu = W(\mu + \mu^*)$ for any $\mu^* \in \mathfrak{N}(W)$.

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

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

Our approach then follows two steps:

1. We use the concept of *the null space of a matrix* to analyze W, the advantage of which is that the tools in linear algebra to characterize and analyze null spaces are well developed.
2. We note that when a pattern in or values of the parameters μ are in the null space of W, then the prior in equation 5.7 has the property of prior indifference given in equation 5.6.

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

We also add slightly novel terminology by referring to the null space $\mathfrak{N}$ of the matrix W in equation 5.7 as *the null space of the prior*. Because the expression $\mu' W \mu$, with W singular and positive semidefinite, defines a seminorm (see appendix B.1.3, page 220), it would be more appropriate, and more in line with some literature, to refer to $\mathfrak{N}$ as "the null space of the seminorm associated with the prior," but this terminology seems unnecessarily complicated, and so we do not adopt it here.

Before proceeding to a full analysis of several classes of priors, we make the key point that partially informative priors can also be used to force the random variables to assume a certain *class* of configurations with high probability, without requiring them to take on any *one* configuration, as would be the case with a proper prior. To see this, consider the prior in equation 5.7, with its null space $\mathfrak{N}(W)$, and let θ assume larger and larger values. This will force the proper part of the prior to become increasingly concentrated around $\mu_\perp = 0$, but still leave $\mu_\circ$ unaffected. Plugging such a prior in the posterior is then equivalent to constraining the solution to the entire null space rather than to a point, as would be the case

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

for a proper prior. Which element of the null space corresponds to the solution will then be determined by the data through the likelihood. If we built the prior in such way that the null space is a set of configurations with "desirable" properties, then we will have found the configuration with these properties that best fit the data. We explore this observation more in detail in the following example.

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

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

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

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

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

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

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

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

5.2 Step 1: A Prior for μ

In this section, we consider the case in which the cross-sectional index is a variable like age, which is intrinsically continuous, although it is discretized in practice. We proceed in two distinct steps, as outlined in section 4.4: in this section, we build a nonparametric (qualitative) prior for the expected value of the dependent variable and then, in section 5.3, use it along with an assumption about the functional form to derive a prior for the regression coefficients $\boldsymbol{\beta}$.

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

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

5.2.1 Measuring Smoothness

Our immediate goal is to find functionals $H_t[\mu]$ defined for any time t that are small when μ is a smooth function of a (remember that a functional is a map from a set of functions to the set of real numbers; see appendix B.1.6, page 223). Functionals of this type are easily constructed using the observation that the oscillating behavior of a function is amplified by the application of any differential operator to it. Therefore, an initial candidate for $H_t[\mu]$ could be

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

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

Example In order to convince ourselves that this functional does what we expect it to do, we compute it for a family of wiggly functions and check that it gets larger if we make the functions more wiggly. Fix $t = 1$ and take the family $\mu_k(a, 1) = \sin(\frac{2\pi k a}{A})$, indexed by the integer k. These are sin waves of frequency proportional to k, so k is a measure of how wiggly these functions are. First, notice that, taking $\mathbb{n}$ even for simplicity, we have

