8 Comparisons and Extensions

In this chapter, we provide some general procedures for understanding the priors built in chapters 5–7. We begin in section 8.1 with a systematic comparison of the priors-on-coefficients approach with our priors on the expected value of the dependent variable. We then prove, in section 8.2, that priors specified in the large literature on hierarchical Bayesian models that feature exchangeable clusters of cross sections are special cases of our models. The results in this section demonstrate that our results about the inappropriateness of putting priors on coefficients in spatial models apply to the hierarchical literature as well. It also demonstrates how our approach has the same attractive features of empirical Bayes but without having to leave the standard Bayesian approach to inference. Section 8.3 shows how our models can also be used for smoothing noisy data to reveal the underlying patterns, even if forecasting is not of interest. Section 8.4 then shows how to modify our methods when the dependent variable changes meaning, such as when the international classification of disease changes definitions.

8.1 Priors on Coefficients versus Dependent Variables

In this section, we compare the prior on coefficients from section 4.2 with that on the expected value of the dependent variable, in section 4.4 and chapters 5 and 7. We provide intuition by comparing the notion of distance in section 8.1.1 and the relevant conditional densities in section 8.1.2. Section 8.1.3 describes connections between the results in our first two sections with theoretical results from the pattern recognition literature.

8.1.1 Defining Distances

In order to facilitate comparison, we write each prior in two equivalent forms with the aid of the quadratic form identity (appendix B.2.6, page 237). Then, we put the prior on $\boldsymbol{\beta}$ in equation 4.5 (page 59) side by side with the prior on μ from equation 7.11 (page 132) as

follows:

$$H^{\mu}[\boldsymbol{\beta}, \theta] = \theta \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j \qquad \Leftrightarrow \quad H^{\beta}[\boldsymbol{\beta}, \Phi] = \sum_{ij} W_{ij} \boldsymbol{\beta}_i' \Phi \boldsymbol{\beta}_j \tag{8.1}$$

$$= \frac{1}{2}\theta \sum_{ij} s_{ij} \|\mu_i - \mu_j\|^2 \quad \Leftrightarrow \quad = \frac{1}{2} \sum_{ij} s_{ij} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}^2, \tag{8.2}$$

where $\mu_i = \mathbf{Z}_i \boldsymbol{\beta}_i$ is a $T \times 1$ vector, $\|\cdot\|$ is the Euclidean norm, and $\|\mathbf{b}\|_{\Phi}^2 = \mathbf{b}'\Phi\mathbf{b}$ is the Mahalanobis norm of the vector $\mathbf{b}$ (the left-hand side of equation 8.2 can be proved to be equal to the left-hand side of equation 8.1 by direct substitution of μ_i and μ_j into the expression). Notice that the matrix s does not have to be the same in the two priors, but because it has similar meaning and its explicit form is irrelevant here, we just take it to be the same to ease notation.

When imposing smoothness on $\boldsymbol{\beta}$, researchers use s_{ij} as a distance in the space of cross-sectional units, but, for fixed i and j, no natural definition of distance between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ exists. Therefore, the usual procedure is to parametrize the distance between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ as the Mahalanobis distance $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}$. This approach obviously cannot be used when $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ have different dimensions, or correspond to different covariates, because the set of coefficients would have no obvious metric structure and so would not be comparable.

In contrast, in our case, rather than comparing the coefficients $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, we compare their predicted patterns for the expected value of the dependent variable, μ_i and μ_j, taking advantage of the fact that μ_i and μ_j are interpretable and there exists a natural distance between them, no matter what covariates are included. This distance is Euclidean (rather than Mahalanobis; see appendix B.1.3, page 220), and so the normalization matrix Φ is not required. In other words, we project $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ into *the same higher-dimensional metric space* through the covariate matrices $\mathbf{Z}_i$ and $\mathbf{Z}_j$ and then compare them. The covariates play here the role of "translators," allowing one to compare vectors of disparate quantities. This they do through the matrices $\mathbf{C}_{ij}$, which allow us to project a vector of "type i" onto a vector of "type j."

This result can be seen more clearly in equation 8.1 where the prior on coefficients contains a sum of scalar products $\boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_j$, which does not have meaning unless $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are of the same type. However, on the right-hand side of equation 8.1, we see that following our approach the scalar products $\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j$ are well defined: vector $\boldsymbol{\beta}_j$ of type j is converted to a vector of type i by the matrix $\mathbf{C}_{ij}$, and then the usual Euclidean scalar product is computed (because $\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j = \boldsymbol{\beta}_j'\mathbf{C}_{ji}'\boldsymbol{\beta}_i$, we can also say that vector $\boldsymbol{\beta}_i$ of type i is converted to a vector of type j by the matrix $\mathbf{C}_{ji}'$). The matrices $\mathbf{C}_{ij}$, despite their simplicity, allow us to impose a metric structure on a set that does not have any existing structure. While the set of coefficients $\boldsymbol{\beta}$ does not even have the structure of a vector space (see section B.1.1, page 218), because the sum of $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ is not defined, a notion of distance is defined between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, when translated to the scale of the expected value, by the expression:

$$d^2(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j | \mathbf{Z}_i, \mathbf{Z}_j) \equiv \|\mathbf{Z}_i\boldsymbol{\beta}_i - \mathbf{Z}_j\boldsymbol{\beta}_j\|^2 = \boldsymbol{\beta}_i'\mathbf{C}_{ii}\boldsymbol{\beta}_i + \boldsymbol{\beta}_j'\mathbf{C}_{jj}\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_i'\mathbf{C}_{ij}\boldsymbol{\beta}_j. \tag{8.3}$$

