Thus $\widehat{Var}(\widehat{\beta}) = \left(X^T\widehat{\Sigma}^{-1}X\right)^{-1}$ where $\widehat{\Sigma} = \widehat{\sigma}^2 H(\widehat{\phi}) + \widehat{\tau}^2 I$ and $X = [1, \texttt{long}, \texttt{lat}]$. Given the estimates of the sill, range, and nugget (from the `nls` function), it is possible to estimate the covariance matrix $\widehat{\Sigma}$, and thereby get $\widehat{Var}(\widehat{\beta})$. Develop an `S-plus` or `R` program to perform this exercise to obtain estimates of standard errors for $\widehat{\beta}$ for the scallops data.

Hint: $\hat{\tau}^2$ is the nugget; $\hat{\sigma}^2$ is the partial sill (the sill minus the nugget). Finally, the correlation matrix $H(\hat{\phi})$ can be obtained from the spherical covariance function, part of your solution to Exercise 1.

Note: It appears that `S+SpatialStats` uses the ordinary Euclidean (not geodesic) metric when computing distance, so you may use this as well when computing $H(\widehat{\phi})$. However, you may also wish to experiment with geodesic distances here, perhaps using your solution to Chapter 1, Exercise 7.

CHAPTER 3

Basics of areal data models

We now present a development of exploratory tools and modeling approaches that are customarily applied to data collected for areal units. We have in mind general, possibly irregular geographic units, but of course include the special case of regular grids of cells (pixels). Indeed, the ensuing models have been proposed for regular lattices of points and parameters, and sometimes even for point-referenced data (see Appendix A, Section A.5 on the problem of inverting very large matrices).

In the context of areal units the general inferential issues are the following:

(i) Is there spatial pattern? If so, how strong is it? Intuitively, "spatial pattern" suggests measurements for areal units that are near to each other will tend to take more similar values than those for units far from each other. Though you might "know it when you see it," this notion is evidently vague and in need of quantification. Indeed, with independent measurements for each unit we expect to see *no pattern*, i.e., a completely random arrangement of larger and smaller values. But again, randomness will inevitably produce some patches of similar values.

(ii) Do we want to smooth the data? If so, how much? Suppose, for example, that the measurement for each areal unit is a count, say, a number of cancers. Even if the counts were independent, and perhaps even after population adjustment, there would still be extreme values, as in any sample. Are the observed high counts more elevated than would be expected by chance? If we sought to present a surface of expected counts we might naturally expect that the high values would tend to be pulled down, the low values to be pushed up. This is the notion of smoothing. No smoothing would present a display using simply the observed counts. Maximal smoothing would result in a single common value for all units, clearly excessive. Suitable smoothing would fall somewhere in between, and take the spatial arrangement of the units into account.

Of course, how much smoothing is appropriate is not readily defined. In particular, for model-based smoothers such as we describe below, it is not evident what the extent of smoothing is, or how to control it.

Specification of a utility function for smoothing (as attempted in Stern and Cressie, 1999) would help to address these questions.

(iii) For a new areal unit or set of units, how can we infer about what data values we expect to be associated with these units? That is, if we modify the areal units to new units, e.g., from zip codes to census block groups, what can we say about the cancer counts we expect for the latter given those for the former? This is the so-called *modifiable areal unit problem (MAUP)*, which historically (and in most GIS software packages) is handled by crude areal allocation. Sections 6.2 and 6.3 propose model-based methodology for handling this problem.

As a matter of fact, in order to facilitate interpretation and better assess uncertainty, we will suggest model-based approaches to treat the above issues, as opposed to the more descriptive or algorithmic methods that have dominated the literature and are by now widely available in GIS software packages. We will also introduce further flexibility into these models by examining them in the context of regression. That is, we will assume that we have available potential covariates to explain the areal unit responses. These covariates may be available at the same or at different scales from the responses, but, regardless, we will now question whether there remains any spatial structure adjusted for these explanatory variables. This suggests that we may not try to model the data in a spatial way directly, but instead introduce spatial association through random effects. This will lead to versions of generalized linear mixed models (Breslow and Clayton, 1993). We will often view such models in the hierarchical fashion that is the primary theme of this text.

3.1 Exploratory approaches for areal data

We begin with the presentation of some tools that can be useful in the initial exploration of areal unit data. The primary concept here is a *proximity matrix*, W. Given measurements $Y_1, \ldots, Y_n$ associated with areal units $1, 2, \ldots, n$, the entries w_{ij} in W spatially connect units i and j in some fashion. (Customarily w_{ii} is set to 0.) Possibilities include binary choices, i.e., $w_{ij} = 1$ if i and j share some common boundary, perhaps a vertex (as in a regular grid). Alternatively, w_{ij} could reflect "distance" between units, e.g., a decreasing function of intercentroidal distance between the units (as in a county or other regional map). But distance can be returned to a binary determination. For example, we could set $w_{ij} = 1$ for all i and j within a specified distance. Or, for a given i, we could get $w_{ij} = 1$ if j is one of the K nearest (in distance) neighbors of i. The preceding choices suggest that W would be symmetric. However, for irregular areal units, this last example provides a setting where this need not be the case. Also, the w_{ij}'s may be standardized by $\sum_j w_{ij} = w_{i+}$. If $\widetilde{W}$ has entries $\widetilde{w}_{ij} = w_{ij}/w_{i+}$, then evidently $\widetilde{W}$ is row stochastic, i.e., $\widetilde{W}\mathbf{1} = \mathbf{1}$, but now $\widetilde{W}$ need not be symmetric.

As the notation suggests, the entries in W can be viewed as weights. More weight will be associated with j's closer (in some sense) to i than those farther away from i. In this exploratory context (but, as we shall see, more generally) W provides the mechanism for introducing spatial structure into our formal modeling.

Lastly, working with distance suggests that we can define distance bins, say, $(0, d_1], (d_1, d_2], (d_2, d_3]$, and so on. This enables the notion of *first-order neighbors* of unit i, i.e., all units within distance d_1 of i, *second-order neighbors*, i.e., all units more than d_1 but at most d_2 from i, *third-order neighbors*, and so on. Analogous to W we can define $W^{(1)}$ as the proximity matrix for first-order neighbors. That is, $w_{ij}^{(1)} = 1$ if i and j are first-order neighbors, and equal to 0 otherwise. Similarly we define $W^{(2)}$ as the proximity matrix for second-order neighbors; $w_{ij}^{(2)} = 1$ if i and j are second-order neighbors, and 0 otherwise, and so on to create $W^{(3)}$, $W^{(4)}$, etc.

Of course, the most obvious exploratory data analysis tool for lattice data is a map of the data values. Figure 3.1 gives the statewide average verbal SAT scores as reported by the College Board and initially analyzed by Wall (2004). Clearly these data exhibit strong spatial pattern, with midwestern states and Utah performing best, and coastal states and Indiana performing less well. Of course, before jumping to conclusions, we must realize there are any number of spatial covariates that may help to explain this pattern; the percentage of eligible students taking the exam, for instance (Midwestern colleges have historically relied on the ACT, not the SAT, and only the best and brightest students in these states would bother taking the latter exam). Still, the map of these raw data show significant spatial pattern.

3.1.1 *Measures of spatial association*

Two standard statistics that are used to measure strength of spatial association among areal units are Moran's I and Geary's C (see, e.g., Ripley, 1981, Sec. 5.4). These are spatial analogues of statistics for measuring association in time series, the lagged autocorrelation coefficient and the Durbin-Watson statistic, respectively. They can also be seen to be areal unit analogues of the empirical estimates for the correlation function and the variogram, respectively. Recall that, for point-referenced data, the empirical covariance function (2.15) and semivariogram (2.9), respectively, provide customary nonparametric estimates of these measures of association.

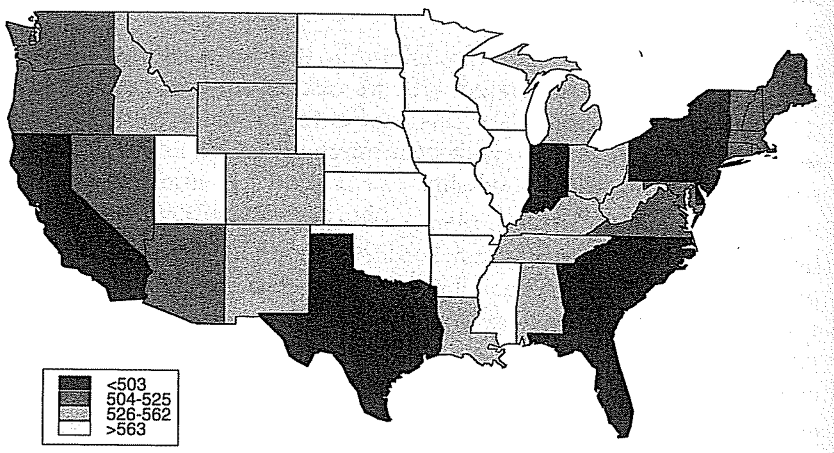

Figure 3.1 *Choropleth map of 1999 average verbal SAT scores, lower 48 U.S. states.*

Moran's I takes the form

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \overline{Y})^2} . \tag{3.1}$$

I is not strictly supported on the interval $[-1, 1]$. It is evidently a ratio of quadratic forms in $\mathbf{Y}$ that provides the idea for obtaining approximate first and second moments through the delta method (see, e.g., Agresti, 2002, Ch. 14). Moran shows under the null model where the Y_i are i.i.d., I is asymptotically normally distributed with mean $-1/(n-1)$ and a rather unattractive variance of the form

$$Var(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2} . \tag{3.2}$$

In (3.2), $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2$, and $S_2 = \sum_k (\sum_j w_{kj} + \sum_i w_{ik})^2$. We recommend the use of Moran's I as an exploratory measure of spatial association, rather than as a "test of spatial significance."

