CBMS Lecture 3

Alan E. Gelfand
Duke University

Basics of areal data models; THE ISSUES

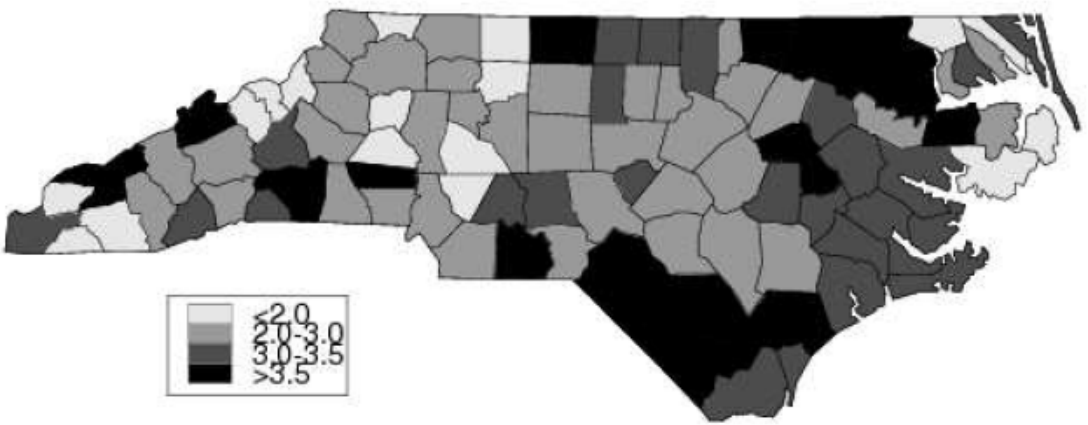
- Here we consider regular grids or lattices and irregular areal units but **NOT** large point referenced datasets
- (i) Is there spatial pattern? If so, how strong is it? Intuitively, "spatial pattern" suggests that measurements for areal units which are near to each other will tend to take more similar values than those for units far from each other
- (ii) Do we want to smooth the data? If so, how much? If the measurement for each unit is a count, even if the counts were independent, and perhaps after population adjustment, there would still be extreme values. Are the observed high counts more elevated than would be expected by chance?
- Under smoothing of counts high values would tend to be pulled down, the low values to be pushed up.
- No smoothing: a display using simply the observed
- Maximal smoothing: a single common value for all units
- *Desired* smoothing: somewhere in between
- How much smoothing is appropriate is not defined.

Issues

- (iii) Inference for new areal units? What data values we expect to be associated with these units?
- If we modify the areal units to new units, e.g., from zip codes to census block groups, what can we say about cancer counts we expect for the latter given those for the former?
- This is the *modifiable areal unit problem (MAUP)*,
- (iv) Descriptive/algorithmic vs. Model-based. We 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 widely available in GIS software packages.

Areal unit data

Actual Transformed SIDS Rates



Proximity matrices

- W, entries w_{ij} (with $w_{ii} = 0$). Choices for w_{ij}:
- $w_{ij} = 1$ if i, j share a common boundary (possibly a common vertex)
- w_{ij} is an *inverse* distance between units
- $w_{ij} = 1$ if distance between units is $\leq K$
- $w_{ij} = 1$ for m nearest neighbors
- W is typically symmetric, but need not be
- $\widetilde{W}$: standardize row i by $w_{i+} = \sum_j w_{ij}$ (so matrix is now row stochastic, but probably no longer symmetric).
- W elements often called "weights"; interpretation
- Could also define first-order neighbors $W^{(1)}$, second-order neighbors $W^{(2)}$, etc.