This expression should be compared with the Mahalanobis distance, which, written in terms of scalar products, is as follows:

$$\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}^2 = \boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_i + \boldsymbol{\beta}_j'\Phi\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_j.$$

This comparison reinforces the point that the expression $\boldsymbol{\beta}_i' \mathbf{C}_{ij} \boldsymbol{\beta}_j$ is the "natural" scalar product between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, and, indeed, it has all the properties of a scalar product (except, of course, the fact that a true scalar product is always defined between elements of the same set; see appendix B.1.2, page 219). Similarly, the distance in equation 8.3 satisfies all the axioms of a distance or, to be precise, of a semidistance, because $d^2(\boldsymbol{\beta}_i, \boldsymbol{\beta}_j | \mathbf{Z}_i, \mathbf{Z}_j) = 0$ does *not* imply that $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ (for a formal definition see appendix B.1.2, page 219).

One way to summarize this discussion is to say that the intuition of putting a prior on coefficients is reasonable, except that the notion of similarity should be defined using prior knowledge about the expected value of the dependent variable.

8.1.2 Conditional Densities

Another useful way to compare the two approaches is to examine the implied conditional priors. Thus, the prior density on the $\boldsymbol{\beta}$'s implies the following conditional prior distribution of the coefficient $\boldsymbol{\beta}_i$, given the values of all the other coefficients $\boldsymbol{\beta}_{-i}$:

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}_{-i}, \Phi \sim \mathcal{N}\left(\sum_j \frac{s_{ij}}{s_i^+} \boldsymbol{\beta}_j \ , \ \frac{\Phi^{-1}}{s_i^+} \right). \tag{8.4}$$

The preceding expression confirms the intuition that *smoothing is achieved by letting $\boldsymbol{\beta}_i$ be a weighted average of the regression coefficients of the neighboring cross sections*, and obviously loses any meaning when the $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are not comparable. Performing a similar calculation for our prior in equation 7.12 (page 132), we obtain the following conditional prior:

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}_{-i}, \theta \sim \mathcal{N}\left(\sum_j \frac{s_{ij}}{s_i^+} \mathbf{C}_{ii}^{-1} \mathbf{C}_{ij} \boldsymbol{\beta}_j, \frac{1}{\theta s_i^+} \mathbf{C}_{ii}^{-1} \right). \tag{8.5}$$

The key to this expression is the presence of two sets of matrices, with different roles: in the conditional mean, the matrix $\mathbf{C}_{ij}$ converts vectors of "type j" into vectors of "type i," but also produces a vector with different measurement units, and so the matrix $\mathbf{C}_{ii}^{-1}$ converts this result to the correct measurement units, to ensure that the coefficients $\boldsymbol{\beta}$ have measurement units that are the inverse of the measurement units of the covariates.

The presence of $\mathbf{C}_{ii}^{-1}$ in the conditional variance ensures that the Bayes estimator 4.3 (page 58) based on the prior in equation 7.12 (page 132) produces forecasts that are invariant for scaling of the covariates *in each cross-sectional unit*. In other words, we can decide to use pounds instead of dollars in some cross-sectional units and still obtain the same forecast (and obviously a different set of appropriately scaled coefficients). If we used the prior on coefficients in equation 4.4 (page 59), not only would we have to make sure that the covariates in the different cross sections are measured in the same units, but, if we changed units, we would also have to change the scale of the covariance parameter Φ.

8.1.3 Connections to "Virtual Examples" in Pattern Recognition

Expression 8.5 has an interesting interpretation in terms of what in the pattern recognition literature are called "virtual examples." The connection to virtual examples is useful here

to clarify the meaning of the prior and, in chapter 10, as a starting point for a fast estimation procedure that does not require Markov Chain Monte Carlo algorithms.

