7 Adding Priors over Time and Space

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

Mathematically, this chapter extends the model for cross sections labeled by indices that vary over sets that are continuous in nature but discretized (like a set of age groups or income brackets), to point continuous without discretizing (like time or distance from the population center), to variables composed of discrete sets with no metric structure (such as a list of countries, diseases, or ethnic groups). The mathematical form for all these priors turns out to be the same as those considered thus far.

7.1 Smoothing over Time

Another form of prior knowledge we are likely to have, and indeed do have in our running example, is that the expected value of the dependent variable μ_{it} varies smoothly over time.[1] Because time is a continuous variable, we can use the same reasoning we developed in section 5.2 for smoothing over age. Hence, denoting by i as a generic cross-sectional

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

index, we use an analogous smoothness functional to smooth over time:

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

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

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

where we have defined the matrix:

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

While equation 7.2 is mathematically similar to equations 4.18 (page 71) and 7.6, it differs in a substantively important way. One way to see this is that equation 7.2 contains no interactions among any cross sections, so that, for example, a random permutation of the cross-sectional index will leave this expression unchanged. From a probabilistic point of view, this means that the coefficients $\boldsymbol{\beta}_i$ are *independent* (not identically distributed) random variables, while the whole point of smoothing over age groups in equation 4.18 (page 71) and countries in equation 7.6 is precisely that the $\boldsymbol{\beta}_i$ are *dependent* in specific interesting ways.

The observation is that equations 4.18 (page 71) and 7.6 are insensitive to any temporal behavior of the covariates, because time enters into those equations only through the product $\mathbf{Z}_i' \mathbf{Z}_j$: a random permutation of the time index will leave this quantity unchanged. In contrast, the whole point of equation 7.2 is to take into account the temporal behavior of the covariates, because it explicitly incorporates the time derivatives of the covariates.

7.1.1 Prior Indifference and the Null Space

The smoothness functional in equation 7.1 is a standard smoothness functional of the type discussed in chapter 5. Therefore, in terms of μ the null space contains profiles of log-mortality, which, in each cross section, evolve over time as polynomials in t of degree $\text{ⓝ} - 1$. What happens when we project on the subspace spanned by the covariates? Because the smoothness functional in equation 7.1 simply sums over the cross sections, the null space of the prior can be studied independently for each cross section, and so for simplicity in the following we assume that there is only one cross section, which we denote with the index i.

Restricted to the subspace defined by our covariates and linear functional form, the null space is simply the null space of the matrix $\mathbf{C}_{ii}^{\text{time},n}$. Because for any matrix V we know that V and $V'V$ share the same null space, the null space of $\mathbf{C}_{ii}^{\text{time},n}$ in equation 7.3

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

1. Suppose we have only one age group and one country but multiple time periods. If $\mathbb{n} = 1$, then only constant levels are in the null space on the scale of μ. If only a constant term is included in the covariate matrix $\mathbf{Z}$, then after the prior is restricted to the subspace defined by the covariates and our functional form $\mathbb{S}_{\mathbf{Z}}$, all patterns are in the null space. That is, because the prior can affect only the constant term, and the constant term can have no effect on the smoothness of μ_t over time, the prior has no parameters to adjust to achieve smoothness and will do nothing. Thus, in this situation, the prior will have no effect on the empirical results, which is equivalent to a likelihood-only analysis, or a Bayesian analysis with an improper uniform prior.
2. If we continue with the first example, but with the change $\mathbb{n} = 2$, then the null space for μ includes constant shifts as well as changes in the slope of μ_t over time. However, because the covariates still include only the constant term, the prior will have no effect on the constant term or the empirical estimates. So nothing changes from the first example.
3. If $\mathbb{n} = 1$, and $\mathbf{Z}$ includes a constant term and GDP, then the null space for μ, and after restriction to the subspace $\mathbb{S}_{\mathbf{Z}}$, includes only constant shifts. This means that the prior will have no effect on the constant term in the regression. The prior smoothes expected log-mortality in this example by requiring the squared first derivative with respect to time to be small. However, the only way the prior can have an effect such as this is by affecting the coefficient on GDP. If GDP varies a lot over time, then this prior can impose smoothness only by reducing the size of its coefficient.
4. If we continue with the previous example but change the degree of smoothness to $\mathbb{n} = 2$, the null space in μ becomes larger: the prior would now be indifferent to changes in both levels and slopes of μ_t over time. However, the null space restricted to $\mathbb{S}_{\mathbf{Z}}$ is the same as the previous example because patterns linear in time are not in the span of the covariates (unless GDP happened to be exactly linear in time). The non-null space has changed from the previous example because the prior now penalizes the second derivative of GDP. In other words, the prior is now sensitive to, and tries to smooth, μ only as affected by the nonlinear portions of GDP. It will do this by reducing the size of the coefficient on GDP. (The fact that GDP may be nearly linear is immaterial, because any nonlinearities are enough to let the prior use its coefficient to achieve the desired degree of smoothness.)
5. If we continue with the previous example, suppose we add a time trend to the constant and GDP in $\mathbf{Z}$. Because $\mathbb{n} = 2$, the null space on the scale of μ includes shifts in both the level and slope, as before. Because the covariates are

sufficiently rich to represent these patterns, the null space restricted to $\mathbb{S}_{\mathbf{Z}}$ also includes level and slope shifts. In this example, the prior would then have an effect only on the coefficient of GDP. This coefficient is adjusted by the prior to keep μ smooth, and so it would be reduced if the second derivatives of this variable were large. The constant and slope on the linear trend are unaffected by the prior.

7.2 Smoothing over Countries

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

In this section, we consider the case where the cross-sectional index i is a label that does not come with a continuous structure naturally associated with it. In this case an appropriate mathematical framework to describe smoothness is graph theory. To keep things simple, we proceed here intuitively, leaving the formal connection to graph theory and our precise definitions to appendix E.