For the data mapped in Figure 3.1, we used the `spatial.cor` function in `S+SpatialStats` (see Section 2.5) to obtain a value for Moran's I of 0.5833, a reasonably large value. The associated standard error estimate of 0.0920 suggests very strong evidence against the null hypothesis of no spatial correlation in these data.

Geary's C takes the form

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \overline{Y})^2} . \tag{3.3}$$

C is never negative, and has mean 1 for the null model; *low* values (i.e., between 0 and 1) indicate *positive* spatial association. Also, C is a ratio of quadratic forms in $\mathbf{Y}$ and, like I, is asymptotically normal if the Y_i are i.i.d. We omit details of the distribution theory, recommending the interested reader to Cliff and Ord (1973), or Ripley (1981, p. 99).

Again using the `spatial.cor` function on the SAT verbal data in Figure 3.1, we obtained a value of 0.3775 for Geary's C, with an associated standard error estimate of 0.1008. Again, the marked departure from the mean of 1 indicates strong positive spatial correlation in the data.

If one truly seeks to run a significance test using (3.1) or (3.3), our recommendation is a Monte Carlo approach. Under the null model the distribution of I (or C) is invariant to permutation of the Y_i's. The exact null distribution of I (or C) requires computing its value under all $n!$ permutation of the Y_i's, infeasible for n in practice. However, a Monte Carlo sample of say 1000 permutations, including the observed one, will position the observed I (or C) relative to the remaining 999, to determine whether it is extreme (perhaps via an empirical p-value). Again using `spatial.cor` function on our SAT verbal data, we obtained empirical p-values of 0 using both Moran's I and Geary's C; *no* random permutation achieved I or C scores as extreme as those obtained for the actual data itself.

A further display that can be created in this spirit is the *correlogram.* Working with say I, in (3.1) we can replace w_{ij} with the previously defined $w_{ij}^{(1)}$ and compute say $I^{(1)}$. Similarly, we can replace w_{ij} with $w_{ij}^{(2)}$ and obtain $I^{(2)}$. A plot of $I^{(r)}$ vs. r is called a correlogram and, if spatial pattern is present, is expected to decline in r initially and then perhaps vary about 0. Evidently, this display is a spatial analogue of a temporal lag autocorrelation plot (e.g., see Carlin and Louis, 2000, p. 181). In practice, the correlogram tends to be very erratic and its information context is often not clear.

With large, regular grids of cells as we often obtain from remotely sensed imagery, it may be of interest to study spatial association in a particular direction (e.g., east-west, north-south, southwest-northeast, etc.). Now the spatial component reduces to one dimension and we can compute lagged autocorrelations (lagged appropriately to the size of the grid cells) in the specific direction. An analogue of this was proposed for the case where the Y_i are binary responses (e.g., presence or absence of forest in the cell) by Agarwal, Gelfand, and Silander (2002). In particular, Figure 3.2 shows rasterized maps of binary land use classifications for roughly 25,000 1 km

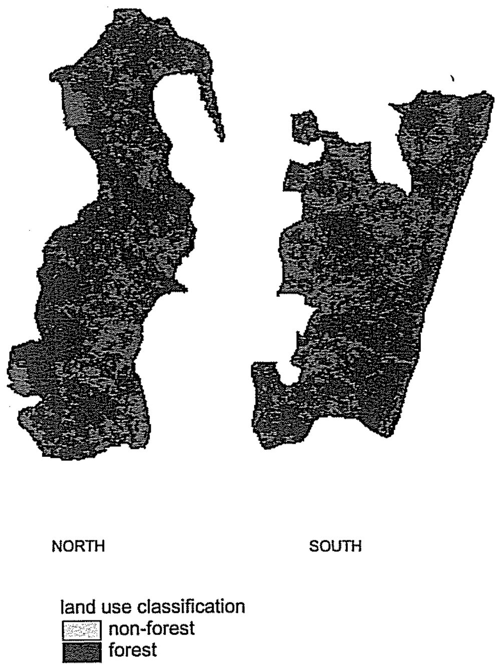

Figure 3.2 *Rasterized north and south regions (1 km × 1 km) with binary land use classification overlaid.*

× 1 km pixels in eastern Madagascar; see Agarwal et al. (2002) as well as Section 6.4 for further discussion.

While the binary map in Figure 3.2 shows spatial pattern in land use, we develop an additional display to provide quantification. For data on a regular grid or lattice, we calculate binary analogues of the sample autocovariances, using the 1 km × 1 km resolution with four illustrative directions: East (E), Northeast (NE), North (N), and Northwest (NW). Relative to a given pixel, we can identify all pixels in the region in a specified direction from that pixel and associate with each a distance (Euclidean distance centroid to centroid) from the given pixel. Pairing the response at the given pixel (X) with the response at a directional neighbor (Y), we obtain a correlated binary pair. Collecting all such (X,Y) pairs at a given direction/distance combination yields a 2×2 table of counts. The resultant log-odds ratio measures the association between pairs in that direction at that distance. (Note that if we followed the same procedure but reversed direction, e.g., changed from E to W, the corresponding log odds ratio would be unchanged.)

In Figure 3.3, we plot log odds ratio against direction for each of the four directions. Note that the spatial association is quite strong, requiring a distance of at least 40 km before it drops to essentially 0. This suggests that we would not lose much spatial information if we work with the lower (4 km × 4 km) resolution. In exchange we obtain a richer response variable (17 ordered levels, indicating number of forested cells from 0 to 16) and a substantial reduction in number of pixels (from 26,432 to 1,652 in the north region, from 24,544 to 1,534 in the south region) to facilitate model fitting.

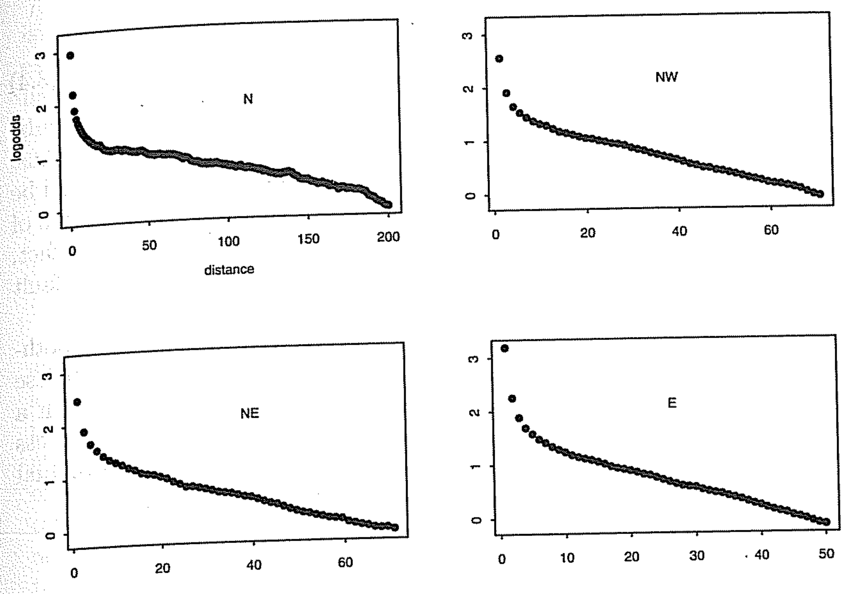

Figure 3.3 *Land use log-odds ratio versus distance in four directions.*

3.1.2 Spatial smoothers

Recall from the beginning of this chapter that often a goal for, say, a choropleth map of the Y_i's is *smoothing*. Depending upon the number of classes used to make the map, there is already some implicit smoothing in such a display (although this is not *spatial* smoothing, of course).

The W matrix directly provides a spatial smoother; that is, we can replace Y_i by $\widehat{Y}_i = \sum_j w_{ij} Y_j / w_{i+}$. This ensures that the value for areal unit i "looks like" its neighbors, and that the more neighbors we use in computing $\widehat{Y}_i$, the more smoothing we will achieve. In fact, $\widehat{Y}_i$ may be viewed as an unusual smoother in that it ignores the value actually observed for unit i. As such, we might revise the smoother to

$$\widehat{Y}_i^* = (1-\alpha)Y_i + \alpha\widehat{Y}_i \ , \tag{3.4}$$

where $\alpha \in (0,1)$. Working in an exploratory mode, various choices may be tried for α, but for any of these, (3.4) is a familiar *shrinkage* form. Thus, under a specific model with a suitable loss function, an optimal α could be sought. Finally, the form (3.4), viewed generally as a linear combination of the Y_j, is customarily referred to as a *filter* in the GIS literature. In fact, such software will typically provide choices of filters, and even a default filter to automatically smooth maps.

In Section 4.1 we will present a general discussion revealing how smoothing emerges as a byproduct of the hierarchical models we propose to use to explain the Y_i. In particular, when W is used in conjunction with a stochastic model (as in Section 3.3), the $\widehat{Y}_i$ are updated across i and across Monte Carlo iterations as well. So the observed Y_i will affect the eventual $\widehat{Y}_i$, and a "manual" inclusion of Y_i as in (3.4) is unnecessary.

3.2 Brook's Lemma and Markov random fields

A useful technical result for obtaining the joint distribution of the Y_i in some of the models we discuss below is *Brook's Lemma* (Brook, 1964). The usefulness of this lemma is exposed in Besag's (1974) seminal paper on conditionally autoregressive models.