In order to simplify the exposition, we do not smooth over time, so that $\mathbf{C}_{ij} = \frac{1}{T}\mathbf{Z}_i'\mathbf{Z}_j$, and let us interpret equation 8.5 as saying that, conditional on the values of all the other coefficients $\boldsymbol{\beta}_{-i}$, we expect $\boldsymbol{\beta}_i$ to be in a neighborhood of the conditional mean, a fact that we write informally as

$$\boldsymbol{\beta}_i \approx \sum_j \frac{s_{ij}}{s_i^+}(\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\mathbf{Z}_j\boldsymbol{\beta}_j.$$

Then, noting that the quantities $\mathbf{Z}_j\boldsymbol{\beta}_j = \mu_j$ are the predicted values for the dependent variable in cross section j, we write

$$\boldsymbol{\beta}_i \approx (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i' \sum_j \frac{s_{ij}}{s_i^+}\mu_j.$$

The sum in the preceding expression is simply the average of the predicted values for the dependent variable in the cross sections that are neighbors of cross section i (excluding i itself because $s_{ii} = 0$), and we call this quantity $\bar{\mu}_i$, rewriting

$$\boldsymbol{\beta}_i \approx (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\bar{\mu}_i. \tag{8.6}$$

This expression is a standard least-squares estimator and has a simple interpretation. Given the values of all the other coefficients $\boldsymbol{\beta}_{-i}$, we could get an a priori likely estimate of $\boldsymbol{\beta}_i$ in two steps:

1. Obtain an estimate for the dependent variable in cross section i by averaging the predicted values of the cross sections that are neighbors of i (the vector $\bar{\mu}_i$).
2. Then, to obtain the coefficients in cross section i, run a least-squares regression of $\bar{\mu}_i$ on $\mathbf{Z}_i$.

Because the vector $\bar{\mu}_i$ is not a vector of observed values, or "examples" but rather is inferred using prior knowledge, we say that it is a vector of "virtual examples," and in this sense we could say that *the role of the prior knowledge we have on the problem is to create suitable sets of virtual examples*. For more discussion of the connection between prior information and virtual examples, see Abu-Mostafa (1992), Bishop (1995), and Niyogi, Girosi, and Poggio (1998).

8.2 Extensions to Hierarchical Models and Empirical Bayes

In this section, we demonstrate two fundamental results. First, our methods incorporate the key attractive features seen in empirical Bayes approaches, but without having to resort to the sometimes problematic empirical Bayes theory of inference. In empirical Bayes, hyperparameters from the last level of a hierarchical model are estimated rather than chosen a priori. Although this procedure might seem better because it brings the data to bear on

the problem of making difficult choices about obscure hyperparameters, many scholars question the inferential validity of this approach: it uses the data twice and inferences must be corrected in various ad hoc ways to avoid underestimating the width of confidence intervals. Despite the inferential problems, however, this procedure is still frequently used, one important reason for which is because using the data in this way turns out to be equivalent to making the prior indifferent to certain chosen parameters (Carlin and Louis, 2000). For example, with empirical Bayes it is possible to achieve shrinkage among a set of parameters without having to specify the mean of the parameters. Of course, our formal approach to prior indifference accomplishes exactly the same task, but entirely within the standard Bayesian framework. We demonstrate this equivalence here.

Second, we show here that Bayesian hierarchical models, with clusters of exchangeable units, are a special case of the Bayesian spatial models we are analyzing in this book. As such, our results about the inappropriateness of putting priors directly on coefficients in spatial models (see section 4.3) also extends to the Bayesian hierarchical modeling literature. Taken together, it would seem that the many Bayesian models that use covariates are using prior densities that inappropriately reflect their prior knowledge. All our techniques for putting priors on the expected value of the dependent variable and developing priors indifferent to chosen features of the parameters apply to hierarchical models as well.

8.2.1 The Advantages of Empirical Bayes without Empirical Bayes

We begin by considering a hierarchical linear model, with N cross sections and N vectors of coefficients $\boldsymbol{\beta}_i$. A common assumption is the following "shrinkage" prior:

$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\gamma,\ \tau^2\right).$$

The "direct" effect of this prior is to shrink the coefficients $\boldsymbol{\beta}_i$ toward the same mean γ. The "indirect" effect is that the coefficients are shrunk toward each other. It is often the case that the indirect effect is more desirable than the direct one: one can be confident that the coefficients $\boldsymbol{\beta}_i$ should be similar to each other without necessarily knowing what value they should assume. In other words, the researcher may be agnostic (indifferent) about the absolute level of the coefficients but may be knowledgeable about their relative size. Let us apply the idea of using subspaces to represent indifference. It is sufficient to work with one-dimensional coefficients, so we assume $\boldsymbol{\beta}_i, \gamma \in \mathbb{R}$ in the following. Taking $\tau = 1$ for simplicity, the preceding prior can be rewritten as

$$\mathcal{P}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\beta}_i - \gamma)^2\right). \tag{8.7}$$

Defining the $N \times 1$ vectors $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)$ and $\boldsymbol{\gamma} \equiv (\gamma, \ldots \gamma)$, we rewrite the preceding expression in vector form:

$$\mathcal{P}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|^2\right). \tag{8.8}$$

While this prior is defined over $\mathbb{R}^N$, there is a whole subspace of $\mathbb{R}^N$ we are indifferent to: this is the set $V \subset \mathbb{R}^N \equiv \{x \mid x = (k, \ldots, k),\ \forall k \in \mathbb{R}\}$, which coincides with the diagonal of the positive orthant[1] in $\mathbb{R}^N$. In other words, we are indifferent between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i + k$, for any $k \in \mathbb{R}$.