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

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

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

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

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

Now that the prior is in a form similar to equation 5.12 (page 86), we repeat the steps of section 5.2 to derive the prior for the coefficients $\boldsymbol{\beta}$. Plugging the specification $\mu_{ct} = \mathbf{Z}_{ct}\boldsymbol{\beta}_c$ into equation 7.5 we obtain, predictably:

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

where we have defined the matrix:

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

and $\mathbf{Z}_c$ is the usual data matrix for country c, that is, the matrix whose rows are the vectors $\mathbf{Z}_{ct}$. The key here is the perfect correspondence between equation 7.6 and equation 5.16 (page 92). There is only one difference: for the prior over age groups the matrix $W^{\text{age},\mathbb{n}}$ was determined easily by the choice of the scalar $\mathbb{n}$, whereas for the matrix W^{cntry} we have to do more work and build the adjacency matrix s^{cntry} by hand, using experts' opinions to figure out which countries should be considered neighbors. Mathematically, then the two forms are the same. The only difference is due to the substantive differences between the two problems.

7.2.1 Null Space and Prior Indifference

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

$$\mathcal{P}(\mu|\theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{t} \mu'_t W^{\text{cntry}} \mu_t\right). \tag{7.7}$$

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

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

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

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

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

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

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

Although using islands to add flexibility to the prior and to expand the null space in this way can be very useful, in practice such data sets can be analyzed separately for the group of countries on each island. As such, we analyze only the case with no islands (i.e., one world island) in the rest of this section. Obviously, the same result applies separately and independently within each island.

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

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

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

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

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

the prior does not change. Therefore, when it comes to these coefficients, the prior carries information only about their relative levels.

7.2.2 Interpretation

In chapter 5, we considered the case of smoothness functionals for functions of variables that are continuous in principle, although discrete in practice. In this case, the key ingredient was the possibility of using the derivative of order $\mathbb{n}$ as a measure of local variation. We also saw that priors written in terms of derivatives, as in equation 5.10, could be written, once discretized, as in equation 5.14 (page 90), a form that we used as a starting point for the prior in equation 7.4. We noticed that when the derivative in equation 5.10 is of order 1, the weights $s_{aa'}^{\text{age},\mathbb{n}}$ that connect one age group to another should be positive. Because the weights in equation 7.4 $s_{cc'}^{\text{cntry}}$ are, by construction, positive, it is natural to ask whether the expression 7.4 can be related to some notion of first derivative with respect to the country label. We now show that this is indeed the case, by giving an intuitive description and leaving the details to appendix E.

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

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

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

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

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

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

This derivation of the smoothness functional does not add anything from a technical point of view. However, it allows us to write a smoothness functional for a discrete variable

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

7.3 Smoothing Simultaneously over Age, Country, and Time

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

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

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

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

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

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

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

$$W_{ij} \equiv W_{ca,c'a'} \equiv \frac{\theta^{\text{age}}}{C} W_{aa'}^{\text{age},\text{ⓝ}} \delta_{cc'} + \frac{\theta^{\text{ctr}}}{A} W_{cc'}^{\text{cntry}} \delta_{aa'}$$

$$\mathbf{C}_{ij} \equiv \frac{1}{T} \mathbf{Z}_i' \mathbf{Z}_j + \frac{\theta^{\text{time}} \delta_{ij}}{CATW_{ii}} \left(\frac{d^{\text{ⓝ}} \mathbf{Z}_i}{dt^{\text{ⓝ}}} \right)' \left(\frac{d^{\text{ⓝ}} \mathbf{Z}_i}{dt^{\text{ⓝ}}} \right).$$

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

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

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

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

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

The remarkable feature of this expression is its simplicity: it is a normal prior, with the same mathematical form as the prior implied by the original prior on coefficients in equation 4.5, and yet it embeds information about smoothness over ages, countries, and time and all on the scale of the dependent variable. In addition, while the weight matrix W^{cntry} has to be constructed by hand, the matrix $W^{\text{age},\text{ⓝ}}$ is automatically determined once the integer ⓝ has been chosen. It follows from the construction in this section that the general structure of the prior in equation 7.12 does not depend on the particular choice that we have made in equation 7.10: we could easily add terms with mixed derivatives with respect to ages and times, or ages and countries, or triple-term interactions that combine ages, countries, and time, and still the final result would be of the type of equation 7.12. We turn to this topic in the next section.

7.4 Smoothing Time Trend Interactions

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

7.4.1 Smoothing Trends over Age Groups

In addition, or as an alternative, to the smoothness functional for age in equation 5.8 (page 81), we now show how to allow the time *trend* of the expected value of the dependent variable to vary smoothly across age groups. For example, we often expect log-mortality to decrease at similar rates for all age groups (except possibly in infants). For example, most demographers would normally be highly skeptical of mortality forecasts for a country, sex, and cause, that trended upward for 25-year-olds but downward for 30-year-olds. In this case, an appropriate smoothness functional can be obtained by replacing the level $\mu(a,t)$ in equation 5.8 (page 81) with the time derivative $\partial\mu(a,t)/\partial t$, and averaging over time as well:

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

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

With a smoothness functional like that in equation 7.13, the prior for the coefficients $\boldsymbol{\beta}$ has exactly the same form as the one in equation 4.19 (page 71), the only difference being that the covariates should be replaced by their time derivatives, and therefore the matrices $\mathbf{C}_{aa'}$ should be replaced by

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

Obviously time derivatives of order $\eth_t > 1$ could be considered too, if desired, by simply replacing the matrix $\mathbf{C}_{aa'}^{\text{time},1}$ with a similarly defined matrix $\mathbf{C}_{aa'}^{\text{time},\eth_t}$, in which the first derivative is replaced with the derivative of order $\eth_t$.

7.4.2 Smoothing Trends over Countries

Just as we sometimes may want to smooth the trend of the expected value of the dependent variable across *age groups*, we may also want to do the same across *countries*. This is often a less restrictive but useful form of prior knowledge, which avoids our having to make statements about the levels of the expected value of the dependent variable. This is especially useful in situations where two countries, with different base levels of the dependent variable, pursue similar policies over time, or benefit from the same relevant technological advances.