It is clear that given $p(y_1, \ldots, y_n)$, the so-called *full conditional* distributions, $p(y_i|y_j, j \neq i)$, $i = 1, \ldots, n$, are uniquely determined. Brook's Lemma proves the converse and, in fact, enables us to constructively retrieve the unique joint distribution determined by these full conditionals. But first, it is also clear that we cannot write down an arbitrary set of full conditional distributions and assert that they uniquely determine the joint distribution. To see this, let $Y_1|Y_2 \sim N(\alpha_0+\alpha_1 Y_2, \sigma_1^2)$ and let $Y_2|Y_1 \sim N(\beta_0+\beta_1 Y_1^3, \sigma_2^2)$, where N denotes the normal (Gaussian) distribution. It is apparent that

$$E(Y_1) = E[E(Y_1|Y_2)] = E[\alpha_0 + \alpha_1 Y_2] = \alpha_0 + \alpha_1 E(Y_2) \ , \tag{3.5}$$

i.e., $E(Y_1)$ is linear in $E(Y_2)$ (hence $E(Y_2)$ is linear in $E(Y_1)$). But it must also be the case that

$$E(Y_2) = E[E(Y_2|Y_1)] = E[\beta_0 + \beta_1 Y_1] = \beta_0 + \beta_1 E(Y_1^3) \ . \tag{3.6}$$

Equations (3.5) and (3.6) could simultaneously hold only in trivial cases, so the two mean specifications are *incompatible*. Thus we can say that $f(y_1|y_2)$ and $f(y_2|y_1)$ are incompatible with regard to determining $p(y_1, y_2)$. We do not propose to examine conditions for compatibility here, although there has been considerable work in this area (see, e.g., Arnold and Strauss, 1991, and references therein).

Another point is that $p(y_1 \ldots, y_n)$ may be improper even if $p(y_i|y_j, j \neq i)$ is proper for all i. As an elementary illustration, consider $p(y_1, y_2) \propto \exp[-\frac{1}{2}(y_1 - y_2)^2]$. Evidently $p(y_1|y_2)$ is $N(y_2, 1)$ and $p(y_2|y_1)$ is $N(y_1, 1)$, but $p(y_1, y_2)$ is improper. Casella and George (1992) provide a similar example in a bivariate exponential (instead of normal) setting.

Brook's Lemma notes that

$$\begin{aligned} p(y_1, \ldots, y_n) \ &= \ \frac{p(y_1|y_2, \ldots, y_n)}{p(y_{10}|y_2, \ldots, y_n)} \cdot \frac{p(y_2|y_{10}, y_3, \ldots, y_n)}{p(y_{20}|y_{10}, y_3, \ldots, y_n)} \\ &\cdots \frac{p(y_n|y_{10}, \ldots, y_{n-1,0})}{p(y_{n0}|y_{10}, \ldots, y_{n-1,0})} \cdot p(y_{10}, \ldots, y_{n0}) \ , \end{aligned} \tag{3.7}$$

an identity you are asked to check in Exercise 1. Here, $\mathbf{y}_0 = (y_{10}, \ldots, y_{n0})'$ is any fixed point in the support of $p(y_1, \ldots, y_n)$. Hence $p(y_1, \ldots, y_n)$ is determined by the full conditional distributions, since apart from the constant $p(y_{10}, \ldots, y_{n0})$ they are the only objects appearing on the right-hand side of (3.7). Hence the joint distribution is determined up to a proportionality constant. If $p(y_1, \ldots, y_n)$ is improper then this is, of course, the best we can do; if $p(y_1, \ldots, y_n)$ is proper then the fact that it integrates to 1 determines the constant. Perhaps most important is the constructive nature of (3.7): we can create $p(y_1, \ldots, y_n)$ simply by calculating the product of ratios. For more on this point see Exercise 2.

Usually, when the number of areal units is very large (say, a large number of small geographic regions, or a regular grid of pixels on a screen), we do not seek to write down the joint distribution of the Y_i. Rather we prefer to work (and model) exclusively with the n corresponding full conditional distributions. In fact, from a spatial perspective we would think that the full conditional distribution for Y_i should really depend only upon the neighbors of cell i. Adopting some definition of a neighbor structure (e.g., the one setting $W_{ij} = 1$ or 0 depending on whether i and j are adjacent or not), let ∂_i denote the set of neighbors of cell i.

Next suppose we specify a set of full conditional distributions for the Y_i such that

$$p(y_i|y_j, j \neq i) = p(y_i|y_j, j \in \partial_i) \tag{3.8}$$

A critical question to ask is whether a specification such as (3.8) uniquely determines a joint distribution for $Y_1, \ldots Y_n$. That is, we do not need to see the explicit form of this distribution. We merely want to be assured that if, for example, we implement a Gibbs sampler (see Subsection 4.3.1) to simulate realizations from the joint distribution, that there is indeed a unique stationary distribution for this sampler.

The notion of using *local* specification to determine a joint (or global) distribution in the form (3.8) is referred to as a *Markov random field* (MRF). There is by now a substantial literature in this area, with Besag (1974) being a good place to start. Geman and Geman (1984) provide the next critical step in the evolution, while Kaiser and Cressie (2000) offer a current view and provide further references.

A critical definition in this regard is that of a *clique*. A clique is a set of cells (equivalently, indices) such that each element is a neighbor of every other element. With n cells, depending upon the definition of the neighbor structure, cliques can possibly be of size 1, 2, and so on up to size n. A *potential function* (or simply *potential*) of order k is a function of k arguments that is exchangeable in these arguments. The arguments of the potential would be the values taken by variables associated with the cells for a clique of size k. For continuous Y_i, a customary potential when $k = 2$ is Y_iY_j if i and j are a clique of size 2. (We use the notation $i \sim j$ if i is a neighbor of j and j is a neighbor of i.) For, say, binary Y_i, a potential when $k = 2$ is

$$I(Y_i = Y_j) = Y_iY_j + (1 - Y_i)(1 - Y_j) \, ,$$

where again $i \sim j$ and I denotes the indicator function. Throughout this book (and perhaps in most practical work as well), only cliques of order less than or equal to 2 are considered.

Next, we define a *Gibbs distribution* as follows: $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of the Y_i only through potentials on cliques. That is,

$$p(y_1, \ldots, y_n) \propto \exp\left\{ \gamma \sum_k \sum_{\boldsymbol{\alpha} \in \mathcal{M}_k} \phi^{(k)}(y_{\alpha_1}, y_{\alpha_2}, \ldots, y_{\alpha_k}) \right\} . \qquad (3.9)$$

Here, $\phi^{(k)}$ is a potential of order k, $\mathcal{M}_k$ is the collection of all subsets of size k from $\{1, 2, \ldots, n\}$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ indexes this set, and $\gamma > 0$ is a scale (or "temperature") parameter.

Informally, the *Hammersley-Clifford Theorem* (see Besag, 1974; also Clifford, 1990) demonstrates that if we have an MRF, i.e., if (3.8) defines a unique joint distribution, then this joint distribution is a Gibbs distribution. That is, it is of the form (3.9), with all of its "action" coming in the form of potentials on cliques. Cressie (1993, pp. 417–18) offers a proof of this theorem, and mentions that its importance for spatial modeling lies in its limiting the complexity of the conditional distributions required, i.e., full conditional distributions can be specified locally.

Geman and Geman (1984) provided essentially the converse of the Hammersley-Clifford theorem. If we begin with (3.9) we have determined an MRF. As a result, they argued that to sample a Markov random field, one could sample from its associated Gibbs distribution, hence coining the term "Gibbs sampler."

If we only use cliques of order 1, then the Y_i must be independent, as is evidenced by (3.9). For continuous data on $\Re^1$, a common choice for the

joint distribution is a pairwise difference form

$$p(y_1, \ldots, y_n) \propto \exp\left\{ -\frac{1}{2\tau^2} \sum_{i,j} (y_i - y_j)^2 I(i \sim j) \right\} . \qquad (3.10)$$

Distributions such as (3.10) will be the focus of the next section. For the moment, we merely note that it is a Gibbs distribution on potentials of order 1 and 2 and that

$$p(y_i \,\big|\, y_j, j \neq i) = N\left(\sum_{j \in \partial_i} y_i / m_i \, , \, \tau^2 / m_i \right) , \qquad (3.11)$$

where m_i is the number of neighbors of cell i. The distribution in (3.11) is clearly of the form (3.8) and shows that the mean of Y_i is the average of its neighbors.

3.3 Conditionally autoregressive (CAR) models

Although they were introduced by Besag (1974) approximately 30 years ago, conditionally autoregressive (CAR) models have enjoyed a dramatic increase in usage only in the past decade or so. This resurgence arises from their convenient employment in the context of Gibbs sampling and more general Markov chain Monte Carlo (MCMC) methods for fitting certain classes of hierarchical spatial models (seen, e.g., in Section 5.4.3).

3.3.1 The Gaussian case

We begin with the Gaussian (or *autonormal*) case. Suppose we set

$$Y_i \mid y_j, j \neq i \sim N\left(\sum_j b_{ij} y_j \, , \, \tau_i^2 \right) , \; i = 1, \ldots, n \, . \qquad (3.12)$$

These full conditionals are compatible, so through Brook's Lemma we can obtain

$$p(y_1, \ldots, y_n) \propto \exp\left\{ -\frac{1}{2} \mathbf{y}' D^{-1} (I - B) \mathbf{y} \right\} , \qquad (3.13)$$

where $B = \{b_{ij}\}$ and D is diagonal with $D_{ii} = \tau_i^2$. Expression (3.13) suggests a joint multivariate normal distribution for $\mathbf{Y}$ with mean $\mathbf{0}$ and variance matrix $\Sigma_{\mathbf{y}} = (I - B)^{-1} D$.