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

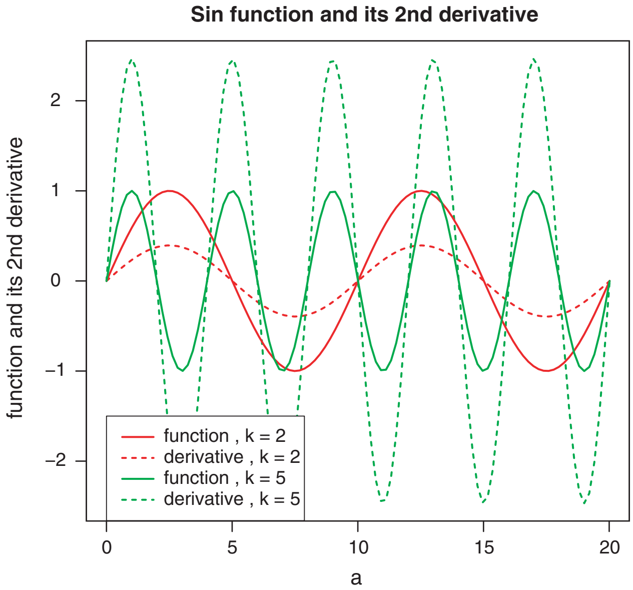

FIGURE 5.1. The sin function and its 2nd derivative, for different frequencies. On the vertical axis we have both the function $\mu_k(a) = \sin(\frac{2\pi k a}{A})$ and its 2nd derivative. Here $A = 20$ and k takes on the values 2 and 5.

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

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

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

For a given k, the value of the functional is increasing with $\mathbb{m}$. Therefore, large values of $\mathbb{m}$ correspond to functionals that are very restrictive, because in these cases even small values of k can lead to a large value for the smoothness functional. This justifies calling $\mathbb{m}$ the degree of smoothness (other justifications for this terminology lie in spline theory and in some other important properties of the preceding smoothness functional but we do not discuss them here). For further information, see Wahba (1990), de Boor (1978), Schumaker (1981), or Eubank (1988).

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

5.2.2 Varying the Degree of Smoothness over Age Groups

Before proceeding, we point out that the smoothness functional in equation 5.8 can be generalized in a way that can be very useful in practice. It is often the case that, while $\mu(a, t)$ is a smooth function of a, it can be smoother in certain regions of its domain than in others. For example, if $\mu(a, t)$ is the expected value of log-mortality from all causes, we know that it will have a fairly sharp minimum at younger ages (and therefore be less smooth), but it will be almost a straight line at older ages (i.e., where it is more smooth). (For example, see figure 2.1, page 22.) Therefore, penalizing the lack of smoothness uniformly across age groups would misrepresent our prior knowledge: younger ages should be penalized relatively less than older ages. This problem is easily fixed by replacing the Lebesgue measure da in the integral in equation 5.8 with a more general measure $dw^{\text{age}}(a)$.[2] For example, we may set $dw^{\text{age}}(a) = a^l da$ for some $l > 0$ in order to penalize older ages more.

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

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

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

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

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

5.2.3 Null Space and Prior Indifference

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

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

density over μ as follows:

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

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

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

These functions have the property that

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

Therefore, the preceding prior has the indifference property:

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

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

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

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

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

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

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

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

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

5.2.4 Nonzero Mean Smoothness Functional

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

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

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

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

5.2.5 Discretizing: From Age to Age Groups

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

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

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

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

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

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

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

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

5.2.6 Interpretation

We now further interpret the smoothness functional in equation 5.12. First, we have seen in section 5.2.3 that the prior associated with the smoothness functional in equation 5.10 is indifferent to polynomials of degree $\mathbb{n} - 1$ in age, with time-dependent coefficients. This important and useful property was derived in the continuous setting, and it also holds in the discretized setting if the derivative operator is discretized properly. In fact, any discretized form of the derivative of order $\mathbb{n}$ should have the property that $\mathrm{D}^{\text{age},\mathbb{n}} \nu = 0$ for any vector ν of the form $\nu_a = a^k$, $k = 0, 1, \ldots, \mathbb{n} - 1$ (and any linear combination of such vectors). This means that the matrix $\mathrm{D}^{\text{age},\mathbb{n}}$ has nullity (see section B.2.1) equal to $\mathbb{n}$ and rank equal to $A - \mathbb{n}$. because the matrix $W^{\text{age},\mathbb{n}}$ is proportional to $(\mathrm{D}^{\text{age},\mathbb{n}})'\mathrm{D}^{\text{age},\mathbb{n}}$, its eigenvalues are simply the squares of the singular values of $\mathrm{D}^{\text{age},\mathbb{n}}$ (see section B.2.4, page 233). As a result, $W^{\text{age},\mathbb{n}}$ has the same rank and nullity as $\mathrm{D}^{\text{age},\mathbb{n}}$:

$$\operatorname{rank}(W^{\text{age},\mathbb{n}}) = A - \mathbb{n}, \quad \operatorname{nullity}(W^{\text{age},\mathbb{n}}) = \mathbb{n}.$$