By simply repeating the argument of section 7.4.1, we can see that a prior corresponding to this kind of knowledge has the same form as the one implied by equation 7.3, in

which the covariates have been replaced by their time derivative, and therefore the matrices $\mathbf{C}_{cc'}$ have been replaced by

$$\mathbf{C}_{cc'}^{\text{time},1} \equiv \frac{1}{T}\left(\frac{d\mathbf{Z}_c}{dt}\right)'\left(\frac{d\mathbf{Z}_{c'}}{dt}\right). \tag{7.15}$$

7.5 Smoothing with General Interactions

In this final section where we build priors, we give detailed calculations for a generic term involving a triple interaction of age, country, and time. By setting appropriate matrices equal to the identity, researchers will be then able to derive formulas for all the other pairwise interactions.

We begin with a smoothness functional of the form:

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\mu,\theta] \equiv \frac{\theta}{C}\sum_c \int_0^T dw^{\text{time}}(t)\int_0^A dw^{\text{age}}(a)\left\|\nabla^c \frac{\partial^{\mathbb{m}_a+\mathbb{m}_t}\mu(c,a,t)}{\partial a^{\mathbb{m}_a}\partial t^{\mathbb{m}_t}}\right\|^2, \tag{7.16}$$

where $dw^{\text{time}}(t)$ and $dw^{\text{age}}(a)$ are probability measures allowing one to impose different degrees of smoothness in different parts of the integration domain, and $\mathbb{m}_a$ and $\mathbb{m}_t$ are integers denoting the order of derivatives with respect to age and time, respectively.

We first discretize the derivatives with respect to age and time:

$$\frac{\partial^{\mathbb{m}_a+\mathbb{m}_t}\mu(c,a,t)}{\partial a^{\mathbb{m}_a}\partial t^{\mathbb{m}_t}} \Rightarrow \mu'_{cat} \equiv \sum_{a't'} \mathrm{D}_{aa'}^{\mathbb{m}_a}\mathrm{D}_{tt'}^{\mathbb{m}_t}\mu_{ca't'},$$

where we use the notation μ' to remind us that this quantity is a derivative. Then we compute the gradient with respect to c of μ'_{cat} and square it:

$$\left\|\nabla^c \mu'_{cat}\right\|^2 = \sum_{c'} s_{cc'}^{\text{cntry}}(\mu'_{cat}-\mu'_{c'at})^2.$$

Now we sum this expression over c, a, and t, weighting the sums over a and t with the weights w_a^{age} and w_t^{time}, which are the discrete versions of the probability measures $dw^{\text{age}}(a)$ and $dw^{\text{time}}(t)$:

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\mu,\theta] = \frac{\theta}{C}\sum_{cat} w_a^{\text{age}} w_t^{\text{time}} \sum_{c'} s_{cc'}^{\text{cntry}}(\mu'_{cat}-\mu'_{c'at})^2 = \theta \sum_{cc'at} W_{cc'}^{\text{cntry}} w_a^{\text{age}} w_t^{\text{time}} \mu'_{cat}\mu'_{c'at},$$

where we define $W^{\text{cntry}} = C^{-1}[(s^{\text{cntry}})^+ - s^{\text{cntry}}]$ as in section 7.2. Now we substitute the expression for μ'_{cat} and obtain

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\mu,\theta] = \theta \sum_{cc'at} W_{cc'}^{\text{cntry}} w_a^{\text{age}} w_t^{\text{time}} \sum_{a't'} \mathrm{D}_{aa'}^{\mathbb{m}_a}\mathrm{D}_{tt'}^{\mathbb{m}_t}\mu_{ca't'} \sum_{a''t''} \mathrm{D}_{aa''}^{\mathbb{m}_a}\mathrm{D}_{tt''}^{\mathbb{m}_t}\mu_{c'a''t''}.$$

Reshuffling the order of the sums, we obtain

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\mu,\theta] = \theta \sum_{cc'aa'tt'} W^{\text{cntry}}_{cc'} \left(\sum_a \mathrm{D}^{\mathbb{m}_a}_{aa'} w^{\text{age}}_a \mathrm{D}^{\mathbb{m}_a}_{aa'} \right) \left(\sum_t \mathrm{D}^{\mathbb{m}_t}_{tt'} w_t \mathrm{D}^{\mathbb{m}_t}_{tt'} \right) \mu_{cat} \mu_{c'a't'}.$$

Defining the following matrices:

$$W^{\text{age},\mathbb{m}_a} \equiv (\mathrm{D}^{\mathbb{m}_a})' \text{diag}[w^{\text{age}}_a] \mathrm{D}^{\mathbb{m}_a} \qquad W^{\text{time},\mathbb{m}_t} \equiv (\mathrm{D}^{\mathbb{m}_t})' \text{diag}[w^{\text{time}}_a] \mathrm{D}^{\mathbb{m}_t}, \tag{7.17}$$

we obtain

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\mu,\theta] = \theta \sum_{cc'aa'tt'} W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{m}_a}_{aa'} W^{\text{time},\mathbb{m}_t}_{tt'} \mu_{cat} \mu_{c'a't'}.$$

Now the prior for $\boldsymbol{\beta}$ can be obtained by simply substituting the specification $\mu_{cat} = \mathbf{Z}_{cat}\boldsymbol{\beta}_{ca}$ in the preceding expression, obtaining

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{cc'aa'tt'} W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{m}_a}_{aa'} W^{\text{time},\mathbb{m}_t}_{tt'} \mathbf{Z}_{cat} \boldsymbol{\beta}_{ca} \mathbf{Z}_{c'a't'} \boldsymbol{\beta}_{c'a'}.$$

By rewriting $\mathbf{Z}_{cat}\boldsymbol{\beta}_{ca}$ as $\boldsymbol{\beta}'_{ca}\mathbf{Z}_{cat}$ and changing the order of the sums, we write