Measures of spatial association

- Moran's I: essentially an "areal covariogram"

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

- Geary's C: essentially an "areal variogram"

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

- Both are asymptotically normal if Y_i are i.i.d.; Moran has mean $-1/(n-1) \approx 0$, Geary has mean 1
- Better significance testing by comparing to a collection of say 1000 random permutations of the Y_i

Measures of spatial association (cont'd)

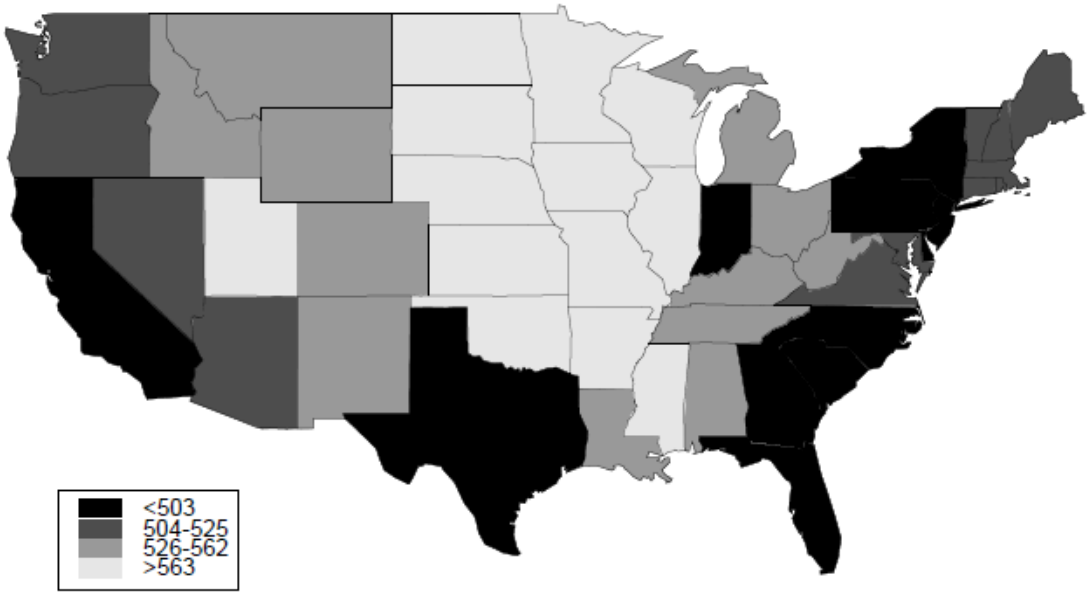
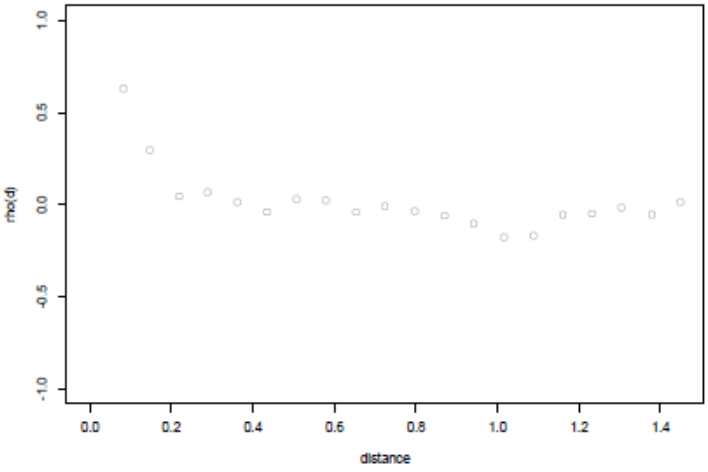


Figure 1: Choropleth map of 1999 average verbal SAT scores, lower 48 U.S. states.

Measures of spatial association (cont'd)