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

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

Consider the cases $\mathbb{n} = 1$ and $\mathbb{n} = 2$, with $\bar{\mu} = 0$ (zero mean) and $\bar{\mu}$ set to some typical age profile (in this case, the one for all-cause male log-mortality). For figure 5.2, we drew three samples from the proper portion of the prior in equation 5.13. Then we added to each an arbitrary element of the null space. Notice that we say an "arbitrary" and not a "random" element of the null space, because we cannot draw at random from the null space since the density over it is improper. Hence, we selected the particular elements here for visual clarity. In the top left panel, we have set $\mathbb{n} = 1$ and $\mu = 0$: the prior has zero mean and the null space is the space of constant functions. Each of the three random samples from the proper portion of the prior is color-coded (red, green, or blue). We then repeat each of the three samples three times by adding to the sample three arbitrary elements of the null space. Hence, in this graph, we can see three red curves, which differ only by a constant. Our prior is indifferent to the choice among these three red curves, in that they have identical prior probabilities. The same holds for the three green curves and three blue curves in the top left graph of the figure. (The samples originally produced by the algorithm that samples from the proper prior are the ones in the middle, which have zero mean over age groups, but this is a minor technical point about how we happen to choose to draw priors.)

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

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

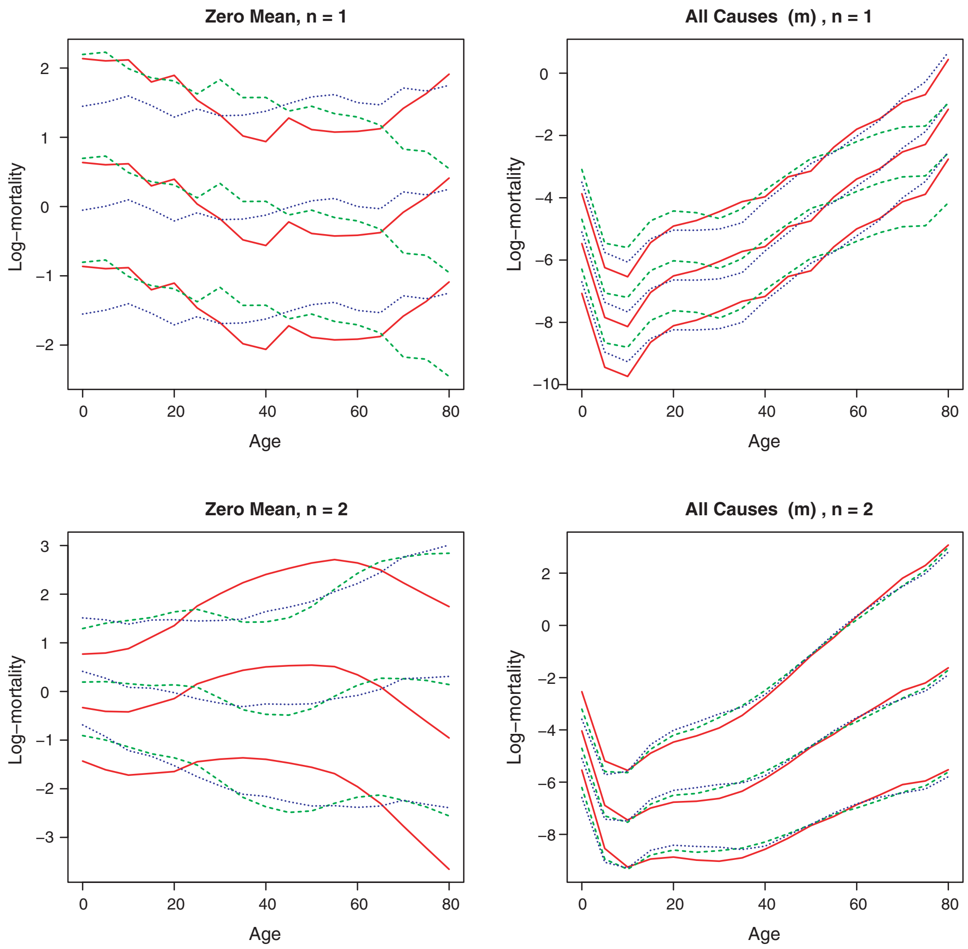

FIGURE 5.2. Age profile samples from smoothness priors with added arbitrary elements of the null space. For each panel, different colors correspond to different samples, while curves of the same color differ by an element of the null space. *Top left*: $\mathbb{n} = 1$ and $\bar{\mu} = 0$; *top right*: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

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

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

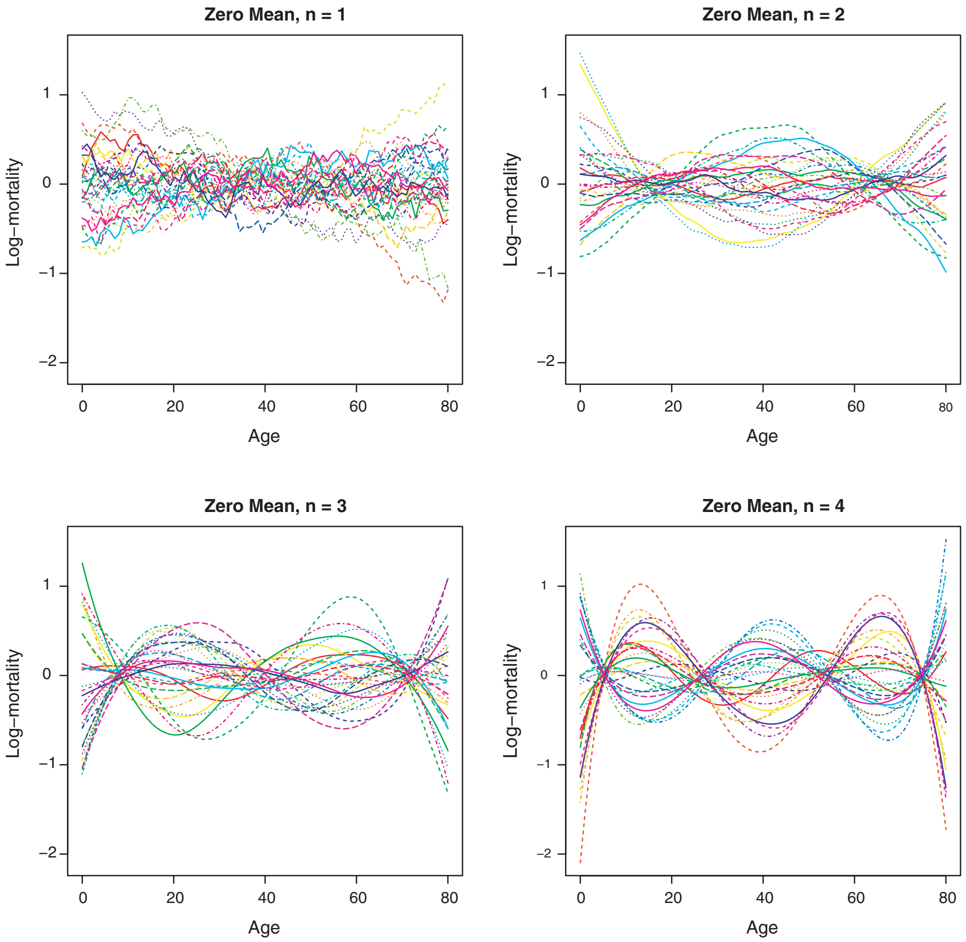

FIGURE 5.3. Age profile samples from the smoothness prior in equation 5.13 for $\mathbb{n} = 1, 2, 3, 4$. Here $A = 80$, and there are 81 age groups, from 0 to 80. The value of θ has been chosen so that the standard deviation of μ_a is 0.3, on average over the age groups, and the scale is the same in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