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{cc'aa'} W^{\text{cntry}}_{c'c''} W^{\text{age},\mathbb{m}_a}_{aa'} \boldsymbol{\beta}'_{ca} \left(\sum_{tt'} \mathbf{Z}'_{cat} W^{\text{time},\mathbb{m}_t}_{tt'} \mathbf{Z}_{c'a't'} \right) \boldsymbol{\beta}_{c'a'}.$$

Now we define the matrix:

$$\mathbf{C}^{\mathbb{m}_t}_{ca,c'a'} \equiv \mathbf{Z}'_{ca} W^{\text{time},\mathbb{m}_t} \mathbf{Z}_{c'a'}, \tag{7.18}$$

where $\mathbf{Z}_{ca}$ is the usual data matrix for cross section ca, which has for each row vector $\mathbf{Z}_{cat}$. If we use the $\mathbf{C}$ matrices defined previously, the smoothness functional for $\boldsymbol{\beta}$ simplifies to

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{cc'aa'} W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{m}_a}_{aa'} \boldsymbol{\beta}'_{ca} \mathbf{C}^{\mathbb{m}_t}_{ca,c'a'} \boldsymbol{\beta}_{c'a'}. \tag{7.19}$$

Defining the multi-indices $i = ca$ and $j = c'a'$, and letting $W^{\mathbb{m}_a,\text{cntry}}_{ij} \equiv W^{\text{cntry}}_{cc'} W^{\text{age},\mathbb{m}_a}_{aa'}$, we simplify the preceding expression further and write it in the usual form:

$$H^{\mathbb{m}_a,\mathbb{m}_t}[\boldsymbol{\beta},\theta] = \theta \sum_{ij} W^{\mathbb{m}_a,\text{cntry}}_{ij} \boldsymbol{\beta}'_i \mathbf{C}^{\mathbb{m}_t}_{ij} \boldsymbol{\beta}_j. \tag{7.20}$$

Expression 7.19, however, is probably the most useful form when deriving new smoothness functionals, because it allows researchers to plug into it the desired values and derive, as special cases, all the priors discussed in this book. Equation 7.20 will often be the form most useful for estimation.

Example Consider the problem of smoothing the time trend over age groups, as described in section 7.4.1. The corresponding smoothness functional is a particular special case of

equation 7.16 in which, instead of the gradient with respect to the country variable, we have the derivative of order 0, in which $\mathbb{n}_t = 1$ and $\mathbb{n}_a$ is arbitrary. As pointed out in section 7.2.2, the derivative of order 0 with respect to countries corresponds to the choice $W^{\text{cntry}} = I$. Plugging these choices in equation 7.19, we obtain the smoothness functional:

$$H^{\mathbb{n}_a,\mathbb{n}_t}[\boldsymbol{\beta}, \theta] = \theta \sum_{caa'} W^{\text{age},\mathbb{n}_a}_{aa'} \boldsymbol{\beta}'_{ca} \mathbf{C}^1_{ca,ca'} \boldsymbol{\beta}_{ca'},$$

where, from equation 7.18, we have defined

$$\mathbf{C}^1_{ca,ca'} = \mathbf{Z}'_{ca} W^{\text{time},1} \mathbf{Z}_{ca'}.$$

In order to compare with equation 7.14, we need to consider the special case, considered in that section, of a uniform measure over time: $dw^{\text{time}}(t) = T^{-1}dt$. Substituting this choice for $dw^{\text{time}}(t)$ in equation 7.17, we obtain

$$W^{\text{time},1} = \frac{1}{T}(\mathrm{D}^1)'\mathrm{D}^1$$

Substituting this expression in the preceding definition of $\mathbf{C}^1_{ca,ca'}$, we obtain, as expected, the same expression of equation 7.14:

$$\mathbf{C}^1_{ca,ca'} = \frac{1}{T}\mathbf{Z}'_{ca}(\mathrm{D}^1)'\mathrm{D}^1\mathbf{Z}_{ca'} = \frac{1}{T}(\mathrm{D}^1\mathbf{Z}_{ca})'(\mathrm{D}^1\mathbf{Z}_{ca'}) = \frac{1}{T}\left(\frac{d\mathbf{Z}_{ca}}{dt}\right)'\left(\frac{d\mathbf{Z}_{ca'}}{dt}\right).$$

⊠

7.6 Choosing a Prior for Multiple Smoothing Parameters

The smoothing parameter θ determines how much weight to put on the prior as compared to the data in the estimation and thus how smooth the forecasts will be. In chapter 6, we showed that, when only one prior is being used, the only information needed to set θ is the average standard deviation of the prior. We also showed in section 6.2.2 that setting the average standard deviation of the prior simultaneously set all the properties of the samples from the prior. When more than one prior is used, as should be the case in many applications, the reasoning of chapter 6 still applies, although the implementation is more involved. We describe these procedures here and then a way we have devised to automate some of them in section 7.6.2.

Suppose we are using K priors, each with a smoothness parameter θ_k ($k = 1, \ldots, K$). If we used only the k-th prior, then we could use the result in chapter 6 that θ_k is uniquely determined by the average standard deviation of the prior, which we denote σ_k (see equation 6.14). Because the parameter σ_k is interpretable and uniquely determines θ_k, we use σ_k to parametrize our single priors. This implies that the expected value of any summary measure $F(\mu)$ is a function of the K parameters σ_k. Therefore, in order to estimate

what the parameters σ_k should be, all we have to do is to find K summary measures F_k $(k = 1, \ldots K)$, for which we have information about in terms of their expected values, and which we denote by $\bar{F}_k$. Then the values of the smoothness parameters are determined by solving the following system of equations:

$$\mathrm{E}_{\perp}[F_k(\mu)|\sigma_1, \ldots, \sigma_K] = \bar{F}_k, \quad k = 1, \ldots, K. \tag{7.21}$$

With more than one prior, it is not possible to solve these equations analytically even for summary measures that are quadratic in μ, and so numerical procedures must be employed. In addition, if more than K summary measures are available, it may be advisable to use all of them. The system of equations (7.21) then becomes overdetermined, and an approximate solution is required, but the advantage is that one gains a better insight into the properties of the prior.

