E. Spatial smoothing using MLM

→ Chapter 13

- RWC consider this as a continuous-by-continuous interaction, or as bivariate smoothing. We'll consider their bivariate penalized spline approach in the next lecture(s).

- This lecture will discuss spatial analysis like you learned in your spatial class.

- The point: Yet another class of models, developed independently, that can be expressed as a MLM at the price of leaving out some (not much?) of the older theory.

<u>WARNING</u>: I am <u>not</u> an expert in the area called "spatial statistics". I know a ~~fair~~ amount about so-called intrinsic CAR models.

IE 1/1   2/11/08

• Spatial stat is about measurements taken at <u>locations</u> and is distinctive in its concern about <u>spatial correlation</u>.

• Two common kinds of locations and data

1. <u>Geostatistical data</u>

Location = x and y coordinates, interpreted literally as distances from an origin along axes

- interpolation between observed points is meaningful and often the main goal
- RWC's 2D splines can be used; I'll briefly show a more traditional approach. (RWC Chapter 13.3 gives a different view.)

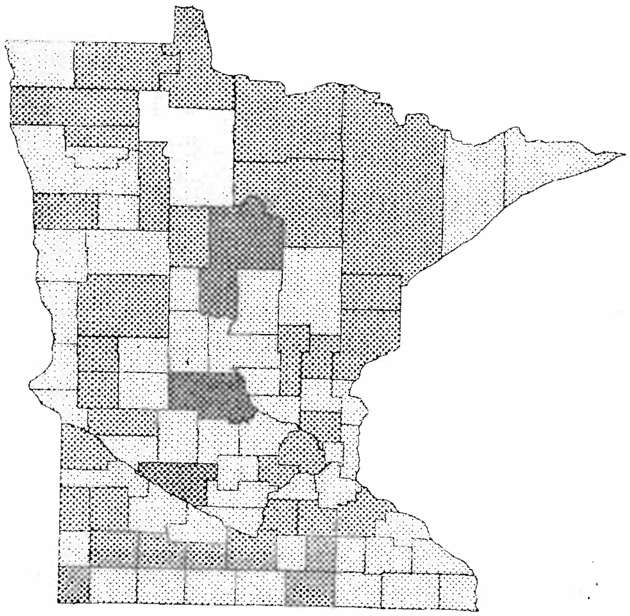
2. <u>Areal data</u>

Location = a specific region on a map, with the measured variable being a total or average over the region.

Distance is not used explicitly; spatial "nearness" is specified by defining <u>pairs of neighbors</u> from among the regions on the map.

IE1/2 2/11/08 | Purpose: Smoothing & "borrowing strength"

Areal data continued: Two examples of regions (areas)