But we are getting ahead of ourselves. First, we need to ensure that $D^{-1}(I - B)$ is symmetric. The simple resulting conditions are

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \quad \text{for all } \; i, j \, . \qquad (3.14)$$

Evidently, from (3.14), B is not symmetric. Returning to our proximity matrix W (which we assume to be symmetric), suppose we set $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$. Then (3.14) is satisfied and (3.12) yields $p(y_i|y_j, j \neq i) = N\left(\sum_j w_{ij}y_j/w_{i+}\ ,\ \tau^2/w_{i+}\right)$. Also, (3.13) becomes

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - W)\mathbf{y}\right\}\ , \tag{3.15}$$

where D_w is diagonal with $(D_w)_{ii} = w_{i+}$.

Now a second problem is noticed. $(D_w - W)\mathbf{1} = \mathbf{0}$, i.e., $\Sigma_{\mathbf{y}}^{-1}$ is singular, so that $\Sigma_{\mathbf{y}}$ does not exist and the distribution in (3.15) is improper. (The reader is encouraged to note the difference between the case of $\Sigma_{\mathbf{y}}^{-1}$ singular and the case of $\Sigma_{\mathbf{y}}$ singular. With the former we have a density function but one that is not integrable; effectively we have too many variables and we need a constraint on them to restore propriety. With the latter we have no density function but a proper distribution that resides in a lower dimensional space; effectively we have too *few* variables.) With a little algebra (3.15) can be rewritten as

$$p(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i \neq j} w_{ij}(y_i - y_j)^2\right\}\ . \tag{3.16}$$

This is a pairwise difference specification slightly more general than (3.10). But the impropriety of $p(\mathbf{y})$ is also evident from (3.16) since we can add any constant to all of the Y_i and (3.16) is unaffected; the Y_i are not "centered." A constraint such as $\sum_i Y_i = 0$ would provide the needed centering. Thus we have a more general illustration of a joint distribution that is improper, but has all full conditionals proper. The specification (3.16) is often referred to as an *intrinsically autoregressive* (IAR) model.

As a result, $p(\mathbf{y})$ in (3.15) cannot be used as a model for data; data could not arise under an improper stochastic mechanism, and we cannot impose a constant center on randomly realized measurements. Hence, the use of an improper autonormal model must be relegated to a *prior* distributional specification. That is, it will be attached to random spatial effects introduced at the second stage of a hierarchical specification (again, see e.g. Section 5.4.3).

The impropriety in (3.15) can be remedied in an obvious way. Redefine $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$ and choose ρ to make $\Sigma_{\mathbf{y}}^{-1}$ nonsingular. This is guaranteed if $\rho \in \left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2}WD_w^{-1/2}$; see Exercise 5. Moreover, since $tr(D_w^{-1/2}WD_w^{-1/2}) = 0 = \sum_{i=1}^n \lambda_{(i)}$, $\lambda_{(1)} < 0$, $\lambda_{(n)} > 0$, and 0 belongs to $\left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$.

Simpler bounds than those given above for the propriety parameter ρ may

be obtained if we replace the adjacency matrix W by the scaled adjacency matrix $\widetilde{W} \equiv Diag(1/w_{i+})W$; recall $\widetilde{W}$ is not symmetric, but it will be row stochastic (i.e., all of its rows sum to 1). $\Sigma_{\mathbf{y}}^{-1}$ can then be written as $M^{-1}(I - \alpha\widetilde{W})$ where M is diagonal. Then if $|\alpha| < 1$, $I - \alpha\widetilde{W}$ is nonsingular. (See the SAR model of the next section, as well as Exercise 7.) Carlin and Banerjee (2003) show that $\Sigma_{\mathbf{y}}^{-1}$ is diagonally dominant and symmetric. But diagonally dominant symmetric matrices are positive definite (Harville, 1997), providing an alternative argument for the propriety of the joint distribution.

Returning to the unscaled situation, ρ can be viewed as an additional parameter in the CAR specification, enriching this class of spatial models. Furthermore, $\rho = 0$ has an immediate interpretation: the Y_i become independent $N(0, \tau^2/w_{i+})$. If ρ is not included, independence cannot emerge as a limit of (3.15). (Incidentally, this suggests a clarification of the role of τ^2, the variance parameter associated with the full conditional distributions: the magnitude of τ^2 should *not* be viewed as in any way quantifying the strength of spatial association. Indeed if all Y_i are multiplied by c, τ^2 becomes $c\tau^2$ but the strength of spatial association among the Y_i is clearly unaffected.) Lastly, $\rho\sum_j w_{ij}Y_j/w_{i+}$ can be viewed as a *reaction function*, i.e., ρ is the expected proportional "reaction" of Y_i to $\sum_j w_{ij}Y_j/w_{i+}$.

With these advantages plus the fact that $p(\mathbf{y})$ (or the Bayesian posterior distribution, if the CAR specification is used to model constrained random effects) is now proper, is there any reason not to introduce the ρ parameter? In fact, the answer may be yes. Under $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$, the full conditional $p(y_i|y_j, j \neq i)$ becomes $N\left(\rho\sum_j w_{ij}y_j/w_{i+}\ ,\ \tau^2/w_{i+}\right)$. Hence we are modeling Y_i not to have mean that is an average of its neighbors, but some *proportion* of this average. Does this enable any sensible spatial interpretation for the CAR model? Moreover, does ρ calibrate very well with any familiar interpretation of "strength of spatial association?" Fixing $\tau^2 = 1$ without loss of generality, we can simulate CAR realizations for a given n, W, and ρ. We can also compute for these realizations a descriptive association measure such as Moran's I or Geary's C. Here we do not present explicit details of the range of simulations we have conducted. However, for a 10 × 10 grid using a first-order neighbor system, when $\rho = 0.8$, I is typically 0.1 to 0.15; when $\rho = 0.9$, I is typically 0.2 to 0.25; and even when $\rho = 0.99$, I is typically at most 0.5. It thus appears that ρ can mislead with regard to strength of association. Expressed in a different way, within a Bayesian framework, a prior on ρ that encourages a consequential amount of spatial association would place most of its mass near 1.

A related point is that if $p(\mathbf{y})$ is proper, the breadth of spatial pattern may be too limited. In the case where a CAR model is applied to random effects, an improper choice may actually enable wider scope for posterior spatial pattern. As a result, we do not take a position with regard to propriety or impropriety in employing CAR specifications (though in the remainder of this text we do sometimes attempt to illuminate relative advantages and disadvantages).

Referring to (3.12), we may write the entire system of random variables as

$$\mathbf{Y} = B\mathbf{Y} + \boldsymbol{\epsilon}\,, \quad \text{or equivalently,} \tag{3.17}$$

$$(I - B)\mathbf{Y} = \boldsymbol{\epsilon}\,. \tag{3.18}$$

In particular, the distribution for $\mathbf{Y}$ induces a distribution for $\boldsymbol{\epsilon}$. If $p(\mathbf{y})$ is proper then $\mathbf{Y} \sim N(\mathbf{0}, (I-B)^{-1}D)$ whence $\boldsymbol{\epsilon} \sim N(\mathbf{0}, D(I-B)')$, i.e., the components of $\boldsymbol{\epsilon}$ are not independent. Also, $Cov(\boldsymbol{\epsilon}, \mathbf{Y}) = D$.

When $p(\mathbf{y})$ is proper we can appeal to standard multivariate normal distribution theory to interpret the entries in $\Sigma_{\mathbf{Y}}^{-1}$. For example, $1/(\Sigma_{\mathbf{Y}}^{-1})_{ii} = Var(Y_i|Y_j, j \neq i)$. Of course with $\Sigma_{\mathbf{Y}}^{-1} = D^{-1}(I-B)$, $(\Sigma_{\mathbf{Y}}^{-1})_{ii} = 1/\tau_i^2$ providing immediate agreement with (3.12). But also, if $(\Sigma_{\mathbf{Y}}^{-1})_{ij} = 0$, then Y_i and Y_j are conditionally independent given $Y_k, k \neq i, j$, a fact you are asked to show in Exercise 8. Hence if any $b_{ij} = 0$, we have conditional independence for that pair of variables. Connecting b_{ij} to w_{ij} shows that the choice of neighbor structure implies an associated collection of conditional independences. With first-order neighbor structure, all we are asserting is a spatial illustration of the local Markov property (Whittaker, 1990, p. 68).

We conclude this subsection with three remarks. First, one can directly introduce a regression component into (3.12), e.g., a term of the form $\mathbf{x}_i'\boldsymbol{\beta}$. Conditional on $\boldsymbol{\beta}$, this does not affect the association structure that ensues from (3.12); it only revises the mean structure. However, we omit details here (the interested reader can consult Besag, 1974), since we will only use the autonormal CAR as a distribution for spatial random effects. These effects are added onto the regression structure for the mean on some transformed scale (again, see Section 5.4.3).

We also note that in suitable contexts it may be appropriate to think of $\mathbf{Y}_i$ as a vector of dependent areal unit measurements or, in the context of random effects, as a vector of dependent random effects associated with an areal unit. This leads to the specification of multivariate conditionally autoregressive (MCAR) models, which is the subject of Section 7.4. From a somewhat different perspective, $\mathbf{Y}_i$ might arise as $(Y_{i1}, \ldots, Y_{iT})'$ where Y_{it} is the measurement associated with areal unit i at time t, $t = 1, \ldots, T$. Now we would of course think in terms of spatiotemporal modeling for Y_{it}. This is the subject of Section 8.5.