We have found in practice that the following summary measures are well suited for our application:

$$\begin{aligned}
\mathrm{SD}(\mu) &\equiv \frac{1}{AT}\sum_{a=1}^{A}\sum_{t=1}^{T}(\mu_{at} - \bar{\mu}_a)^2 \\
F_{\text{age}}(\mu) &\equiv \frac{1}{AT}\sum_{t=1}^{T}\sum_{a=2}^{A}|\mu_{at} - \mu_{a-1,t}| \\
F_{\text{time}}(\mu) &\equiv \frac{1}{AT}\sum_{t=2}^{T}\sum_{a=1}^{A}|\mu_{at} - \mu_{a,t-1}| \\
F_{\text{age/time}}(\mu) &\equiv \frac{1}{AT}\sum_{t=2}^{T}\sum_{a=2}^{A}|(\mu_{at} - \mu_{a,t-1}) - (\mu_{a-1,t} - \mu_{a-1,t-1})|
\end{aligned} \tag{7.22}$$

The summary measure SD is the average standard deviation of the prior, which measures how much samples from the prior vary around the average age profile $\bar{\mu}$. Summary measure F_{age} measures how much log-mortality changes going from one age group to the next, and F_{time} summarizes the changes in log-mortality from one time period to the next. Finally, $F_{\text{age/time}}$ is a measure of how much the time trend changes from one age group to the next. These quantities are easily interpretable, and with some clarification about what they mean, we find that demographers and other experts often have a reasonable estimate of their expected values. If such expert knowledge is not available, one can still get an idea of the expected values of these quantities using a procedure that has the flavor of empirical Bayes, and which we describe in the following section.

An important point is that for certain choices of $\bar{F}_i$, equations 7.21 may have no solution. But instead of this posing a methodological problem, it indicates that the prior is unable to produce samples with the desired characteristics or, in other words, that some of the expert's (or analyst's) choices were logically inconsistent. Learning about such logical inconsistencies can be helpful to an expert in making these choices. Furthermore, it is easy to imagine how this can happen. Suppose, for example, that one prefers a prior with a tiny overall standard deviation, but also one that allows wide variations in log-mortality from one year to the next, or from one age group to the next. These choices are clearly

not compatible, and no set of prior parameters will produce such a result. Problems of this type are more likely to arise when relatively few covariates are being used, because samples from the prior are more constrained in those cases. With more covariates, the prior has more coefficients to adjust to produce patterns consistent with a wider range of patterns.

For these reasons, we recommend, rather than trying to solve equations 7.21 numerically, that analysts study the behavior of the expected values of the summary measures as a function of the parameters σ_k, and in particular the range of values that they can assume. This can be done, for example, using multiple scatterplots, as we illustrate here. The following list outlines our recommended strategy for choosing appropriate values of σ_k:

1. Define at least N_s ($N_s \geq K$) summary measures for which the expected values are approximately known.
2. Assign a reasonably wide range of variation to each σ_k (e.g., 0.01 to 2) and use it to define a grid in the space of the σ_k (e.g., each interval [0.01, 2] could be divided in five subintervals).
3. For each combination of σs corresponding to a point in the grid, which we label γ, draw a large number of samples from the prior. Use these samples to compute numerically the implied expected value of the summary measures, which we denote by $\tilde{F}_i$. Store the results in a table whose rows have the structure:

$$(\sigma_1^\gamma, \ldots, \sigma_K^\gamma, \ \tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma).$$

4. Produce all pairs of scatterplots of the summary measures as a function of the σ's and of each other. Qualitatively assess where the target values $\bar{F}_i$ of the summary measures fall in the scatterplots.
5. Define a distance $D(\cdot\,;\,\cdot)$ in the space of the N_s summary measures and use it to find the combination of prior parameters $\hat{\sigma}_k$ that produces empirical values of the summary measures that are closest to the target. Formally, define

$$\hat{\gamma} \equiv \arg\min_{\gamma} D(\tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma; \bar{F}_1, \ldots, \bar{F}_{N_s})$$

and then set $\hat{\sigma}_k = \sigma_k^{\hat{\gamma}}$.

This procedure depends on the choice of the distance measure $D(\cdot;\,\cdot)$. The usual Euclidean distance in general will not work well because the target values $\bar{F}_1, \ldots, \bar{F}_{N_s}$ may have quite different scales. Therefore, we suggest to use the Euclidean distance on the relative errors:

$$D(\tilde{F}_1^\gamma, \ldots, \tilde{F}_{N_s}^\gamma; \bar{F}_1, \ldots, \bar{F}_{N_s}) \equiv \sum_i \left(\frac{\tilde{F}_i^\gamma - \bar{F}_i}{\bar{F}_i} \right)^2 .$$

A detailed example of the implementation of this procedure is described in section 11.2. In order to fix ideas, here we simply offer a preview of what the previously mentioned scatterplots may look like and what kind of information they convey.

7.6.1 Example

Here we consider the special case of "deterministic forecasts," that is, where the covariates are known to have a well-defined analytical form as a function of time alone. This is useful when realistic covariates are missing, but a linear time trend is known not to be sufficient and a nonlinear trend may be needed. We consider a specification of the form:

$$\mu_{at} = \beta_a^{(0)} + \beta_a^{(1)} t + \beta_a^{(2)} \log(t - \alpha),$$

where α is a given number,[3] and assume that we are interested in forecasting mortality by lung cancer in males. The details of this example are provided in section 11.2. For this choice, the target values of the summary measures 7.22 have been derived with the procedure described in the next section, and are as follows:

$$\bar{\text{SD}} \approx 0.3, \quad \bar{F}_{\text{age}} \approx 0.53, \quad \bar{F}_{\text{time}} \approx 0.033, \quad \bar{F}_{\text{age/time}} \approx 0.007.$$