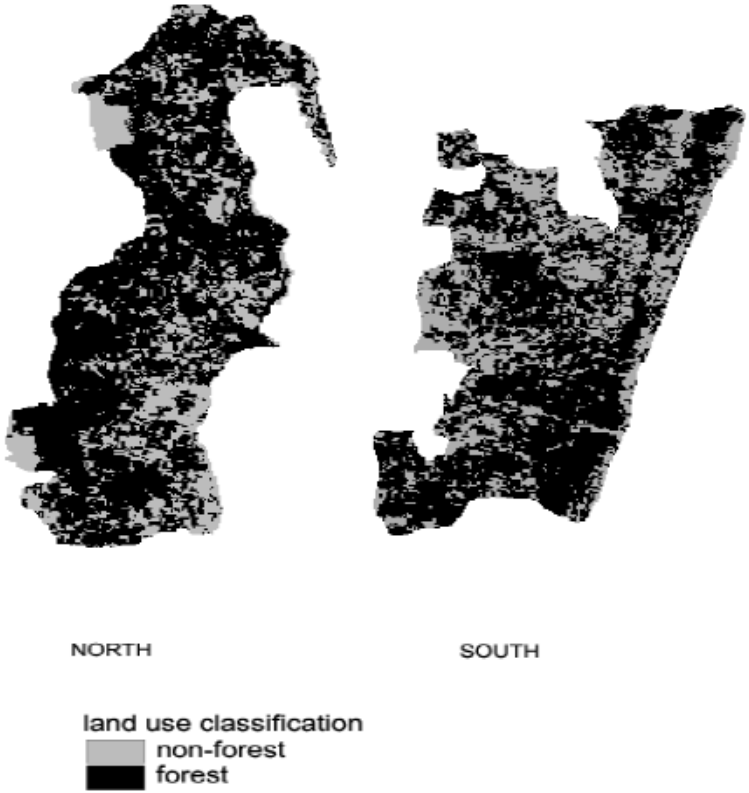
- For these data, we obtain a Moran's I of 0.5833, with associated standard error estimate 0.0920 $\Rightarrow$ very strong evidence against H_0 : no spatial correlation
- We obtain a Geary's C of 0.3775, with associated standard error estimate 0.1008 $\Rightarrow$ again, very strong evidence against H_0 (departure from 1)
- Warning: These data have not been adjusted for covariates, such as the proportion of students who take the exam. (Midwestern colleges have historically relied on the ACT, not the SAT; only the best and brightest students in these states would bother taking the SAT)

Correlogram (via Moran's I)



- Replace w_{ij} with $w_{ij}^{(1)}$ taken from $W^{(1)} \Rightarrow I^{(1)}$
- Replace w_{ij} with $w_{ij}^{(2)}$ taken from $W^{(2)} \Rightarrow I^{(2)}$, etc.
- Plot $I^{(r)}$ vs. r; expect an initial decline across r followed by variation around 0 $\Rightarrow$ spatial pattern!
- spatial analogue of the temporal lag autocorrelation plot

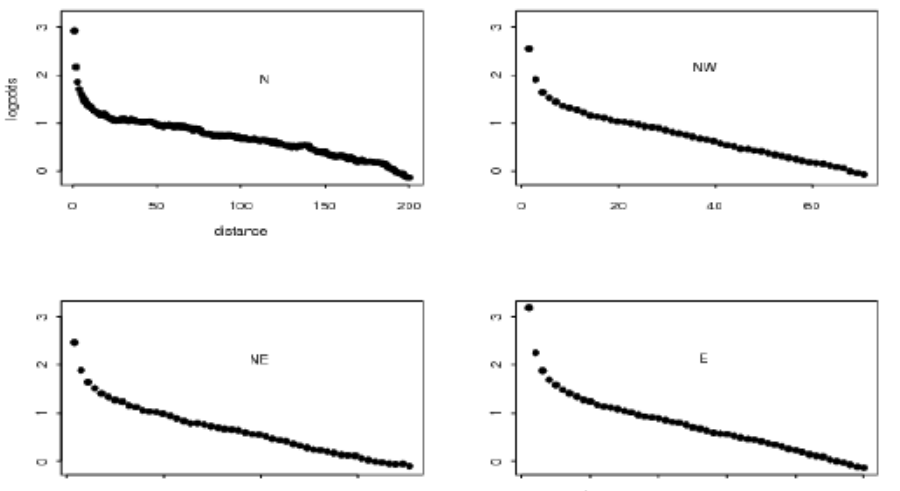
Rasterized binary data map



Binary data correlogram

- With large, regular grids, may seek 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 a specific direction.
- An analogue for the case where the Y_i are binary responses (e.g., presence or absence of forest in the cell) (Agarwal et al.)
- A version of a correlogram for a binary map, using two-way tables and log odds ratios at pixel level