Another very evident feature of these graphs is that, as $\mathbb{n}$ increases, the samples acquire more and more large "bumps" (or global changes in direction). If we think of the number of bumps as a measure of oscillation then this implies that, as $\mathbb{n}$ increases, the samples oscillate more. This sounds like a contradiction: the point of building smoothness functionals is to penalize functions that oscillate too much, and we have been claiming all along that, as $\mathbb{n}$ increases, the functionals become more restrictive, and therefore their samples should oscillate less. The contradiction is apparent, however, only because we have been mixing two kinds of oscillations: one is *local oscillation*, measured locally by the derivative of order $\mathbb{n}$, and the other is *global oscillation*, measured by the number of bumps, or, better, by the number of zero crossings. The smoothness functional in equation 5.10 is built to penalize the local amount of oscillation, on average, and it does not care about the global shape of a function. In fact, we can dramatically alter the global shape of a function by adding to it a polynomial of degree $\mathbb{n} - 1$ without changing the value of the smoothness functional at all. Take, for example, $\mathbb{n} = 4$, so that the null space is the 4-dimensional space of polynomials of degree 3. Polynomials of degree 3 can have 2 "bumps," but they are the smoothest possible curve according to this smoothness functional. Therefore, it should not be surprising that samples from the prior often have one more bump, as is the case for most of the samples in the bottom right panel of figure 5.3.[3]

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

Finally, we compare the smoothness functionals in equation 5.12 derived in this section with the "bare bones" smoothness functional in equation 4.15 (page 70). Because $\mathbb{n} \geq 1$, the constant vector $\nu = (1, 1, \ldots, 1)$ is always in the null space of $W^{\text{age},\mathbb{n}}$, implying that the rows and columns of $W^{\text{age},\mathbb{n}}$ always sum to 0. In turn, this implies (via the result in appendix B.2.6, page 237) that it is always possible to find a matrix $s^{\text{age},\mathbb{n}}$ such that we can write the smoothness functional in equation 5.12 (page 86) in the same form as equation 4.15 (page 70):

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

Because the derivative is a local operator, the matrix $W^{\text{age},\mathbb{n}}$ will usually have a "band" structure, such that $W^{\text{age},\mathbb{n}}_{aa'}$ is different from 0 only if a and a' are "close" to each other (although not necessarily first neighbors). This structure is reflected in the matrix $s^{\text{age},\mathbb{n}}$, which makes clear that the smoothness functional in equation 5.12 is a sum of "local" contributions, obtained by comparing the value of μ in a certain age group with the values in nearby age groups. In this respect, the smoothness functional in equation 5.12 is like the one in the previous chapter, which resulted from pairwise comparisons between neighboring age groups. An important difference between equations 5.14 and 4.15, however, is that in equation 4.15 the "weights" $s_{aa'}$ were chosen to be positive. This allows us to interpret the smoothness functional as a way to penalize configurations in which similar age groups do not have similar values of μ. As soon as $\mathbb{n}$ becomes larger than 1, however, many of the elements of $s^{\text{age},\mathbb{n}}$ become negative, which may appear counterintuitive. Yet this must be the case because, if the elements of $s^{\text{age},\mathbb{n}}$ were all positive, then the null space of the functional could only be the set of constants, independent of the values of $s^{\text{age},\mathbb{n}}$, while we know that the size of the null space increases with $\mathbb{n}$. In other words, the elements of $s^{\text{age},\mathbb{n}}$ become negative in order for some cancellations to occur, cancellations necessary to ensure that the null space has the correct structure.