We use a smoothness functional consisting of three terms:

$$\begin{aligned} H[\mu, \theta_{\text{age}}, \theta_{\text{time}}, \theta_{\text{age/time}}] \equiv & \frac{\theta_{\text{age}}}{AT} \int_0^T dt \int_0^A da \left(\frac{d^2}{da^2} (\mu(a,t) - \bar{\mu}(a)) \right)^2 \\ & + \frac{\theta_{\text{time}}}{AT} \int_0^T dt \int_0^A da \left(\frac{d^2}{dt^2} \mu(a,t) \right)^2 \\ & + \frac{\theta_{\text{age/time}}}{TA} \int_0^T dt \int_0^A da \left(\frac{\partial^3 \mu(a,t)}{\partial a \partial t^2} \right)^2, \end{aligned}$$

and therefore need to estimate three smoothness parameters, θ_{age}, θ_{time}, and $\theta_{\text{age/time}}$. As explained at the beginning of this section, it is convenient to reparametrize the smoothness functional using the standard deviations of the prior, σ_{age}, σ_{time}, and $\sigma_{\text{age/time}}$. We remind the reader that σ_{age} is simply the average standard deviation of the prior over age groups, if it were used in isolation, and it is linked to θ_{age} by equation 6.14 (page 102), which we rewrite here (see equation 6.7, page 101, and equation 6.9, page 101, for the definition of other quantities in this formula) as

$$\theta_{\text{age}} = \frac{\text{Tr}\left(\mathbf{Z} D_{\text{age}}^{+} \mathbf{Z}' \right)}{AT \sigma_{\text{age}}^2}.$$

The same formula applies, with the obvious modifications, for θ_{time} and $\theta_{\text{age/time}}$.

Once the smoothness functional and the specification have been chosen, the prior is defined, and all we have to do is to draw samples from it and compute empirically the expected value of the summary measures for many different values of the prior parameters σ_{age}, σ_{time}, and $\sigma_{\text{age/time}}$. Scatterplots of the prior parameters against the empirical values of the summary measures are shown in figure 7.1.

An important message emerging from the analysis of the scatterplots is that the target value of the summary measure $\bar{\text{SD}} = 0.3$ is not compatible with the value of the summary measure $\bar{F}_{\text{time}} = 0.033$. This is easily seen in the scatterplot of SD against F_{time}

[3] In this example, we set $\alpha = 1876$. A justification for this choice is given in section 11.2.

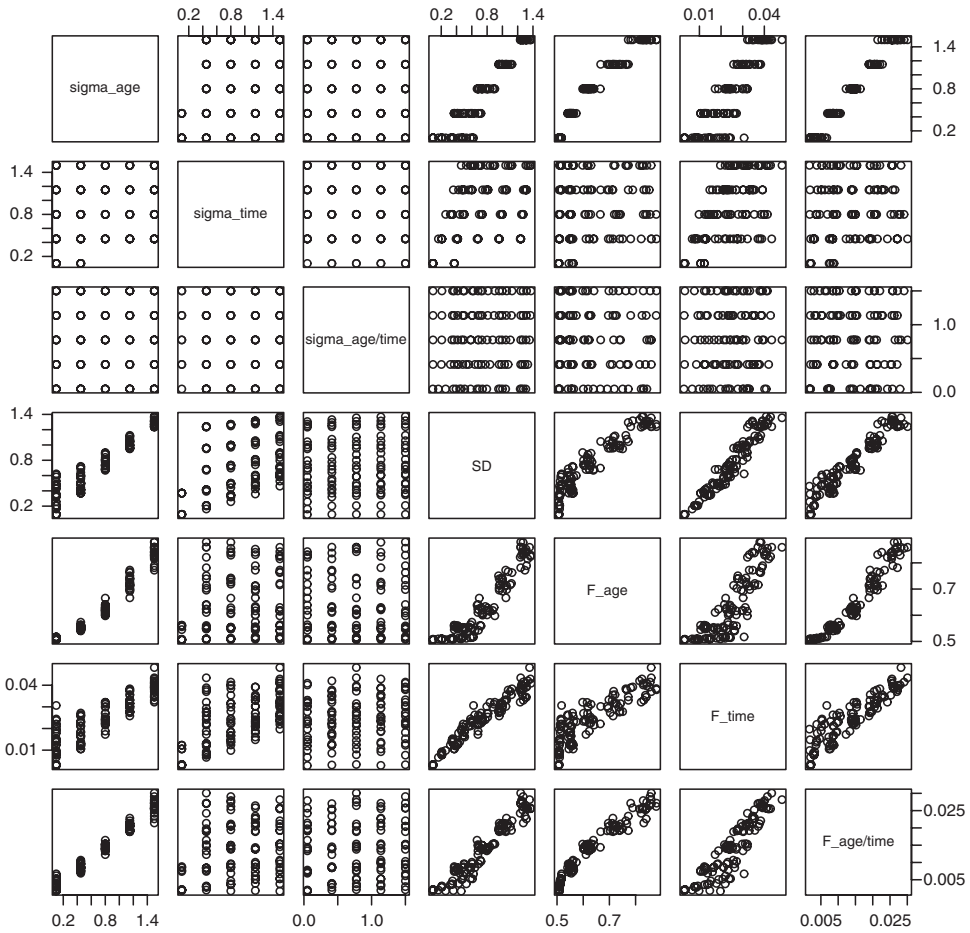

FIGURE 7.1. Scatterplots of summary measures by prior parameters. The plot shows the relationship between the prior parameters σ_{age}, σ_{time}, and $\sigma_{\text{age/time}}$ and summary measures SD, F_{age}, F_{time}, and $F_{\text{age/time}}$.