Binary data correlogram



Spatial smoothers

- To smooth Y_i, replace with $\hat{Y}_i = \frac{\sum_i w_{ij} Y_j}{w_{i+}}$
- More generally, we could include the value actually observed for unit i, and revise our smoother to

$$(1 - \alpha) Y_i + \alpha \hat{Y}_i$$

- For $0 < \alpha < 1$, this is a linear (convex) combination in "shrinkage" form. How to choose α?
- Finally, we could try model-based smoothing, i.e., based on $E(Y_i|Data)$, i.e., the mean of the predictive distribution. Smoothers then emerge as byproducts of the hierarchical spatial models we use to explain the Y_i's

Markov random fields

- Consider $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ and the set of densities $\{p(y_i|y_j,\, j \neq i)\}$
- We know $p(y_1, y_2, ...y_n)$ determines $\{p(y_i|y_j, j \neq i)\}$ (the set of full conditional distributions)
- Does $\{p(y_i|y_j, j \neq i)\}$ determine $p(y_1, y_2, ...y_n)$???
- We need the notion of *compatibility*. With two variables, when are $p(y_1|y_2)$ and $p(y_2|y_1)$ compatible? Not always, e.g., $p(y_1|y_2) = N(a + by_2, \sigma_1^2)$ and $p(y_2|y_1) = N(c + dy_1^3, \sigma_2^2)$

Brook's Lemma

- If the full conditionals are compatible, then Brook's Lemma provides a way to *construct* the joint distribution from the full conditionals
- We can write the joint distribution as

$$p(y_1, \ldots, y_n) = \frac{p(y_1, y_2, \ldots, y_n)}{p(y_{10}, y_2, \ldots, y_n)} \frac{p(y_2, y_{10}, y_3, \ldots, y_n)}{p(y_{20}, y_{10}, y_3, \ldots, y_n)} \cdots \frac{p(y_n, y_{10}, \ldots, y_{n-1,0})}{p(y_{n0}, y_{10}, \ldots, y_{n-1,0})} p(y_{10}, \ldots, y_{n0})$$

- Replacing each joint distributions with conditional $\times$ marginal, the marginal terms cancel and we have

$$p(y_1, \ldots, y_n) = \frac{p(y_1 | y_2, \ldots, y_n)}{p(y_{10} | y_2, \ldots, y_n)} \frac{p(y_2 | y_{10}, y_3, \ldots, y_n)}{p(y_{20} | y_{10}, y_3, \ldots, y_n)} \cdots \frac{p(y_n | y_{10}, \ldots, y_{n-1,0})}{p(y_{n0} | y_{10}, \ldots, y_{n-1,0})} p(y_{10}, \ldots, y_{n0})$$

Brook's Lemma cont.

- We have the joint distribution on the left side in terms of the full conditional distributions on the right side
- And, if left side is proper, since it integrates to 1, the normalizing constant is determined by integrating the right side and then rescaling to 1
- We have a constructive way to retrieve the joint distribution from the full conditional distributions
- Useful in many other problems

"Local" modeling

- Suppose we specify the full conditionals such that

$$p(y_i|y_j, j \neq i) = p(y_i|y_j \in \partial_i) \ ,$$

  where ∂_i is the set of neighbors of cell (region) i.
  When does $\{p(y_i|y_j \in \partial_i)\}$ determine $p(y_1, y_2, ...y_n)$?
- Def'n: a clique is a set of cells such that each element is a neighbor of every other element
- Def'n: a potential function of order k is a positive function of k arguments that is exchangeable in these arguments. A potential of order 2 is $Q(y_i, y_j)$ with $Q(y_i, y_j) = Q(y_j, y_i)$
- Def'n: $p(y_1, \ldots, y_n)$ is a Gibbs distribution if, as a function of the y_i, it is a product of potentials on cliques. With potentials of order 2, $p(y_1, \ldots, y_n) = \Pi_{i<j} Q(y_i, y_j)$

"local" modeling, cont.