Lastly, a (proper) CAR model can in principle be used for point-level data, taking w_{ij} to be, say, an inverse distance between points i and j. However, unlike the spatial prediction described in Section 2.4, now spatial prediction becomes *ad hoc*. That is, to predict at a new site Y_0, we might specify the distribution of Y_0 given $Y_1, \ldots, Y_n$ to be a normal distribution, such as a $N\left(\rho \sum_j w_{0j} y_j / w_{0+}\,,\, \tau^2 / w_{0+}\right)$. Note that this determines the joint distribution of $Y_0, Y_1, \ldots, Y_n$. However, this joint distribution is *not* the CAR distribution that would arise by specifying the full conditionals for $Y_0, Y_1, \ldots, Y_n$ and using Brook's Lemma, as in constructing (3.15).

3.3.2 The non-Gaussian case

If one seeks to model the data directly using a CAR specification then in many cases a normal distribution would not be appropriate. Binary response data and sparse count data are two examples. In fact, one can select any exponential family model as a first-stage distribution for the data and propose

$$p\left(y_i | y_j, j \neq i\right) \propto \exp\left(\{\psi\left(\theta_i y_i - \chi\left(\theta_i\right)\right)\}\right)\,, \tag{3.19}$$

where, adopting a canonical link, $\theta_i = \sum_{j \neq i} b_{ij} y_j$ and ψ is a non-negative dispersion parameter. In fact (3.19) simplifies to

$$p\left(y_i | y_j, j \neq i\right) \propto \exp\left(\psi \sum_{j \neq i} w_{ij} y_i y_j\right)\,. \tag{3.20}$$

Since the data are being modeled directly, it may be appropriate to introduce a nonautoregressive linear regression component to (3.20). That is, we can write $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j \neq i} b_{ij} y_j$, for some set of covariates $\mathbf{x}_i$. After obvious reparametrization (3.20) becomes

$$p\left(y_i | y_j, j \neq i\right) \propto \exp\left(\mathbf{x}_i^T \gamma + \psi \sum_{j \neq i} b_{ij} y_j\right)\,. \tag{3.21}$$

In the case where the Y_i are binary, a particular version, which has received recent attention in the literature, is the *autologistic* model; see, e.g., Heikkinen and Hogmander (1994), Hogmander and Møller (1995), and Hoeting et al. (2000). Here,

$$\log \frac{P\left(Y_i = 1\right)}{P\left(Y_i = 0\right)} = \mathbf{x}_i^T \gamma + \psi \sum w_{ij} y_j\,, \tag{3.22}$$

where $w_{ij} = 1$ if $i \sim j$, $= 0$ otherwise. Using Brook's Lemma the joint distribution of $Y_1, \ldots, Y_n$ can be shown to be

$$p\left(y_1, \ldots, y_n\right) \propto \exp\left(\gamma^T \left(\sum_i y_i \mathbf{x}_i\right) + \psi \sum_{i,j} w_{ij} y_i y_j\right)\,. \tag{3.23}$$

Expression (3.23) shows that f is indeed a Gibbs distribution and appears to be an attractive form. But for likelihood or Bayesian inference, the normalizing constant is required, since it is a function of γ and ψ. However, computation of this constant requires summation over all of the 2^n possible values that $(Y_1, Y_2, ..., Y_n)$ can take on. Even for moderate sample sizes this will present computational challenges. Hoeting et al. (2000) propose approximations to the likelihood using a pseudo-likelihood and a normal approximation.

The case where Y_i can take on one of several categorical values presents a natural extension to the autologistic model. If we label the (say) L possible outcomes as simply $1, 2, ..., L$, then we can define

$$P\left(Y_i = l \mid Y_j, j \neq i\right) \propto \exp\left(\psi \sum_{j \neq i} w_{ij} I\left(Y_j = l\right)\right) , \qquad (3.24)$$

with w_{ij} as above. The distribution in (3.24) is referred to as a *Potts model*. It obviously extends the binary case and encourages Y_i to be like its neighbors. It also suffers from the normalization problem. It can also be employed as a random effects specification, as an alternative to an autonormal; see Green and Richardson (2002) in this regard.

3.4 Simultaneous autoregressive (SAR) models

Returning to (3.17), suppose that instead of letting $\mathbf{Y}$ induce a distribution for $\boldsymbol{\epsilon}$, we let $\boldsymbol{\epsilon}$ induce a distribution for $\mathbf{Y}$. Imitating usual autoregressive time series modeling, suppose we take the ϵ_i to be independent innovations. For a little added generality, assume that $\boldsymbol{\epsilon} \sim N\left(0, \tilde{D}\right)$ where $\tilde{D}$ is diagonal with $\left(\tilde{D}\right)_{ii} = \sigma_i^2$. (Note $\tilde{D}$ has no connection with D in Section 3.3; the B we use below may or may not be the same as the one we used in that section.) Analogous to (3.12), now $Y_i = \sum_j b_{ij} Y_j + \epsilon_i$, $i = 1, 2, ..., n$, with $\epsilon_i \sim N\left(0, \sigma_i^2\right)$. Therefore, if $(I - B)$ is full rank,

$$\mathbf{Y} \sim N\left(\mathbf{0}\,,\, (I - B)^{-1} \tilde{D} \left((I - B)^{-1}\right)'\right) . \qquad (3.25)$$

Also, $Cov(\boldsymbol{\epsilon}, \mathbf{Y}) = \tilde{D}(I - B)^{-1}$. If $\tilde{D} = \sigma^2 I$ then (3.25) simplifies to $\mathbf{Y} \sim N\left(\mathbf{0}\,,\, \sigma^2 \left[(I - B)(I - B)'\right]^{-1}\right)$. In order that (3.25) be proper, $I - B$ must be full rank. Two choices are most frequently discussed in the literature (e.g., Griffith, 1988). The first assumes $B = \rho W$, where W is a so-called contiguity matrix, i.e., W has entries that are 1 or 0 according to whether or not unit i and unit j are direct neighbors (with $w_{ii} = 0$). So W is our familiar first-order neighbor proximity matrix. Here ρ is called a *spatial autoregression parameter* and, evidently, $Y_i = \rho \sum_j Y_j I(j \in \partial_i) + \epsilon_i$, where ∂_i denotes the set of neighbors of i. In fact, any proximity matrix can be used and, paralleling the discussion below (3.15), $I - \rho W$ will be nonsingular if $\rho \in \left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$ where now $\lambda_{(1)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of W.

Alternatively, W can be replaced by $\widetilde{W}$ where now, for each i, the ith row has been normalized to sum to 1. That is, $\left(\tilde{W}\right)_{ij} = w_{ij}/w_{i+}$. Again, $\widetilde{W}$ is not symmetric, but it is row stochastic, i.e., $\widetilde{W}\mathbf{1} = \mathbf{1}$. If we set $B = \alpha\widetilde{W}$, α is called a *spatial autocorrelation parameter* and, were W a contiguity matrix, now $Y_i = \alpha \sum_j Y_i I(j \in \partial_i)/w_{i+} + \epsilon_i$. With a very regular grid the w_{i+} will all be essentially the same and thus α will be a multiple of ρ. But, perhaps more importantly, with $\widetilde{W}$ row stochastic the eigenvalues of $\widetilde{W}$ are all less than or equal to 1 (i.e., max $|\lambda_i| = 1$). Thus $I - \alpha\widetilde{W}$ will be nonsingular if $\alpha \in (-1, 1)$, justifying referring to α as an autocorrelation parameter; see Exercise 7.

A SAR model is customarily introduced in a regression context, i.e., the *residuals* $\mathbf{U} = \mathbf{Y} - X\boldsymbol{\beta}$ are assumed to follow a SAR model, rather than $\mathbf{Y}$ itself. But then, following (3.17), if $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, we obtain the attractive form

$$\mathbf{Y} = B\mathbf{Y} + (I - B)\, X\boldsymbol{\beta} + \boldsymbol{\epsilon} . \qquad (3.26)$$

Expression (3.26) shows that $\mathbf{Y}$ is modeled through a component that provides a spatial weighting of neighbors and a component that is a usual linear regression. If B is the zero matrix we obtain an OLS regression; if $B = I$ we obtain a purely spatial model.

We note that from (3.26) the SAR model does not introduce any spatial effects; the errors in (3.26) are independent. Expressed in a different way, if we modeled $\mathbf{Y} = X\boldsymbol{\beta}$ as $\mathbf{U} + \mathbf{e}$ with $\mathbf{e}$ independent errors, we would have $\mathbf{U} + \mathbf{e} = B\mathbf{U} + \boldsymbol{\epsilon} + \mathbf{e}$ and $\boldsymbol{\epsilon} + \mathbf{e}$ would result in a redundancy. As a result, in practice a SAR specification is not used in conjunction with a GLM. To introduce $\mathbf{U}$ as a vector of spatial adjustments to the mean vector, a transformed scale creates redundancy between the independent Gaussian error in the definition of the U_i and the stochastic mechanism associated with the conditionally independent Y_i.