(see the fourth row and sixth column in figure 7.1). Fixing the value of SD around 0.3, we see that only very small values of F_{time} can be realized (around 0.010). Therefore, if we want a prior with a summary measure F_{time} closer to its target value of 0.033, we will need to settle for a higher value of the average standard deviation SD. The reason underlying this behavior of the prior is that we have very few covariates, each of which is a fixed function of time. Because the prior can operate only by influencing the coefficients on these variables, there just is not much room to maneuver. (The same issue would occur if we had a prior that suggested that log-mortality move according to a quadratic but only a linear term for time was included among the covariates; no amount of adjusting the coefficients would produce the desired effect.) Therefore, samples from the prior are very constrained to begin with, and we will need to settle for an approximate solution of equation 7.21, which can be found using the procedure described above in this section. Alternatively, we could drop the average standard deviation SD from the list of summary measures, realizing that it is not compatible with the other measures. In tables 7.1 and 7.2, we display the combination of parameters σ_{age}, σ_{time}, and $\sigma_{\text{age/time}}$ that lead to the empirical values of the summary

TABLE 7.1.
Summary Measures and Parameter Values: Combinations of Different Values of the Parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ Together with the Corresponding Value of the Summary Measures SD, F_{age}, F_{time} and $F_{\text{age/time}}$.

σ_{age}	σ_{time}	$\sigma_{\text{age/time}}$	SD	F_{age}	F_{time}	$F_{\text{age/time}}$
0.10	0.80	1.50	0.35	0.51	0.015	0.0047
0.45	0.10	1.50	0.37	0.56	0.012	0.0085
0.45	0.45	0.05	0.39	0.55	0.012	0.0074
0.45	0.10	0.41	0.37	0.56	0.012	0.0088
0.45	0.10	0.78	0.37	0.56	0.011	0.0074
0.45	0.10	1.14	0.37	0.55	0.011	0.0077
0.45	0.45	0.78	0.41	0.55	0.013	0.0077
0.10	1.15	1.50	0.49	0.52	0.022	0.0061
0.45	0.45	1.14	0.41	0.55	0.013	0.0083
0.45	0.45	1.50	0.42	0.55	0.013	0.0084
0.45	0.45	0.41	0.41	0.56	0.013	0.0085
0.45	0.80	0.05	0.44	0.55	0.014	0.0069
0.10	1.15	1.14	0.48	0.52	0.021	0.0052
0.10	1.15	0.78	0.46	0.52	0.019	0.0051
0.10	0.80	0.78	0.34	0.51	0.014	0.0038
0.45	0.80	0.41	0.48	0.55	0.017	0.0085
0.10	1.15	0.41	0.43	0.51	0.018	0.0039
0.10	0.80	1.14	0.35	0.51	0.014	0.0037
0.45	0.80	0.78	0.49	0.55	0.017	0.0087
0.10	0.80	0.41	0.32	0.51	0.012	0.0036
0.45	1.15	0.05	0.51	0.54	0.016	0.0073
0.45	0.80	1.50	0.50	0.55	0.018	0.0094
0.45	0.80	1.14	0.50	0.55	0.019	0.0096
0.45	1.15	0.41	0.56	0.56	0.021	0.0087
0.10	1.50	0.41	0.53	0.51	0.022	0.0038

Note: The rows are sorted according to their distance to the target values for the summary measures. We use only four target values, one for each summary measure.

measures closest to the target values. The tables show the 25 best values, starting from the top, sorted according to their distance to the target. The difference between the tables is that in table 7.1 we include SD in the list of target summary measures, whereas in table 7.2 we do not. Notice that in the top row of table 7.1, the summary measures F_{time} and $F_{\text{age/time}}$ are quite far from the target values: this is the price to pay if we want to match the summary measure SD. The top row of table 7.2, instead, matches well the targets for the summary measures F_{age}, F_{time}, and $F_{\text{age/time}}$, but it matches poorly the target for SD, because SD was not included in the list of targets.

7.6.2 Estimating the Expected Value of the Summary Measures

In order to use the procedure just outlined, we need to begin with some substantively reasonable ranges for the expected values of the summary measures. While it would be possible to elicit some of these measures from subject matter experts, here we pretend that expert opinion is unavailable and get our estimates using a procedure that has the flavor of empirical Bayesian analysis. In practice, we recommend that this procedure be used in

TABLE 7.2.
Summary Measures and Parameter Values: Combinations of Different Values of the Parameters σ_{age}, σ_{time} and $\sigma_{\text{age/time}}$ Together with the Corresponding Value of the Summary Measures SD, F_{age}, F_{time} and $F_{\text{age/time}}$.

σ_{age}	σ_{time}	$\sigma_{\text{age/time}}$	SD	F_{age}	F_{time}	$F_{\text{age/time}}$
0.10	1.50	1.50	0.62	0.52	0.031	0.0067
0.45	1.50	0.41	0.64	0.54	0.026	0.0085
0.45	1.50	0.05	0.58	0.56	0.023	0.0078
0.10	1.15	1.50	0.49	0.52	0.022	0.0061
0.10	1.50	0.78	0.58	0.51	0.024	0.0053
0.45	1.15	0.78	0.59	0.55	0.021	0.0086
0.10	1.50	1.14	0.61	0.52	0.024	0.0053
0.45	1.15	0.41	0.56	0.56	0.021	0.0087
0.45	1.50	0.78	0.69	0.54	0.026	0.0096
0.45	1.15	1.50	0.61	0.55	0.023	0.0096
0.10	1.15	1.14	0.48	0.52	0.021	0.0052
0.45	1.15	0.05	0.51	0.54	0.016	0.0073
0.45	1.50	1.50	0.72	0.56	0.027	0.0103
0.45	0.80	0.41	0.48	0.55	0.017	0.0085
0.45	0.80	0.78	0.49	0.55	0.017	0.0087
0.45	1.15	1.14	0.60	0.57	0.023	0.0101
0.10	1.15	0.78	0.46	0.52	0.019	0.0051
0.45	0.80	1.50	0.50	0.55	0.018	0.0094
0.45	0.80	1.14	0.50	0.55	0.019	0.0096
0.45	0.80	0.05	0.44	0.55	0.014	0.0069
0.45	1.50	1.14	0.71	0.56	0.026	0.0107
0.45	0.45	0.78	0.41	0.55	0.013	0.0077
0.45	0.45	1.14	0.41	0.55	0.013	0.0083
0.45	0.45	1.50	0.42	0.55	0.013	0.0084
0.45	0.45	0.41	0.41	0.56	0.013	0.0085

Note: The rows are sorted according to their distance to the target values for the summary measures. We use only three target values, that is $\bar{F}_{\text{age}}$, $\bar{F}_{\text{time}}$ and $F_{\text{age/time}}$.

conjunction with expert involvement, perhaps as a starting point to orient the experts. If used with related or earlier data sets, this procedure can be thought of as formalization and measurement of the *source* of expert knowledge.