How do we modify this prior so that it expresses the indifference we seek? Denoting by $P_\perp$ the projector onto $V_\perp$, the orthogonal complement of V, the prior 8.8 can be made indifferent to V by simply projecting its argument onto $V_\perp$. Therefore, we define a new prior:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|P_\perp(\boldsymbol{\beta} - \boldsymbol{\gamma})\|^2\right).$$

Because $\boldsymbol{\gamma} \in V$ by construction, then $P_\perp \boldsymbol{\gamma} = 0$ and the preceding prior becomes simply:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|P_\perp \boldsymbol{\beta}\|^2\right). \tag{8.9}$$

This expression makes clear that the only part of $\boldsymbol{\beta}$ we have prior knowledge about is $P_\perp \boldsymbol{\beta}$, that can be interpreted as the portion of $\boldsymbol{\beta}$ that contains only "relative" information. Let us find an explicit expression for $P_\perp$. By the properties of projection operators in appendix B.1.13 (page 226), we have $P_\perp = I - P_\circ$, where $P_\circ$ is the projector onto V, which we now recognize as the null space of the prior. $P_\circ$ is easily built in terms of an orthonormal basis for the subspace V, which is given by the constant row vector $\mathbf{v} = \frac{1}{\sqrt{N}}(1, \ldots, 1)$, where the factor $\sqrt{N}$ ensures normalization. Then the projector $P_\circ$ is given by $P_\circ \equiv \mathbf{v}'\mathbf{v}$ (see page 226). The form of both $P_\perp$ and $P_\circ$ is given as

$$P_\circ \equiv \frac{1}{N}\begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix} \qquad P_\perp \equiv \begin{pmatrix} 1-\frac{1}{N} & -\frac{1}{N} & \ldots & -\frac{1}{N} \\ -\frac{1}{N} & 1-\frac{1}{N} & \ldots & -\frac{1}{N} \\ \vdots & \vdots & \ldots & \vdots \\ -\frac{1}{N} & -\frac{1}{N} & \ldots & 1-\frac{1}{N} \end{pmatrix}.$$

Therefore the projector $P_\perp$ operates on a vector $\boldsymbol{\beta}$ as

$$P_\perp \boldsymbol{\beta} = \boldsymbol{\beta} - \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\beta}_i (1, \ldots, 1) \equiv \boldsymbol{\beta} - \bar{\boldsymbol{\beta}}(1, \ldots, 1),$$

where $\bar{\boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{\beta}_i / N$ is the average of the elements of $\boldsymbol{\beta}$. Using this notation, we can rewrite the prior 8.9 as follows:

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})^2\right). \tag{8.10}$$

[1] An "orthant" is a quadrant in three or more dimensions.

This expression should be compared to the prior in equation 8.7: the crucial difference between these two is that while the prior in equation 8.7 shrinks $\boldsymbol{\beta}_i$ to a common, *predetermined* value γ, the prior in equation 8.10 simply shrinks them to some common value, which is not known a priori, but is determined by the data. The prior in equation 8.10 is similar to the empirical Bayes prior, with the difference that in empirical Bayes the value $\bar{\boldsymbol{\beta}}$ is replaced by an average of *empirical estimates* of the $\boldsymbol{\beta}_i$. It shares with the empirical Bayes prior the property of being independent of the absolute scale of $\boldsymbol{\beta}$, but it obviously does not require the empirical Bayesian theory of inference.

8.2.2 Hierarchical Models as Special Cases of Spatial Models

It is instructive to rewrite the prior in equation 8.10 in a way that makes it more similar to the conditionally autoregressive priors described earlier in this chapter. We notice that because $P_\perp$ is symmetric and is a projection operator, then $\|P_\perp \boldsymbol{\beta}\|^2 = \boldsymbol{\beta}' P_\perp' P_\perp \boldsymbol{\beta} = \boldsymbol{\beta}' P_\perp \boldsymbol{\beta}$. Because the rows of $P_\perp$ sum to 0, we can use the quadratic form identity of appendix B.2.6 (page 237) to rewrite the prior in equation 8.10 as

$$\mathcal{P}_\perp(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2N}\sum_{i,j=1}^{N}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)^2\right). \tag{8.11}$$

This prior has the same form of the priors described by the left column of equation 8.2, in which we have set $s_{ij} = 1$ for all $i, j = 1, \ldots N$, and therefore it is the simplest form of conditionally autoregressive prior. *This proves that hierarchical models are special cases of spatial models in which all elements of a cluster are defined to be "neighbors" of all other elements.* All results in this book described in the context of spatial models thus also apply to hierarchical models.

8.3 Smoothing Data without Forecasting

In several purposes researchers may be interested in smoothing observed mortality patterns, rather than forecasting future values. More precisely, they might have noisy and or incomplete mortality data and are interested in removing the noise or imputing the missing values.