We briefly note the somewhat related spatial modeling approach of Langford et al. (1999). Rather than modeling the residual vector $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, they propose that $\mathbf{U} = \tilde{B}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 I\right)$, i.e., that $\mathbf{U}$ be modeled as a spatially motivated linear combination of independent variables. This induces $\Sigma_U = \sigma^2 \tilde{B}\tilde{B}^T$. Thus, the U_i and hence the Y_i will be dependent and given $\tilde{B}$, $cov\left(Y_i, Y_{i'}\right) = \sigma^2 \sum_j b_{ij} b_{i'j}$. If B arises through some proximity matrix W, the more similar rows i and i' of W are, the stronger the association between Y_i and $Y_{i'}$. However, the difference in nature between this specification and that in (3.26) is evident. To align the two, we would set $(I - B)^{-1} = \tilde{B}$, i.e. $B = I - \tilde{B}^{-1}$ (assuming $\tilde{B}$ is of full rank). $I - \tilde{B}^{-1}$ would not appear to have any interpretation through a proximity matrix.

Perhaps the most important point to note with respect to SAR models is that they are well suited to maximum likelihood estimation but not at all for MCMC fitting of Bayesian models. That is, the log likelihood associated with (3.26) (assuming $\tilde{D} = \sigma^2 I$) is

$$\frac{1}{2}\log\left|\sigma^{-1}(I-B)\right| - \frac{1}{2\sigma^2}(\mathbf{Y}-X\beta)^T(I-B)(I-B)^T(\mathbf{Y}-X\beta)\,. \quad (3.27)$$

Though B will introduce a regression or autocorrelation parameter, the quadratic form in (3.27) is quick to calculate (requiring no inverse) and the determinant can usually be calculated rapidly using diagonally dominant, sparse matrix approximations (see, e.g., Pace and Barry, 1997a,b). Thus maximization of (3.27) can be done iteratively but, in general, efficiently.

Also, note that while the form in (3.27) can certainly be extended to a full Bayesian model through appropriate prior specifications, the absence of a hierarchical form with random effects implies straightforward Bayesian model fitting as well. Indeed, the general spatial slice Gibbs sampler (see Appendix Section A.6, or Agarwal and Gelfand, 2002) can easily handle this model. However, suppose we attempt to introduce SAR random effects in some fashion. Unlike CAR random effects that are defined through full conditional distributions, the full conditional distributions for the SAR effects have no convenient form. For large n, computation of such distributions using a form such as (3.25) will be expensive.

SAR models as in (3.26) are frequently employed in the spatial econometrics literature. With point-referenced data, B is taken to be ρW where W is the matrix of interpoint distances. Likelihood-based inference can be implemented in `S+SpatialStats` as well as more specialized software, such as that from the Spatial Analysis Laboratory (`sal.agecon.uiuc.edu`)). Software for large data sets is supplied there, as well as through the website of Prof. Kelley Pace, `www.spatial-statistics.com`. An illustrative example is provided in Exercise 10.

CAR versus SAR models

Cressie (1993, pp. 408–10) credits Brook (1964) with being the first to make a distinction between the CAR and SAR models, and offers a comparison of the two. To begin with, we may note from (3.13) and (3.25) that the two forms are equivalent if and only if

$$(I-B)^{-1}D = (I-\tilde{B})^{-1}\tilde{D}((I-\tilde{B})^{-1})'\,,$$

where we use the tilde to indicate matrices in the SAR model. Cressie then shows that any SAR model can be represented as a CAR model (since D is diagonal), but gives a counterexample to prove that the converse is not true. For the "proper" CAR and SAR models that include spatial correlation parameters ρ, Wall (2004) shows that the correlations between neighboring regions implied by these two models can be rather different; in particular, the first-order neighbor correlations increase at a slower rate

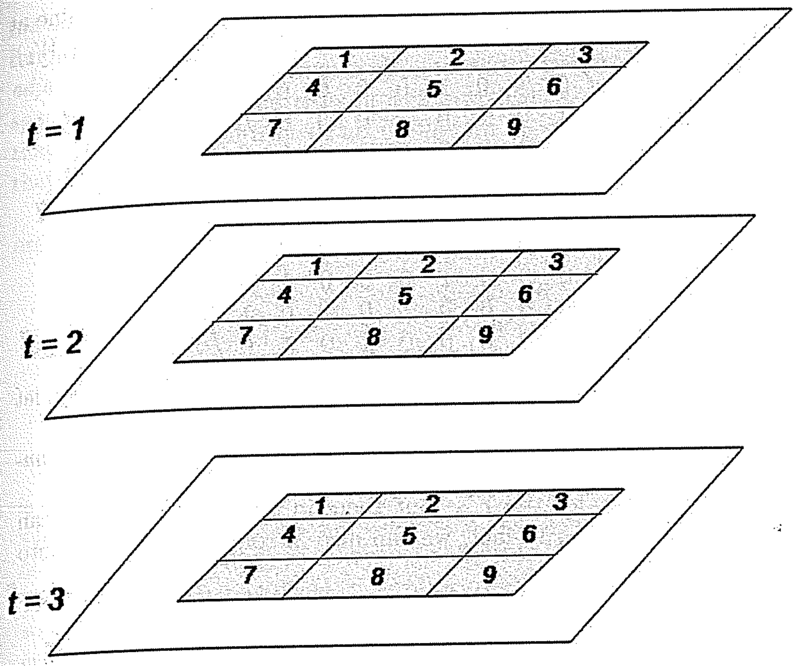

Figure 3.4 *Illustration of spatiotemporal areal unit setting for STAR model.*

as a function of ρ in the CAR model than they do for the SAR model. (As an aside, she notes that these correlations are not even monotone for $\rho < 0$, another reason to avoid negative spatial correlation parameters.) Also, correlations among pairs can switch in nonintuitive ways. For example, when working with the adjacency relationships generated by the lower 48 contiguous U.S. states, she finds that when $\rho = .49$ in the CAR model, $Corr(Alabama, Florida) = .20$ and $Corr(Alabama, Georgia) = .16$. But when ρ increases to .975, we instead get $Corr(Alabama, Florida) = .65$ and $Corr(Alabama, Georgia) = .67$, a slight reversal in ordering.

STAR models

In the literature SAR models have frequently been extended to handle spatiotemporal data. The idea is that in working with proximity matrices, we can define neighbors in time as well as in space. Figure 3.4 shows a simple illustration with 9 areal units, 3 temporal units for each areal unit yielding $i = 1, \ldots, 9$, $t = 1, 2, 3$, labeled as indicated.

The measurements Y_{it} are spatially associated at each fixed t. But also, we might seek to associate, say, Y_{i2} with Y_{i1} and Y_{i3}. Suppose we write Y

as the 27×1 vector with the first nine entries at $t = 1$, the second nine at $t = 2$, and the last nine at $t = 3$. Also let $W_S = BlockDiag(W_1, W_1, W_1)$, where

$$W_1 = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} .$$

Then W_S provides a spatial contiguity matrix for the Y's. Similarly, let $W_T = \begin{pmatrix} 0 & W_2 & 0 \\ W_2 & 0 & W_2 \\ 0 & W_2 & 0 \end{pmatrix}$, where $W_2 = I_{3\times3}$. Then W_T provides a *temporal* contiguity matrix for the Y's. But then, in our SAR model we can define $B = \rho_s W_S + \rho_t W_T$. In fact, we can also introduce $\rho_{ST} W_S W_T$ into B and note that

$$W_S W_T = \begin{pmatrix} 0 & W_1 & 0 \\ W_1 & 0 & W_1 \\ 0 & W_1 & 0 \end{pmatrix} .$$

In this way, we introduce association across both space and time. For instance Y_{21} and Y_{41} affect the mean of Y_{12} (as well as affecting Y_{11}) from W_S by itself. Many more possibilities exist. Models formulated through such more general definitions of B are referred to as *spatiotemporal autoregressive* (STAR) models. See Pace et al. (2000) for a full discussion and development. The interpretation of the ρ's in the above example measures the relative importance of first-order spatial neighbors, first order temporal neighbors, and first-order spatiotemporal neighbors.

3.5 Computer tutorials

In this section we outline the use of the S+SpatialStats package in constructing spatial neighborhood (adjacency) matrices, fitting CAR and SAR models using traditional maximum likelihood techniques, and mapping the results for certain classes of problems. Here we confine ourselves to the modeling of Gaussian data on areal units. As in Section 2.5, we adopt a tutorial style.

3.5.1 *Adjacency matrix construction in* S+SpatialStats

The most common specification for a SAR model is obtained by setting $B = \rho W$ and $\tilde{D} = Diag\left(\sigma_i^2\right)$, where W is some sort of spatial dependence matrix, and ρ measures the strength of spatial association. As such, of fundamental importance is the structure of W, which is often taken as an *adjacency* (or *contiguity*) matrix. It is therefore important to begin with the specification of such matrices in S+SpatialStats. Our discussion follows that outlined in Kaluzny et al. (1998, Ch.5), and we refer the reader to that text for further details.