- For a continuous variable, with $k = 2$, we might take $Q(y_i, y_j) = exp(-w_{i,j}(y_i - y_j))^2$
- For binary data, $k = 2$, we might take $Q(y_i, y_j) = I(y_i = y_j) = y_i y_j + (1 - y_i)(1 - y_j)$
- Cliques of size 1 $\Leftrightarrow$ independence
- Cliques of size 2 with above Q for continuous variables and $w_{i,j} = I(i \sim j) \Leftrightarrow$ pairwise difference form

$$p(y_1, y_2, ...y_n) \propto \exp\left\{ -\frac{1}{2\tau^2} \sum_{i,j} (y_i - y_j)^2 I(i \sim j) \right\}$$

  and therefore $p(y_i | y_j, j \neq i) = N(\sum_{j \in \partial_i} y_j / m_i, \tau^2 / m_i)$, where m_i is the number of neighbors of i
- No interest in $k > 2$.

Two primary results

- Hammersley-Clifford Theorem: If we have a Markov Random Field (i.e., $\{p(y_i|y_j \in \partial_i)\}$ uniquely determine $p(y_1, y_2, ...y_n)$), then the latter is a Gibbs distribution
- Geman and Geman result : If we have a joint Gibbs distribution, i.e., as defined above, then we have a Markov Random Field

Conditional autoregressive (CAR) model

- Gaussian (autonormal) case

$$p(y_i|y_j, j \neq i) = N\left(\sum_j b_{ij} y_j, \tau_i^2\right)$$

Using Brook's Lemma we can obtain

$$p(y_1, y_2, \ldots y_n) \propto \exp\left\{-\frac{1}{2}\mathbf{y}' D^{-1}(I - B)\mathbf{y}\right\}$$

where $B = \{b_{ij}\}$ and D is diagonal with $D_{ii} = \tau_i^2$.

- $\Rightarrow$ suggests a multivariate normal distribution with $\mu_Y = 0$ and $\Sigma_Y = (I - B)^{-1} D$
- $D^{-1}(I - B)$ symmetric requires $\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}$ for all i, j

CAR Model (cont'd)

- Returning to W, let $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$, so

$$p(y_1, y_2, \ldots y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - W)\mathbf{y}\right\}.$$

  D_w diagonal with $(D_w)_{ii} = w_{i+}$ and with algebra,

$$p(y_1, y_2, \ldots y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i\neq j} w_{ij}(y_i - y_j)^2\right\}$$

  Intrinsic autoregressive (IAR) model!

- Improper since $(D_w - W)\mathbf{1} = 0$, so requires a constraint – say, $\sum_i y_i = 0$
- So, not a data model, a random effects model!
- τ^2 represents both dispersion and spatial dependence

CAR Model Issues

- With τ^2 unknown, what to do with power of τ^2 in joint distribution? ($n-$ # of "islands")
- "Proper version:" replace $D_w - W$ by $D_w - \rho W$, and choose ρ so that $\Sigma_{\mathbf{y}} = (D_w - \rho W)^{-1}$ exists!
  This in turn implies $Y_i | Y_{j \neq i} \sim N(\rho \sum_j w_{ij} Y_j, \tau^2 / m_i)$

"To ρ or not to ρ?"

- **Advantages**:
- makes distribution proper
- adds parametric flexibility
- $\rho = 0$ interpretable as independence

CAR Models with ρ parameter

Disadvantages:

- why should we expect Y_i to be a proportion of average of neighbors – a sensible spatial interpretation?
- calibration of ρ as a correlation, e.g.,

$$\rho = 0.80 \text{ yields } 0.1 \leq \text{ Moran's } I \leq 0.15,$$
$$\rho = 0.90 \text{ yields } 0.2 \leq \text{ Moran's } I \leq 0.25,$$
$$\rho = 0.99 \text{ yields } \text{ Moran's } I \leq 0.5$$

- So, used with random effects, scope of spatial pattern may be limited

More CAR modeling

- Again, CAR is an improper prior for modeling random effects with inverse covariance matrix $\frac{1}{\tau^2}Q$ where $Q = (D_W - W)$
- To make a data model, $Y_i = \mu_i + \phi_i + \epsilon_i$ where ϕ_i from a CAR and ϵ_i are i.i.d. $N(0, \gamma^2)$, i.e., $Y_i|\mu_i, \phi_i \sim N(\mu_i + \phi_i, \gamma^2)$
- Two variance components - τ^2 captures structured spatial variation, γ^2 captures unstructured variation
- If Q were full rank $\Sigma_{\boldsymbol{\phi}+\boldsymbol{\epsilon}} = \tau^2 Q^{-1} + \gamma^2 I_n$. With reparametrization becomes $\sigma^2(\lambda Q^{-1} + (1-\lambda)I_n)$

cont.