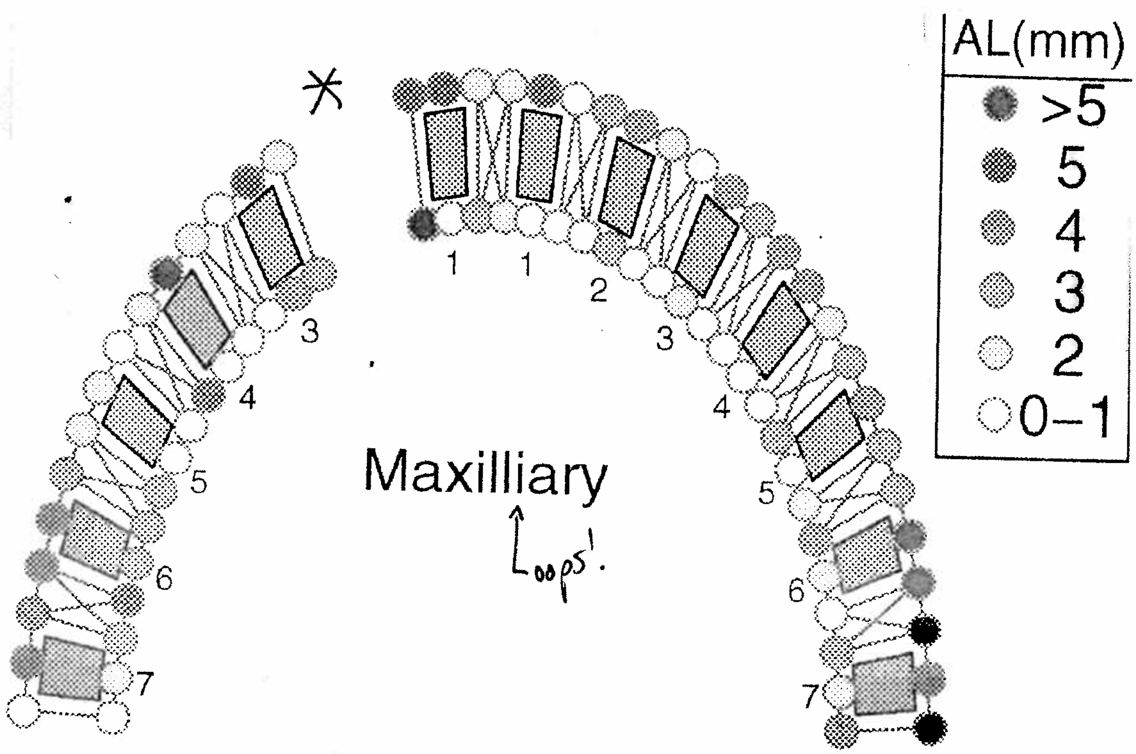


Common definition:

2 counties are a neighbor pair if they share a boundary

Periodontal data:

2 measurement sites "O" are a neighbor pair if they are touching or are joined by a line.



This "map" has 2 "islands" because a tooth is missing at "*"

IE 1/3 2/11/08

Fitting these models into MLM: Geostat

MLM $\underset{n\times 1}{y} = X\beta + \underset{n\times q}{Z}\,\underset{q\times 1}{u} + \underset{n\times 1}{\varepsilon}$

① $\text{cov}(\varepsilon) = R$ captures spatial correlation

e.g., $R = [R_{ij}]$ $\quad R_{ij} = \sigma_\varepsilon^2 I_n + \sigma_c^2 \exp(-d_{ij}/\theta)$

"nugget"

d_{ij} = distance between observations i, j

② $\text{cov}(\varepsilon) = \sigma_\varepsilon^2 I_n \quad Z = I_n \leftarrow q = n$

$\text{cov}(u)$ has a spatial structure

e.g. $G = [G_{ij}] \quad G_{ij} = \sigma_u^2 \exp(-d_{ij}/\theta)$

gives y the same marginal covariance as approach ①

Fitting areal models into MLM

Simplest case: $y = \delta + \varepsilon \qquad \varepsilon \sim N_n(0, \sigma_\varepsilon^2 I_n)$

δ has a CAR model/distribution/prior
(Note: other models exist, e.g. SAR)

The CAR model can be specified several ways

Pair wise differences (Besag 1974 [?])

$f(\underset{\sim}{\delta} \mid \tau^2) \propto \exp\left(\frac{-1}{2\tau^2} \sum_{i \sim j} (\delta_i - \delta_j)^2\right)$ "$i \sim j$" = "i & j are neighbors"

It is easy to show $f(\delta \mid \tau^2) \propto (\tau^2)^{-\frac{(n-G)}{2}} \exp\left(\frac{-1}{2\tau^2} \delta' Q \delta\right)$

G = # of islands in the spatial map

oops! $Q = (q_{ij})$: q_{ii} = region i's number of neighbors

$q_{ij} = \begin{cases} -1 & \text{if } i \sim j \\ 0 & \text{otherwise.} \end{cases}$

- Q has G zero eigenvalues, one for each island.
- $\forall\, Q$, 1_n is an eigenvector with zero eigenvalue.

IE 2/5 2/11/08

Conditional mean and variance

$$E(\delta_i \mid \delta_{(-i)}) = \frac{1}{m_i} \sum_{i \sim j} \delta_j \qquad m_i = \begin{cases} \text{\# of neighbors} \\ \text{of region } i \end{cases}$$

$$\mathrm{var}(\delta_i \mid \delta_{(-i)}) = \tau^2 / m_i$$

This implies the same (singular) joint density of δ.

A more general specification for the CAR

$$\delta \sim N(0, \tau^2 (D - \rho C)^{-1})$$

- D is diagonal
- C has $c_{ii} = 0$, $c_{ij} = c_{ji} \geq 0$; $d_{ii} = \sum_{j=1}^{n} c_{ij}$
- $\rho \in [0,1]$ describes spatial dependence (vaguely!)
- Reduces to our ICAR if $\rho = 1$ and

$$c_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

IE 1/6 2/11/08

• How to fit this into the MLM?

If $y = \delta + \varepsilon$, $\delta \sim CAR(\tau^2, Q)$

oops!