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

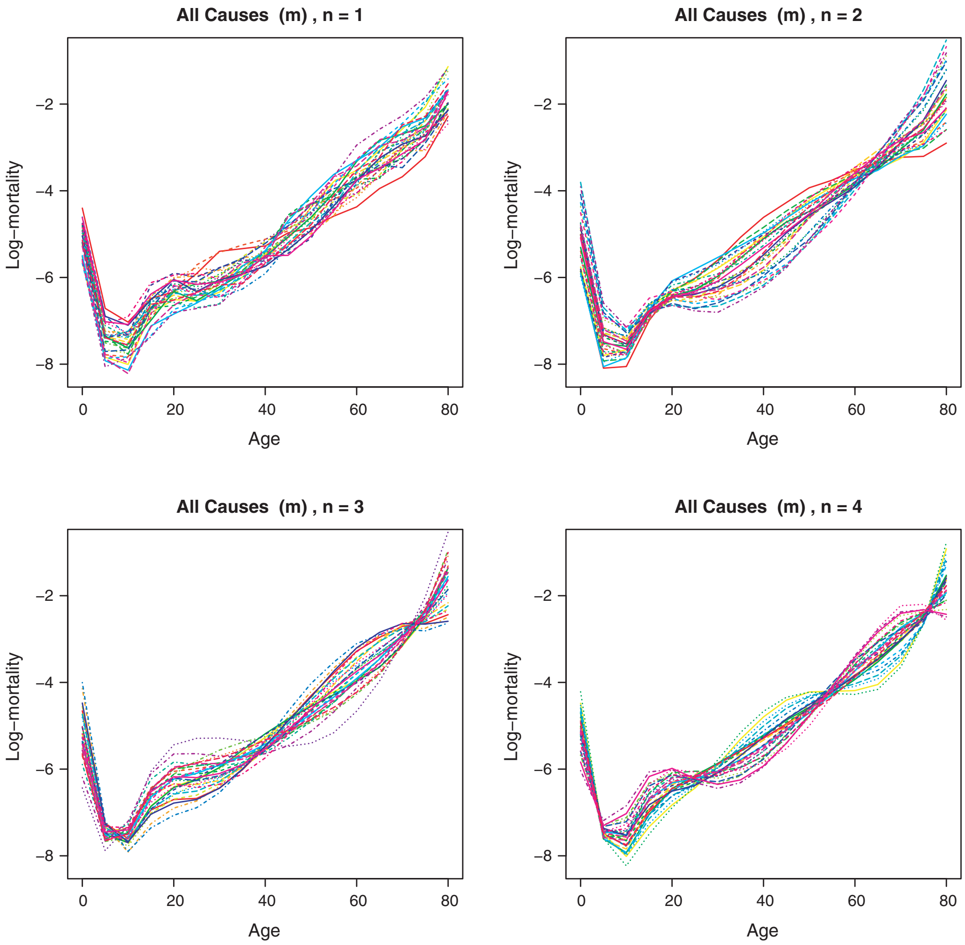

FIGURE 5.4. Age profile samples from the smoothness prior in equation 5.13 for $\mathbb{n} = 1, 2, 3, 4$ and a typical age Profile for all-causes log-mortality in males. There are 17 age groups ($A = 17$), at 5-years intervals, labeled $0, 5, \ldots, 80$. The value of θ has been chosen so that the standard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

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

5.3 Step 2: From the Prior on μ to the Prior on β

5.3.1 Analysis

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

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

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

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

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

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

5.3.2 Interpretation

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

Second, the weights $W^{\text{age},\mathbb{n}}_{aa'}$ in equation 5.16 are fully specified once we choose, from prior information, the order $\mathbb{n}$ of the smoothness functional in equation 5.10 (see section 6.1). That is, all A^2 elements of this matrix—all elements of which, under previous

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

approaches, would need to be specified by hand—are uniquely determined by the single scalar $\mathbb{n}$.

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