- But Q not full rank so instead write (Leroux et al.) $\Sigma^{-1}_{\boldsymbol{\phi}+\boldsymbol{\epsilon}} = \frac{1}{\sigma^2}(\lambda Q + (1-\lambda)I_n)$
- Now $\Sigma^{-1}_{\boldsymbol{\phi}+\boldsymbol{\epsilon}}$ full rank
- $\Sigma_{\boldsymbol{\phi}+\boldsymbol{\epsilon}} = \sigma^2(\lambda Q + (1-\lambda)I_n)^{-1}$
- Corresponds to *marginalization*. As a result, we have random effects η_i with $\boldsymbol{\eta} \sim N(0, \Sigma_{\boldsymbol{\phi}+\boldsymbol{\epsilon}})$
- Now $E(\eta_i|\eta_j, j \neq i) = \frac{\lambda}{1-\lambda+\lambda m_i} \sum_{j\sim i} \eta_j$
- Now $\text{Var}(\eta_i|\eta_j, j \neq i) = \frac{\sigma^2}{1-\lambda+\lambda m_i}$
- A reparametrization to $\boldsymbol{\phi} + \boldsymbol{\epsilon}$
- $\lambda = 1$ is IAR model, $\lambda = 0$ is independence model

cont.

- Another variation allows spatially varying, directional, adaptive weights $\tilde{w}_{ij} = w_{ij} e^{Z_i}$ where $Z_i = Z(s_i^*)$ with s_i^* the centroid of areal unit i and $Z(\cdot)$ a mean 0 Gaussian process (Berrocal et al.)
- A conditional CAR model, given $\{Z_i\}$
- More general proximities
- Model weights based on the forward difference analogue of penalizing the derivatives of a surface under a thin plate spline.
- Consider twelve neighbors of a given point.
- North, east, south, and west neighbors each weight $+8$, the northeast, southeast, southwest, and northwest neighbors, each weight -2 and the "two away" north, east, south, and west neighbors, each weight -1. Thus, $w_{i+} = 20$
- Unusual for spatial smoothing but a probabilistic justification through the two-dimensional random walk on the lattice.

Comments

- The CAR specifies Σ_Y^{-1}, **NOT** Σ_Y (as in the point-level modeling case), so does not directly model association
- $(\Sigma_Y^{-1})_{ii} = 1/\tau_i^2$; $(\Sigma_Y^{-1})_{ij} = 0 \Leftrightarrow$ conditional independence
- Prediction at new sites is ad hoc; if

$$p(y_0|y_1, y_2, \ldots y_n) = N(\sum_j w_{0j} y_j / w_{0+}, \tau^2 / w_{0+})$$

then $p(y_0, y_1, \ldots y_n)$ well-defined but not CAR

NonGaussian versions

- To model the data directly using a CAR specification in many cases a normal distribution would not be appropriate.
- Binary response, sparse counts, categorical data are examples.
- Here, focus on case where the Y_i are binary variables, the so-called autologistic CAR model
- Ignoring covariates, consider the joint distribution $p(y_1, y_2, ..., y_n; \psi)$

$$\propto \exp(\psi \sum_{i,j} w_{ij} 1(y_i = y_j)) = \exp(\psi \sum_{i,j} w_{ij}(y_i y_j + (1-y_i)(1-y_j))).$$

- A Gibbs distribution with a potential on cliques of order $k = 2$.
- Always proper since it can take on only 2^n values. However, ψ is an unknown parameter and need to calculate the normalizing constant $c(\psi)$ in order to infer about ψ
- Computation of this constant requires summation over all of the 2^n possible values that $(Y_1, Y_2, ..., Y_n)$ can take on.

cont.