We can't write $\delta \sim N_n(0, G^*)$ because the CAR (ICAR, actually) specifies a singular distribution: $G^{*-1} = \frac{1}{\tau^2} Q$ has zero eigenvalues.

• A little cleverness is needed

spectral decomposition

$Q = VDV'$ for V orthogonal, D diagonal

$$D = \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & 0_{G\times G} \end{array}\right]$$

D_1 is $(N-G)\times(N-G)$ diagonal, with $D_1 = diag(d_1, \ldots, d_{N-G})$ and $d_j > 0$

$V = [V_1 | V_2]$ V_1 has $N-G$ columns, V_2 has G

• So $\delta \sim CAR(\tau^2, Q)$ implies

$u = V_1'\delta \sim N_{N-G}(0, \tau^2 D_1^{-1})$, $\beta = V_2'\delta$ has a flat prior, i.e., it's a fixed effect.

$col(V_2)$ includes 1_n; more generally $col(V_2)$ fits the island means

$\delta \sim CAR(\tau^2, Q) \quad Q = VDV', \quad V = \begin{array}{c} \left[V_1 \mid V_2 \right] \\ {\scriptstyle N-G \quad G} \end{array}$

$u = V_1' \delta \sim N_{N-G}(0, \tau^2 D_1^{-1}) \qquad D = \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & 0_{G\times G} \end{array} \right]$

$\beta = V_2' \delta$ is fixed effects (G of them)

So
$$
\begin{aligned}
y &= \delta + \varepsilon \\
&= VV'\delta + \varepsilon \\
&= V\left(\frac{V_1'\delta}{V_2'\delta}\right) + \varepsilon \\
&= V\left(\frac{u}{\beta}\right) + \varepsilon \\
&= V_2 \underset{G\times 1}{\beta} + V_1 \underset{(N-G)\times 1}{u} + \varepsilon
\end{aligned}
$$

$\mathrm{cov}(\varepsilon) = R = \sigma_\varepsilon^2 I_n$

$\mathrm{cov}(u) = \cancel{G} = \tau^2 D_1^{-1}$

Loops!

- V_2 is the main effect for island means
  - if $G = 1$ island, $V_2 = 1_n$ is an intercept.

IE 1/8 2/11/08

- This is easily extended to include fixed effects besides the island means.

- If X's columns are the other fixed effects (<u>without</u> an intercept), then the MLM expression is

$$y = V_2 \overset{\text{intercepts}}{\beta^{(I)}} + X \overset{\text{predictors}}{\beta^{(P)}} + V_1 u + \varepsilon$$

oops!

$$\operatorname{cov}(u) = G^* = \tau^2 D_1^{-1}$$
$$\operatorname{cov}(\varepsilon) = R = \sigma_\varepsilon^2 I_n$$

What are the columns of V_1?

How do we interpret them?

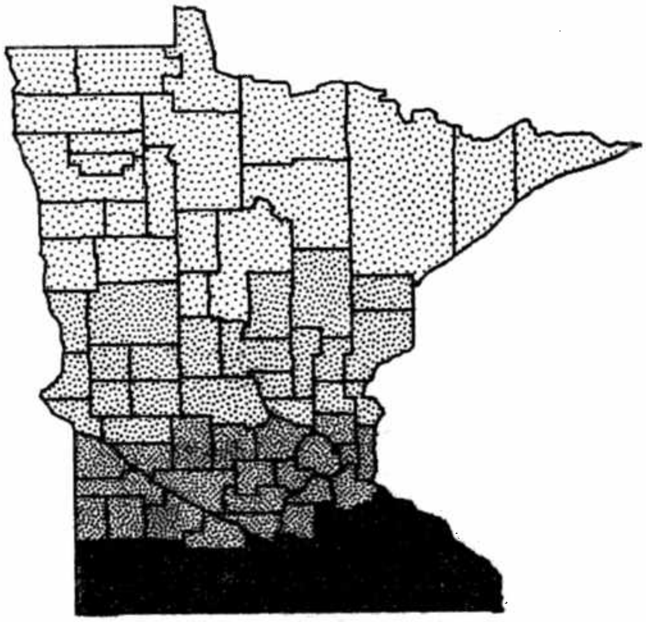
We have some intuition based on examples.