This problem is easily handled in our framework and does not require the development of any new technique. As we describe here, it simply requires coding a set of dummy variables, one for each observation, and then applying our existing priors. As such, our existing software designed for forecasting can be used without modification.

We consider the case in which there is only one country. We have already defined a suitable set of priors for the expected value of the dependent variable:

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{1}{2}H[\mu, \theta]\right), \tag{8.12}$$

where the smoothness functional $H[\mu, \theta]$ will have, in general, a component for smoothing over age groups and a component for smoothing over time. If smoothness functionals over age groups and time as those in equations 5.8 (page 81) and 7.1 (page 125), respectively, are used, the discretized version of the smoothness functional is[2]

$$H[\mu, \theta] = \frac{\theta^{\text{age}}}{T} \sum_{aa't} W_{aa'}^{\text{age},\mathbb{n}} \mu_{at} \mu_{a't} + \frac{\theta^{\text{time}}}{A} \sum_{att'} W_{tt'}^{\text{time},\mathbb{k}} \mu_{at} \mu_{at'}.$$

In the preceding expression, $\mathbb{n}$ and $\mathbb{k}$ are the order of smoothness of the smoothness functional over age and time, respectively.

Unlike in the regression case, the quantity we are interested in is μ itself, and we do not need to link μ to a set of covariates here. Therefore, the smoothing problem consists simply of estimating μ given the prior 8.12 and the likelihood for this specification:

$$m_{at} \sim \mathcal{N}\left(\mu_{at}, \frac{\sigma_a^2}{b_{at}}\right) \quad a = 1, \ldots, A, \quad t = 1, \ldots T.$$

For clarity, we explicitly write down the negative log-posterior distribution for μ:

$$\log \mathcal{P}(\mu | m, \theta, \sigma) \propto \sum_{at} \frac{b_{at}}{\sigma_a^2} (m_{at} - \mu_{at})^2 + \left[\frac{\theta^{\text{age}}}{T} \sum_{aa't} W_{aa'}^{\text{age},\mathbb{n}} \mu_{at} \mu_{a't} + \frac{\theta^{\text{time}}}{A} \sum_{att'} W_{tt'}^{\text{time},\mathbb{k}} \mu_{at} \mu_{at'} \right]. \tag{8.13}$$

This expression fits squarely in the standard framework of nonparametric smoothing and can also be seen as a simple application of standard Bayesian smoothing theory.[3] Usually the estimate for μ is obtained by maximizing the posterior distribution, that is, by minimizing the expression in equation 8.13 over μ, using a variety of methods, including cross validation, to determine the parameters θ and σ. This approach takes advantage of the fact that the log-posterior is quadratic in μ, and linear methods can be used to solve part of the problem. Alternatively, one can develop a full Gibbs sampling strategy for the computation of the mean of the posterior.

In our case we do not need to develop new methods or even write new code. We observe that any estimation strategy used to solve the regression problem can be immediately applied to solve the smoothing problem by constructing an artificial set of covariates such that the regression coefficients can be interpreted as estimates of the expected value of the dependent variable.

In order to see this, consider the regression problem with the usual specification $\mu_{at} = \mathbf{Z}_{at} \boldsymbol{\beta}_a$. Now choose as covariates a set of T dummy variables, with one dummy variable associated with each year from 1 to T.[4] This is equivalent to using a covariate

[2] We are considering a zero mean prior here. If a nonzero mean prior is needed, the rest of the analysis remains the same, but μ is interpreted as the mean-centered age profile.

[3] If m_{at} includes missing values, we can simply fill them with an arbitrary number and set the weight b_{at} to zero.

[4] This implies that we drop the constant term for the specification.

matrix $\mathbf{Z}_a$ equal to the T-dimensional identity matrix. The specification $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$ can now be rewritten as

$$\mu_{at} = \beta_a^{(t)}.$$

In this way, estimates of the coefficients are easily translated into estimates for μ. A user who wishes to smooth the data and is not interested in forecasting has simply to decide the priors to use, create a set of dummy variables, and run any estimation algorithm.

8.4 Priors When the Dependent Variable Changes Meaning

We now consider an application of smoothing, as described in section 8.3, to a case not often considered in standard smoothing theory—smoothing in the presence of discontinuities.

A common problem in the analysis of cause-specific mortality rates is that the International Classification of Diseases (ICD), which is used to classify causes of death, changes roughly every decade. If a change is large enough, it could lead to visible discontinuities in the log-mortality time series, violating the assumptions that observed log-mortality is smooth over time.

For example, figure 8.1 presents the time series of log-mortality for "other infectious diseases" in males aged 0 to 4, for four different countries. The jumps in years 1968 and 1979 do not correspond to the sudden beginning and end of some worldwide epidemic with instant starting and stopping times, but rather a change in the way some infectious diseases have been coded. In particular they appear to reflect the adoption of ICD-8 codes in 1968 and of ICD-9 in 1979 (which we designate with vertical lines in each figure).