One way of specifying neighborhood structures is through lists in an ordinary text (ASCII) file. For example,

```
1   2  4
2   1  3  5  6
3   2  4  5
4   1  3  6
5   2  3  7
6   2  4
7   5
```

is a typical text listing of adjacencies. Here we have 7 sites, where site 1 has sites 2 and 4 as neighbors, site 2 has sites 1, 3, 5, and 6 as neighbors, and so on. The function read.neighbor in S+SpatialStats reads such a text file and converts it to a spatial.neighbor object, the fundamental adjacency-storage object in the language. Thus, if we write the above matrix to a file called Neighbors.txt, we may create a spatial.neighbor object (say, ngb) as

```
>ngb <- read.neighbor(‘‘Neighbors.txt’’, keep=F)
```

By default, the spatial.neighbor object ngb is larger than required, since the symmetry of the neighbors is not accounted for. To correct this, the size can be reduced using the spatial.condense function:

```
>ngb <- spatial.condense(ngb, symmetry=T)
```

Another, perhaps more direct method of creating neighbor objects is by invoking the spatial.neighbor function directly on an $n \times n$ contiguity matrix. Note that the neighbor relations listed above are equivalent to the (symmetric, 0-1) contiguity matrix

$$\begin{matrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{matrix}$$

Suppose these relations are stored in a file called Adjacency.txt. A spatial.neighbor object may be created from this contiguity matrix as follows:

```
>no.sites <- 7
```

```
>ngb.mat <- matrix(scan(‘‘Adjacency.txt’’),
   ncol=no.sites,byrow=T)
>ngb2 <- spatial.neighbor(neighbor.matrix=ngb.mat,
   nregion=no.sites, symmetric=T)
```

Note that the "symmetry" above refers to the spatial dependence matrix W, and may not always be appropriate. For example, recall that a common specification is to take W as the *row-normalized* adjacency matrix. In such cases, each element is scaled by the sum of the corresponding row, and the resulting W matrix is not symmetric. To form a row-normalized spatial dependence matrix W, we modify the above example to

```
>ngb2 <- spatial.neighbor(neighbor.matrix=ngb.mat,
   nregion=no.sites, weights=1/c(2,4,3,3,3,2,1))
```

Here, the weights are the number of neighbors (i.e., the number of elements in each row of `Neighbors.txt`).

3.5.2 *SAR and CAR model fitting in* S+SpatialStats

We next turn to fitting Gaussian linear spatial models using the `slm` (spatial linear model) function in `S+SpatialStats`. A convenient illustration is offered by the SIDS (sudden infant death syndrome) data, analyzed by Cressie (1993, Sec. 6.2) and Kaluzny et al. (1998, Sec. 5.3), and already loaded into the `S+SpatialStats` package. This data frame contains counts of SIDS deaths from 1974 to 1978 along with related covariate information for the 100 counties in the U.S. state of North Carolina. We fit two spatial autoregressive models with the dependent variable as `sid.ft` (a Freedman-Tukey transformation of the ratio of the number of SIDS cases to the total number of births in each county). Further information about the data frame can be obtained by typing

```
>help(sids)
```

We first fit a null model (no covariates). Note that `sids.neighbor` (built into `S+SpatialStats`) is a `spatial.neighbor` object containing the contiguity structure for the 100 North Carolina counties. Specifically, `region.id` is the variable that identifies the way the regions are numbered, while `weights` specifies the elements of the W matrix. We follow Cressie (1993) and assign the reciprocal of the births in the county as the weights. The resulting model fit (without covariates) is obtained by typing

```
>sids.nullslm.SAR <- slm(sid.ft~1, cov.family=SAR,
   data=sids, spatial.arglist=list(neighbor=sids.neighbor,
   region.id=1:100, weights=1/sids$births))
>null.SAR.summary <- summary(sids.nullslm.SAR)
```

To fit a Gaussian spatial regression model with a regressor (say, the ratio of non-white to total births in each county between 1974 and 1978), we simply modify the above to

```
>sids.raceslm.SAR <- slm(sid.ft~nwbirths.ft,
   cov.family=SAR, data=sids, spatial.arglist
   =list(neighbor=sids.neighbor, region.id=1:100,
   weights=1/sids$births))
>race.SAR.summary <- summary(sids.raceslm.SAR)
```

The output contained in `race.SAR.summary` is as follows:

```
Call:
slm(formula = sid.ft ~ nwbirths.ft, cov.family = SAR,
data = sids, spatial.arglist = list(neighbor = sids.neighbor,
region.id = 1:100, weights = 1/sids\$births))}

Residuals:
   Min     1Q  Median    3Q   Max
-106.9 -18.28   4.692 25.53 79.09

Coefficients:
                Value  Std. Error t value Pr(>|t|)
(Intercept) 1.6729        0.2480  6.7451   0.0000
nwbirths.ft 0.0337        0.0069  4.8998   0.0000

Residual standard error: 34.4053 on 96 degrees of freedom

Variance-Covariance Matrix of Coefficients
              (Intercept)    nwbirths.ft
(Intercept)   0.061513912  -1.633112e-03
nwbirths.ft -0.001633112    4.728317e-05

Correlation of Coefficient Estimates
             (Intercept)  nwbirths.ft
(Intercept ) 1.000000      -0.957582
nwbirths.ft -0.957582       1.000000
```

Note that a county's non-white birth rate does appear to be significantly associated with its SIDS rate, but this covariate is strongly negatively associated with the intercept. We also remark that the `slm` function can also fit a CAR (instead of SAR) model simply by specifying `cov.family=CAR` above.

Next, instead of defining a neighborhood structure completely in terms of spatial adjacency on the map, we may want to construct neighbors using a distance function. For example, given centroids of the various regions, we could identify regions as neighbors if and only if their intercentroidal distance is below a particular threshold.

We illustrate using `www.biostat.umn.edu/~brad/data/Columbus.dat`, a data set offering neighborhood-level information on crime, mean home value, mean income, and other variables for 49 neighborhoods in Columbus, OH, during 1980. More information on these data is available from Anselin (1988, p.189), or in Exercise 10.

We begin by creating the data frame:

```
>columbus <- read.table(‘‘Columbus.dat’’, header=T)
```

Suppose we would like to have regions with intercentroidal distances less than 2.5 units as neighbors. We first form an object, `columbus.coords`, that contain the centroids of the different regions. The function that we use is `find.neighbor`, but a required intermediate step is making a *quad tree*, which is a matrix providing the most efficient ordering for the nearest neighbor search. This is accomplished using the `quad.tree` function in `S+SpatialStats`. The following steps will create a `spatial.neighbor` object in this way:

```
>columbus.coords <- cbind(columbus$X, columbus$Y)
>columbus.quad <- quad.tree(columbus.coords)
>columbus.ngb <- find.neighbor(x=columbus.coords,
   quadtree=columbus.quad, max.dist=2.5)
>columbus.ngb <- spatial.neighbor(row.id=columbus.ngb[,1],
   col.id=columbus.ngb[,2])
```

Once our neighborhood structure is created, we proceed to fit a CAR model (having crime rate as the response and house value and income as covariates) as follows:

```
>columbus.CAR <- slm(CRIME ~ HOVAL + INC, cov.family=CAR,
   data=columbus, spatial.arglist=list(neighbor=
   columbus.ngb, region.id=1:49))
>columbus.CAR.summary <- summary(columbus.CAR)
```

The output from `columbus.CAR.summary`, similar to that given above for the SAR model, reveals both covariates to be significant (both p-values near .002).

3.5.3 Choropleth mapping using the `maps` *library in* `S-plus`

Finally, we describe the drawing of choropleth maps in `S+SpatialStats`. In fact, `S-plus` is all we need here, thanks to the `maps` library originally described by Becker and Wilks (1993). This map library, invoked using the command,

```
>library(maps)
```

contains the geographic boundary files for several maps, including county boundaries for every state in the U.S. However, other important regional boundary types (say, zip codes) and features (rivers, major roads, and railroads) are generally not available. As such, while `S-plus` is not nearly as versatile as `ArcView` or other GIS packages, it does offer a rare combination of GIS and statistical analysis capabilities.

We will now map the actual transformed SIDS rates along with their fitted values under the SAR model of the previous subsection. Before we can do this, however, a special feature of the polygon boundary file of the `S-plus` North Carolina county map must be accounted for. Specifically, county #27, Currituck county, is apparently comprised of not one but three separate regions. Thus, when mapping the raw SIDS rates, Kaluzny et al. (1998) propose the following solution: form a modified vector for the mapping variable (`sid.ft`), but with Currituck county appearing three times:

```
>sids.map <- c(sids$sid.ft[1:26],rep(sids$sid.ft[27],3),
   sids$sid.ft[28:100])
```

We next form a vector of the cutpoints that determine the different bins into which the rates will be classified, and assign the county rates to these bins:

```
>cutoff.sids <- c(0.0,2.0,3.0,3.5,7.0)
>sids.mapgrp <- cut(sids.map, breaks.sids)
```

We now must assign a color (or shade of gray) to each bin. An oddity in the default postscript color specification of `S-plus` is that color "1" is black, and then increasingly lighter shades are given by colors 3, 2, and 4 (not 2, 3, and 4, as you might expect). While this problem may be overcome by careful work with the `ps.options` command, here we simply use the nonintuitive 4-2-3-1 lightest to darkest grayscale ordering, which is obtained here simply by swapping categories 1 and 4:

```
>sids.mapgrp[(sids.mapgrp==1)] <- 0
>sids.mapgrp[(sids.mapgrp==4)] <- 1
>sids.mapgrp[(sids.mapgrp==0)] <- 4
```

Now the map of the actual (transformed) SIDS rates can be obtained as

```
>map("county", "north carolina", fill=T, color=sids.mapgrp)
>map("county", "north carolina", add=T)
>title(main="Actual Transformed SIDS Rates")
>legend(locator(1), legend=
   c("<2.0","2.0-3.0","3.0-3.5",">3.5"), fill=c(4,2,3,1))
```

In the first command, the modified mapping vector `sids.mapgrp` is specified as the grouping variable for the different colors. The `fill=T` option automates the shading of regions, while the next command (with `add=T`) adds the county boundaries. Finally, the `locator(1)` option within the `legend` command waits for the user to click on the position where the legend is desired; Figure 3.5(a) contains the result we obtained. We hasten to add that one can automate the placing of the legend by replacing the