Instead of looking at the data to determine the parameters of the prior, we look at smoothed versions of the data, obtained by making forecasts of the time series and then selecting the in-sample portion of the predictions. But to examine a set of forecasts, we need to start with some baseline model, although the particular model we choose should not matter much. Our recommendation is to use a simple version of our model, although one could also use least squares when it happens to produce reasonable in-sample fits in a particular application.

Once a smooth and realistic version of the data is available, we can then use these time series to compute an estimate of the expected value of the summary measures in equation 7.22. For example, denoting by $\hat{\mu}_{cat}$ the in-sample prediction of the model, the expected value of the summary measure F_{age} can be estimated as

$$\bar{F}_{\text{age}} \approx \frac{1}{CAT} \sum_{c=1}^{C} \sum_{t=1}^{T} \sum_{a=2}^{A} |\hat{\mu}_{cat} - \hat{\mu}_{c,a-1,t}|.$$

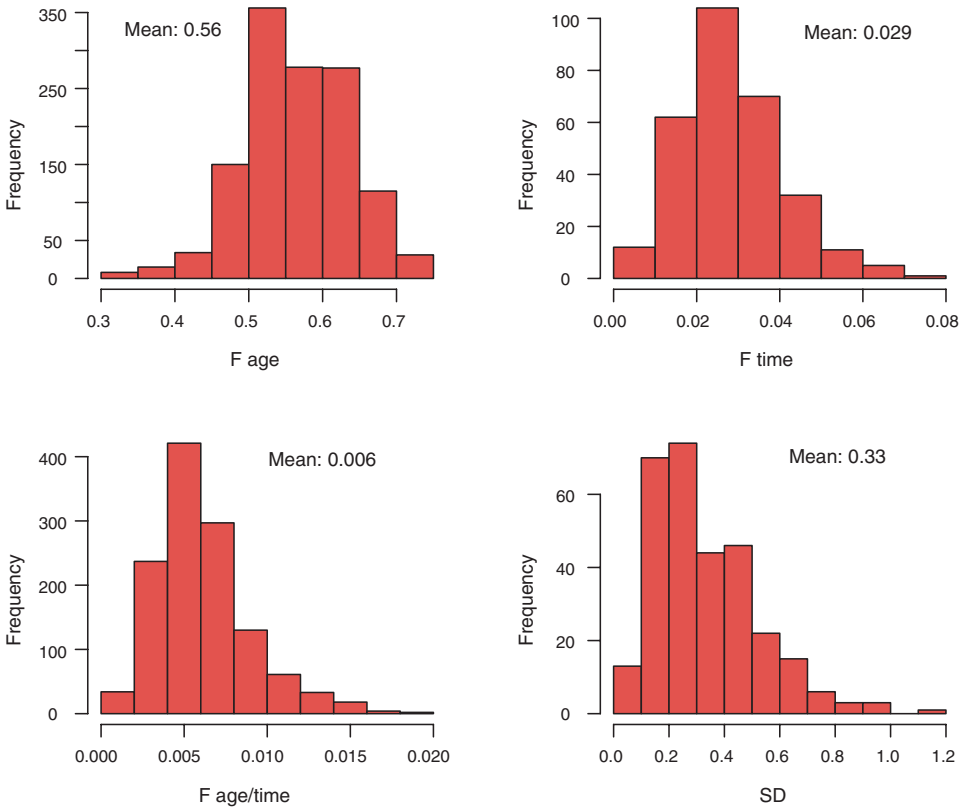

FIGURE 7.2. Result of the empirical Bayes-like procedure for setting summary measure target values: Empirical distribution of the quantities involved in the computation of the summary measures in equation 7.22. The dependent variable is log-mortality for lung cancer in males, for 25 countries.

In addition to the mean, it may be useful to look at the entire distribution of the terms in the preceding sum, in order to get an idea of its spread and possible skewness.

In order to provide an example of such distributions, we consider the case of death by lung cancer in males. We use a simple specification with a linear trend and a logarithmic trend in order to get a basic set of forecasts. The basic forecasts are initially obtained using our Bayesian method with a prior that smoothes over age groups only (using a second derivative and constant weights) for all the countries with more than 15 observations, using a standard deviation of the prior equal to 0.3. Because not all the forecasts look reasonable, we eliminate those which do not, and rerun our method, trying a few alternative values for the standard deviation of the prior. The whole point of this procedure is to create a fairly large number of smoothed versions of the in-sample data that look realistic. In these data, we find that a value of the standard deviation of the prior equal to 0.2 produces a reasonable in-sample prediction, which we use as baseline starting point.

The distributions of the quantities involved in the computation of the summary measures 7.22 are shown in figure 7.2, together with their mean values.

While the procedure outlined here is not rigorous, it is an example of the kind of qualitative analysis one can perform to set reasonable starting values for the standard

deviation of the prior to help orient experts. The main point here is that it is possible to link the standard deviation of the prior to other quantities that are, at least in principle, observable, and about which we might have real prior knowledge. In our example this link is provided by figure 7.1.

7.7 Summary

This chapter offers a rich set of priors for analyzing mortality rates. But it also offers a set of tools researchers can use to adapt new priors in new substantive problems. The key features of our approach in this chapter involve the application of the two-step method (introduced in chapter 4) for specifying priors on the expected value of the dependent variable, rather than on the coefficients directly; new methods for the analysis of prior indifference via null spaces; and ways we developed to set priors with genuine prior knowledge.