Several ways exist for dealing with data with one or more such jumps (other than ignoring the problem). One consists of fixing the problem by preprocessing, that is, modifying the time series in order to make it comparable across the whole period of analysis. This can sometimes be done using "comparability ratios," which attempt to translate one meaning (or ICD code change) into another. However, comparability ratios are often unavailable (after all, if such a simple translation were possible, the international public health establishment would probably not have gone to such lengths to change the ICD code in the first place), and so we are often stuck with discontinuous time series. In addition, a discontinuity may exist for other reasons than ICD revisions: for example, a country that was previously unable to report deaths from certain isolated regions might suddenly find the resources to increase coverage. Civil wars and other events often lead a country to sharply change its reporting practices (Murray et al., 2002).

In principle, the dependent variable changes meaning after every jump, and because we are interested in forecasting only the last meaning, the obvious thing to do is to discard all the data before the last jump. This, however, is extreme, because it assumes no correlation between meanings. An alternative consists of making some assumptions about how log-mortality before and after the jump are related. Here we consider the simplest assumption: the dynamics of log-mortality remain unchanged, except for a shift and a change in slope at

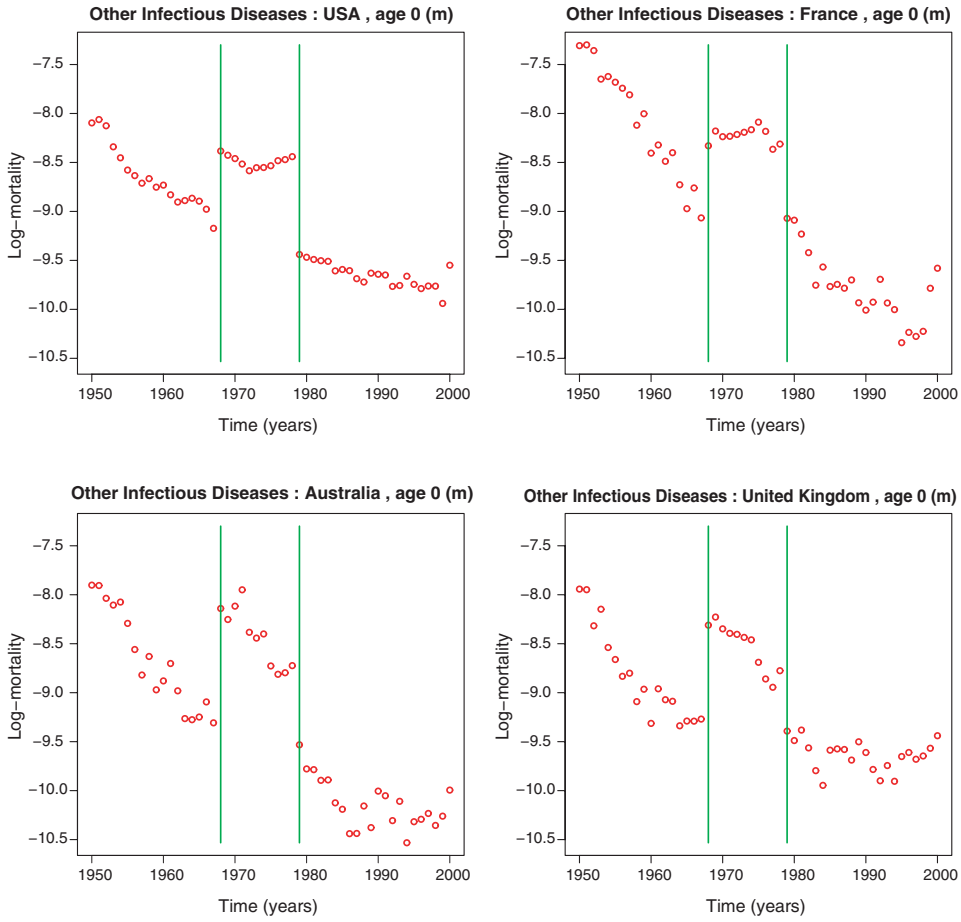

FIGURE 8.1. The effects of changes in ICD codes. Log-mortality for "Other Infectious Diseases" in males, aged 0–4, for four countries. The discontinuous behavior is likely due to changes in ICD codes. The green lines mark the years of the change.

the time of the change (which we assume known). This can be incorporated into the model by including two new variables among the covariates: one is an indicator variable that is 1 before the change and 0 after, and the other is linear before the change and 0 (or constant) after (i.e., an indicator variable for the change and an interaction between a time trend and the indicator variable).