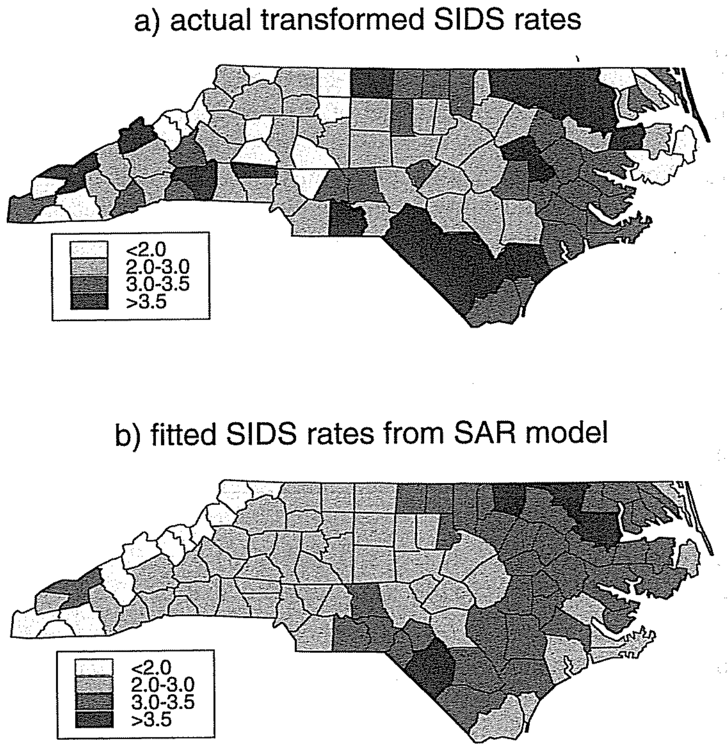
a) actual transformed SIDS rates
<2.0
2.0-3.0
3.0-3.5
>3.5
b) fitted SIDS rates from SAR model
<2.0
2.0-3.0
3.0-3.5
>3.5

Figure 3.5 *Unsmoothed raw (a) and spatially smoothed fitted (b) rates, North Carolina SIDS data.*

`locator(1)` option with actual (x, y) coordinates for the upper left corner of the legend box.

To draw a corresponding map of the fitted values from our SAR model (using our parameter estimates in the mean structure), we must first create a modified vector of the fits (again due to the presence of Currituck county):

```
>sids.race.fit <- fitted(sids.raceslm.SAR)
>sids.race.fit.map <- c(sids.race.fit[1:26],
   rep(sids.race.fit[27],3), sids.race.fit[28:100])
>sids.race.fit.mapgrp <- cut(sids.race.fit.map,
   cutoff.sids)
```

where `cutoff.sids` is the same color cutoff vector as earlier. The map is then drawn as follows:

```
>sids.race.fit.mapgrp <- cut(sids.race.fit.map,
   breaks.sids)
```

```
>sids.race.fit.mapgrp[(sids.race.fit.mapgrp==1)] <- 0
>sids.race.fit.mapgrp[(sids.race.fit.mapgrp==4)] <- 1
>sids.race.fit.mapgrp[(sids.race.fit.mapgrp==0)] <- 4
>map("county", "north carolina", fill=T,
   color=sids.race.fit.mapgrp)
>map("county", "north carolina", add=T)
>title(main="Fitted SIDS Rates from SAR Model")
>legend(locator(1), legend=
   c("<2.0","2.0-3.0","3.0-3.5",">3.5"), fill=c(4,2,3,1))
```

Figure 3.5(b) contains the result. Note that the SAR model has resulted in significant smoothing of the observed rates, and clarified the generally increasing pattern as we move from west to east.

Finally, if a map of predicted (rather than fitted) values is desired, these values can be formed as

```
>noise <- 1/sqrt(sids$births)*resid(sids.raceslm.SAR)
>signal <- sids$sid.ft - sids.race.fit - noise
>sids.race.pred <- signal + sids.race.fit
>sids.race.pred.map <- c(sids.race.pred[1:26],
   rep(sids.race.pred[27],3), sids.race.pred[28:100])
>sids.race.pred.mapgrp <- cut(sids.race.pred.map,
   cutoff.sids)
```

The actual drawing of the maps then proceeds exactly as before.

3.6 Exercises

1. Verify Brook's Lemma, equation (3.7).

2.(a) To appreciate how Brook's Lemma works, suppose Y_1 and Y_2 are both binary variables, and that their joint distribution is defined through conditional logit models. That is,

$$\log \frac{P(Y_1 = 1|Y_2)}{P(Y_1 = 0|Y_2)} = \alpha_0 + \alpha_1 Y_2 \quad \text{and} \quad \log \frac{P(Y_2 = 1|Y_1)}{P(Y_2 = 0|Y_1)} = \beta_0 + \beta_1 Y_1 \, .$$

Obtain the joint distribution of Y_1 and Y_2.

(b) This result can be straightforwardly extended to the case of more than two variables, but the details become increasingly clumsy. Illustrate this issue in the case of *three* binary variables, Y_1, Y_2, and Y_3.

3. Returning to (3.13) and (3.14), let $B = ((b_{ij}))$ be an $n \times n$ matrix with positive elements; that is, $b_{ij} > 0$, $\sum_j b_{ij} \leq 1$ for all i, and $\sum_j b_{ij} < 1$ for at least one i. Let $D = Diag\left(\tau_i^2\right)$ be a diagonal matrix with positive elements τ_i^2 such that $D^{-1}(I - B)$ is symmetric; that is, $b_{ij}/\tau_i^2 = b_{ji}/\tau_j^2$, for all i, j. Show that $D^{-1}(I - B)$ is positive definite.

4. Looking again at (3.13), obtain a simple sufficient condition on B such

that the CAR prior with precision matrix $D^{-1}(I - B)$ is a pairwise difference prior, as in (3.16).

5. Show that $\Sigma_{\mathbf{y}}^{-1} = D_w - \rho W$ is nonsingular (thus resolving the impropriety in (3.15)) if $\rho \in \left(1/\lambda_{(1)}, 1/\lambda_{(n)}\right)$, where $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{-1/2} W D_w^{-1/2}$.

6. Show that if all entries in W are nonnegative and $D_w - \rho W$ is nonsingular with $\rho > 0$, then all entries in $(D_w - \rho W)^{-1}$ are nonnegative.

7. Recalling the SAR formulation using the scaled adjacency matrix $\widetilde{W}$ just below (3.25), prove that $I - \alpha\widetilde{W}$ will be nonsingular if $\alpha \in (-1, 1)$, so that α may be sensibly referred to as an "autocorrelation parameter."

8. In the setting of Subsection 3.3.1, if $(\Sigma_{\mathbf{y}}^{-1})_{ij} = 0$, then show that Y_i and Y_j are conditionally independent given $Y_k, k \neq i, j$.

9. The file `www.biostat.umn.edu/~brad/data/state-sat.dat` gives the 1999 state average SAT data (part of which is mapped in Figure 3.1), while `www.biostat.umn.edu/~brad/data/contig-lower48.dat` gives the contiguity (adjacency) matrix for the lower 48 U.S. states (i.e., excluding Alaska and Hawaii, as well as the District of Columbia).

   (a) Use the `S+SpatialStats` software to construct a `spatial.neighbor` object from the contiguity file.

   (b) Use the `slm` function to fit the SAR model of Section 3.4, taking the verbal SAT score as the response Y and the percent of eligible students taking the exam in each state as the covariate X. Use row-normalized weights based on the contiguity information in `spatial.neighbor` object. Is knowing X helpful in explaining Y?

   (c) Using the `maps` library in `S-plus`, draw choropleth maps similar to Figure 3.1 of both the fitted verbal SAT scores and the spatial residuals from this fit. Is there evidence of spatial correlation in the response Y once the covariate X is accounted for?

   (d) Repeat your SAR model analysis above, again using `slm` but now assuming the CAR model of Section 3.3. Compare your estimates with those from the SAR model and interpret any changes.

   (e) One might imagine that the percentage of eligible students taking the exam should perhaps affect the variance of our model, not just the mean structure. To check this, refit the SAR model replacing your row-normalized weights with weights equal to the reciprocal of the percentage of students taking the SAT. Is this model sensible?

10. Consider the data `www.biostat.umn.edu/~brad/data/Columbus.dat`, taken from Anselin (1988, p. 189). These data record crime information for 49 neighborhoods in Columbus, OH, during 1980. Variables measured include NEIG, the neighborhood id value (1–49); HOVAL, its mean housing value (in \$1,000); INC, its mean household income (in \$1,000); CRIME, its number of residential burglaries and vehicle thefts per thousand households; OPEN, a measure of the neighborhood's open space; PLUMB, the percentage of housing units without plumbing; DISCBD, the neighborhood centroid's distance from the central business district; X, an x-coordinate for the neighborhood centroid (in arbitrary digitizing units, not polygon coordinates); Y, the same as X for the y-coordinate; AREA, the neighborhood's area; and PERIM, the perimeter of the polygon describing the neighborhood.

   (a) Use `S+SpatialStats` to construct `spatial.neighbor` objects for the neighborhoods of Columbus based upon centroid distances less than

      i. 3.0 units,
      ii. 7.0 units,
      iii. 15 units.

   (b) For each of the four spatial neighborhoods constructed above, use the `slm` function to fit SAR models with CRIME as the dependent variable, and HOVAL, INC, OPEN, PLUMB, and DISCBD as the covariates. Compare your results and interpret your parameter estimates in each case.

   (c) Repeat your analysis using Euclidean distances in the B matrix itself. That is, in equation (3.26), set $B = \rho W$ with the W_{ij} the Euclidean distance between location i and location j.

   (d) Repeat part (b) for CAR models. Compare your estimates with those from the SAR model and interpret them.