- We can obtain the full conditional distributions for the Y_i's. In fact, $P(Y_i = 1|y_j, j \neq i) = e^{\psi S_{i,1}}/(e^{\psi S_{i,0}} + e^{\psi(S_{i,1})})$ where $S_{i,1} = \sum_{j\sim i} 1(y_j = 1)$ and $S_{i,0} = \sum_{j\sim i} 1(y_j = 0)$ and $P(Y_i = 0|y_j, j \neq i) = 1 - P(Y_i = 1|y_j, j \neq i)$.
- $S_{i,1}$ is the number of neighbors of i equal to 1 and $S_{i,0}$ is the number of neighbors of i equal to 0; larger values of ψ place more weight on matching.
- Since the full conditional distributions take on only two values, there are no normalizing issues
- Bringing in covariates is natural on the log scale, i.e.,

$$\log \frac{P(Y_i = 1|y_j, j \neq i)}{P(Y_i = 0|y_j, j \neq i)} = \psi(S_{i,1} - S_{i,0}) + \mathbf{X}_i^T \boldsymbol{\beta}.$$

Potts model

- The case where Y_i can take on one of several categorical values is a natural extension
- If we label the (say) L possible outcomes as simply $1, 2, ..., L$, then we can define the joint distribution for $(Y_1, Y_2, ..., Y_n)$ as above, i.e.

$$p(y_1, y_2, ..., y_n; \psi) \propto \exp(\psi \sum_{i,j} w_{ij} 1(y_i = y_j))$$

  with w_{ij} as above.
- This distribution is referred to as a *Potts model*
- Now the distribution takes on L^n values; now, calculation of the normalizing constant is even more difficult.

Simultaneous autoregressive models (SAR)

- We can write the system of CAR model conditional distributions as $\mathbf{Y} = B\mathbf{Y} + \boldsymbol{\epsilon}$ or equivalently as $(I - B)\mathbf{Y} = \boldsymbol{\epsilon}$.
- The distribution for $\mathbf{Y}$ induces a distribution for $\boldsymbol{\epsilon}$. If $[\mathbf{Y}]$ is proper, $\mathbf{Y} \sim N(\mathbf{0}, (I - B)^{-1}D)$ so $\boldsymbol{\epsilon} \sim N(\mathbf{0}, D(I - B)^T)$ Suppose we reverse this, specify a (normal) distribution for $\boldsymbol{\epsilon}$ to induce a distribution for $\mathbf{Y}$. This is the SAR model
- Imitating usual autoregressive time series modeling, suppose we take the ϵ_i to be independent innovations.
- For added generality, assume that $\boldsymbol{\epsilon} \sim N\left(0, \tilde{D}\right)$ where $\tilde{D}$ is diagonal with $\left(\tilde{D}\right)_{ii} = \sigma_i^2$.
- Now $Y_i = \sum_j b_{ij} Y_j + \epsilon_i,\ i = 1, 2, ..., n$, with $\epsilon_i \sim N\left(0, \sigma_i^2\right)$
- Equivalently, $(I - B)\mathbf{Y} = \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon}$ distributed as above.
- If $(I - B)$ is full rank,

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

cont.

- 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
- If $\mathbf{U} = B\mathbf{U} + \boldsymbol{\epsilon}$, we obtain

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

- Nice interpretation: a spatial weighting of neighbors and a component that is a usual linear regression.
- SAR models are frequently employed in the spatial econometrics literature.
- The SAR model does not introduce any spatial effects; the errors are independent.
- Important point: SAR models are well suited to maximum likelihood estimation but not for MCMC fitting of Bayesian models

CAR versus SAR models

- Under propriety, the two specifications are equivalent if and only if

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

  where we use the tilde to indicate matrices in the SAR model.
- So, any SAR model can be represented as a CAR model (since D is diagonal, we can straightforwardly solve for B)
- The converse is not true (Cressie).