Once these variables have been introduced, however, we also have to change our prior, because we no longer expect log-mortality to vary smoothly over time, and our smoothness assumption must be replaced by something weaker (less constraining). To do this, we denote by t^* the year in which the discontinuity occurs (the extension to data with more than one jump will be obvious). The new prior knowledge can be formulated as follows: log-mortality varies smoothly over time before t^* and after t^*, with no assumption imposed at the discontinuity. We now write a smoothness functional that encodes this knowledge. We do this by rewriting the generic smoothness functional over time

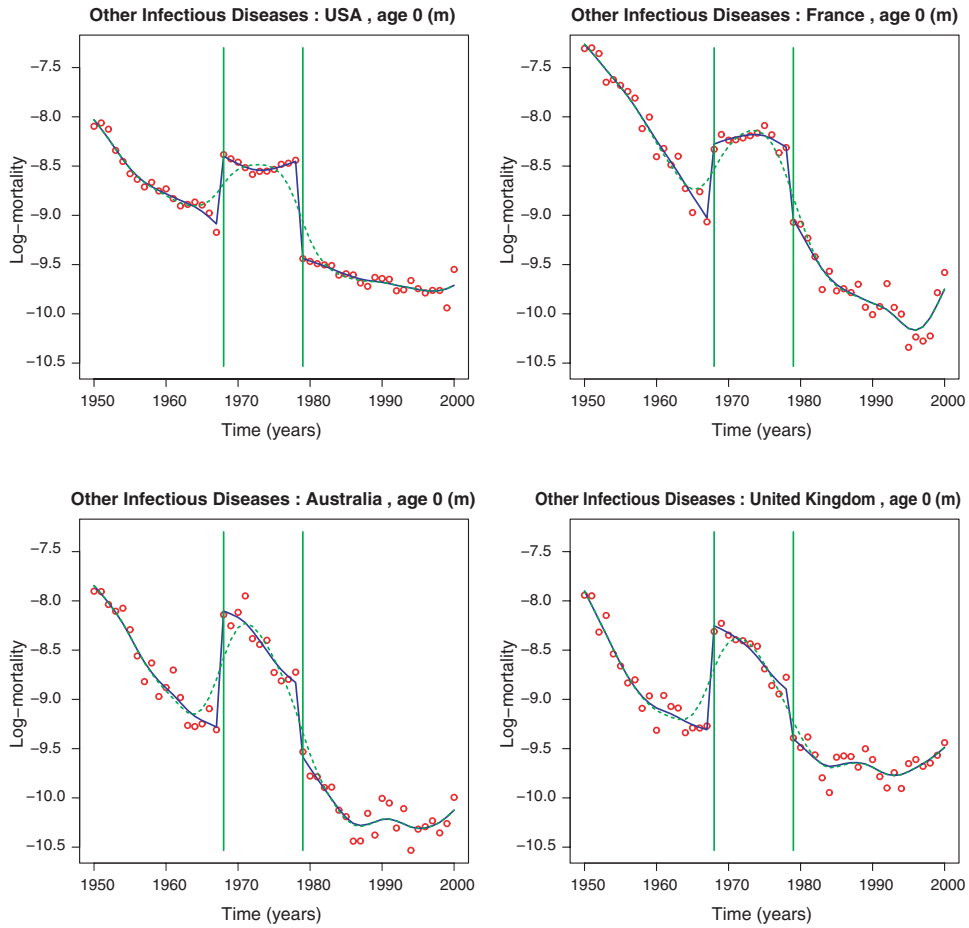

FIGURE 8.2. Modeling the effects of changes in ICD codes. Log-mortality for "Other Infectious Diseases" in males, aged 0–4, for four countries. The green curve smoothes the data with the standard smoothness functional, while the blue curve smoothes with the modified smoothness functional, allowing for discontinuities. The smoothness parameter θ has been set to 10, which is probably close to optimal.

of equation 7.1 (page 125) as follows:

$$H[\mu,\theta] \equiv \frac{\theta}{N}\sum_i \left[\int_0^{t^*} dw^{\text{time}}(t)\left(\frac{d^{\mathbb{m}}\mu(i,t)}{dt^{\mathbb{m}}}\right)^2 + \int_{t^*}^{T} dw^{\text{time}}(t)\left(\frac{d^{\mathbb{m}}\mu(i,t)}{dt^{\mathbb{m}}}\right)^2\right]. \quad (8.14)$$

The two-part integral in this smoothness functional has the desired property, because it enforces smoothness independently before and after the jump but does not penalize functions that have a jump at time t^*. The null space for this functional is the set of *piecewise* polynomials of degree $\mathbb{m}-1$, where the two "pieces" correspond to the period before and after t^*. Take, for example, the standard choice $\mathbb{m}=2$: this implies that we are indifferent to patterns of mortality that are linear in time, but with different slopes and

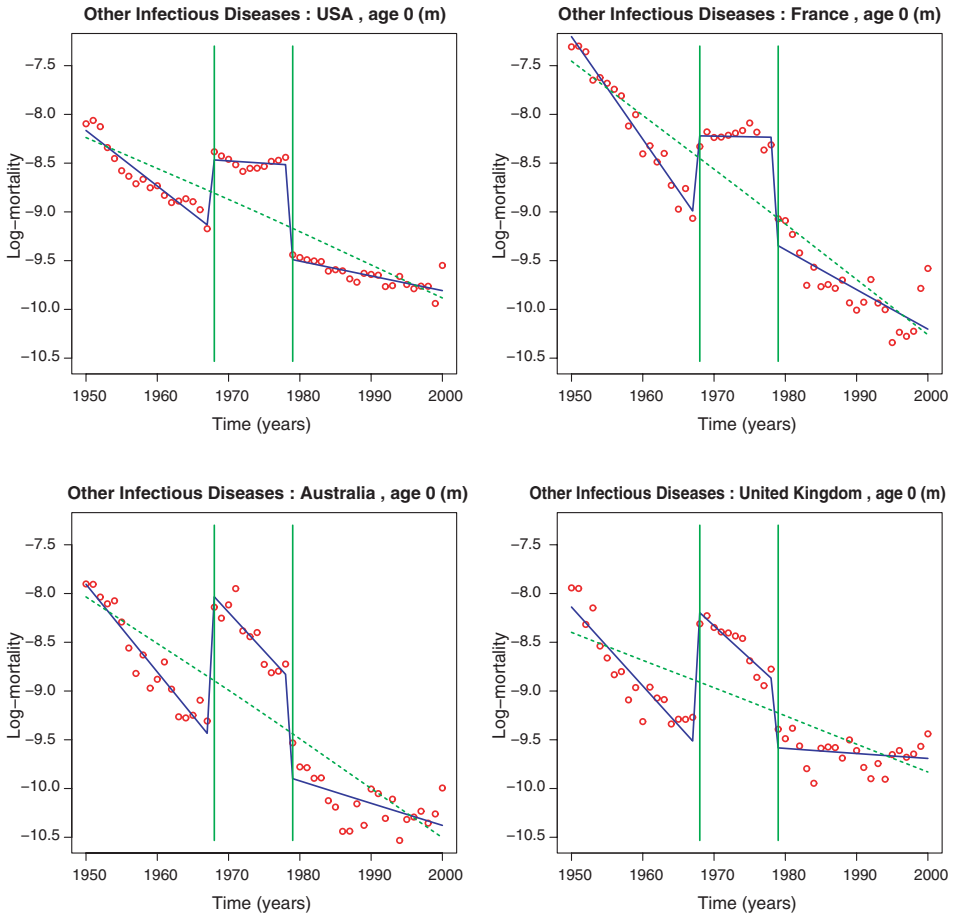

FIGURE 8.3. The null space for models of changes in ICD codes. Log-mortality for "Other Infectious Diseases" in males, aged 0–4, for four countries. The green curve smooths the data with the standard smoothness functional, while the blue curve smooths with the modified smoothness functional, allowing for discontinuities. The smoothness parameter θ has been set to 100,000, forcing the smoothed curve into the null space of the smoothness functional.

intercepts before and after the change. In other words, we make no assumptions about the coefficients of the two new variables.

Equation 8.14 underscores a key point that should always be kept in mind when choosing a prior: we must be clear about its domain of definition, that is, the set of functions we can plug into it. When we write the prior of equation 7.1 (page 125), we implicitly assume that log-mortality is at least continuous, because, if not, the functional assumes an infinite value. However, the prior in equation 8.14 is defined also for patterns of log-mortality that are discontinuous at t^*. In other words, the domain of definition of the functional in equation 8.14 is larger then the functional in equation 7.1, although the two functionals coincide at the domain of equation 7.1. Put differently, in principle, if we did not add the two variables there would be nothing gained by using functional 8.14 rather than functional 7.1 (in practice there would be something lost, due to the discretization

of the derivative operator, which is always poorer at the extrema of the domain of the integral).

In order to understand the difference between using the smoothness functionals in equations 8.14 and 7.1, we use both functionals, with $\mathbb{n} = 2$, to smooth (rather than forecast) the log-mortality patterns of figure 8.1 Because the standard smoothness functional 7.1 is "unaware" of the jumps, and it assumes that the underlying function is continuous, we expect it to make large errors around the jumps, resulting in oversmoothing in those areas. The functional in equation 8.14, which has been modified to include two jumps rather than one—one in year 1968 and one in year 1978—smoothes in the three regions independently. We report the results in figure 8.2 The red dots represent the data; the green dashed line, the results of smoothing with functional 7.1, which ignores the discontinuity; and the blue continuous line, the results of the functional 8.14. The smoothness parameter θ has been chosen large enough that the differences between the two smoothed curves are clear. These results are very pleasing, because the modified smoothness functional does exactly what it is supposed to do: it smoothes the data while preserving the discontinuities.

In order to check that the modified smoothness functional also has the right null space, we smooth the data with the same smoothness functionals, but with a near-infinity value of the smoothness parameter. In so doing, we force the smoothed curve to lie in the null space of the functional, thus ignoring the data wherever the prior has information and providing the best approximation to the data from the null space when the prior is completely uninformative. We report these results in figure 8.3, using the same color coding as before. Note that the green curve is a straight line, because the null space of the smoothness functional 7.1 with $\mathbb{n} = 2$ is the set of polynomials of degree 1. For the modified smoothness functional in equation 8.14, the smoothed curve is a *piecewise* polynomial of degree 1, as predicted by the theory.