IE 1/9 2/11/08

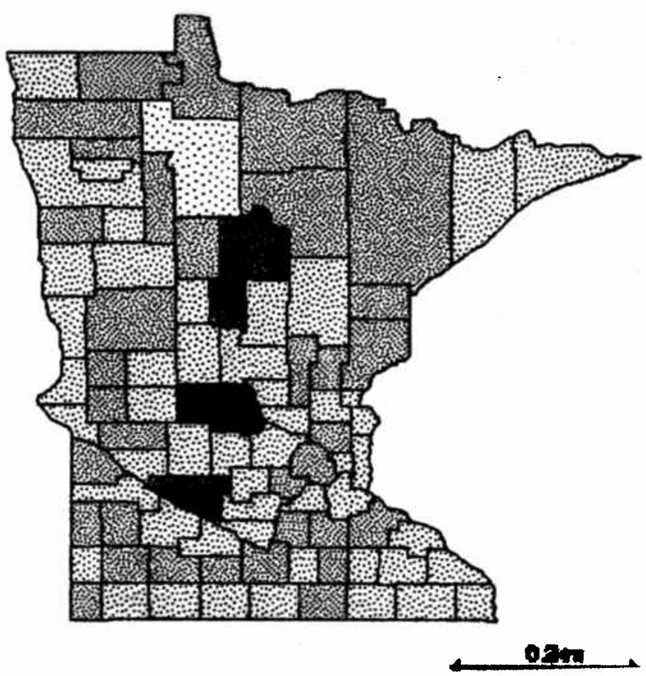
Counties of Minnesota

Recall $Q = [V_1 | V_2] \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix}$

Prior precision is $\frac{1}{\tau^2} Q$



- Column of V_1 corresponding to smallest $d_i > 0$ i.e., weakest prior constraint



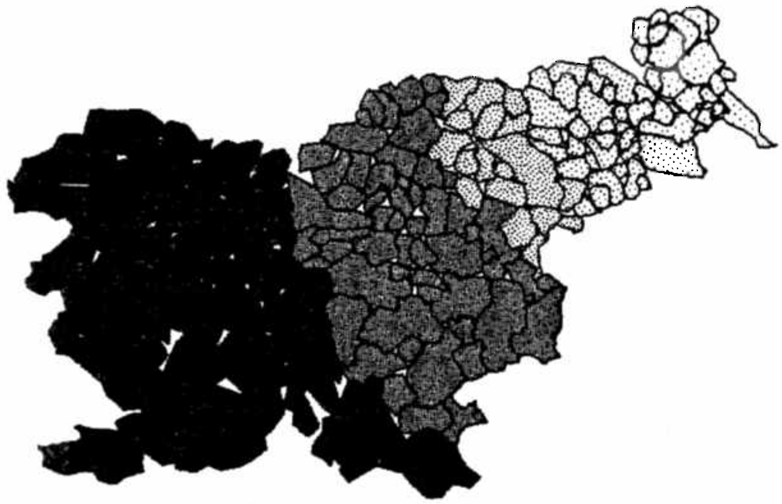
Column of V_1 corresponding to largest $d_i > 0$, i.e., strongest prior constraint.

IE I/10 2/13/08

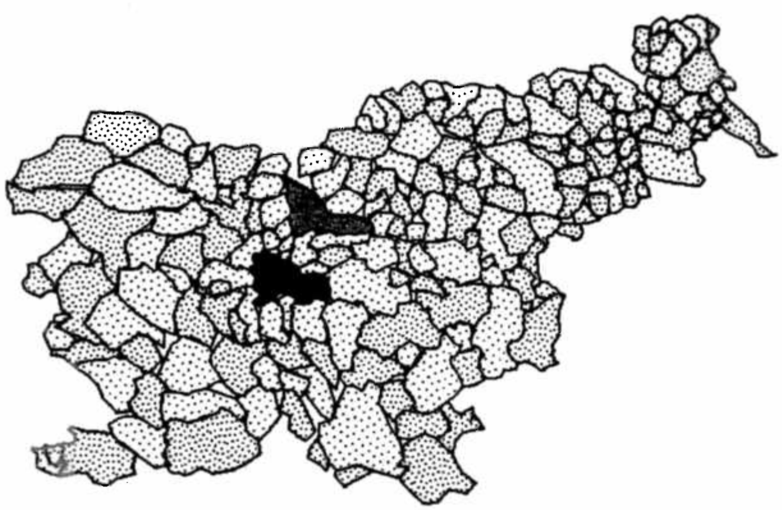
Municipalities of Slovenia

$V_1 u$ is RE

$\text{cov}(u) = \tau^2 D_1^{-1}$



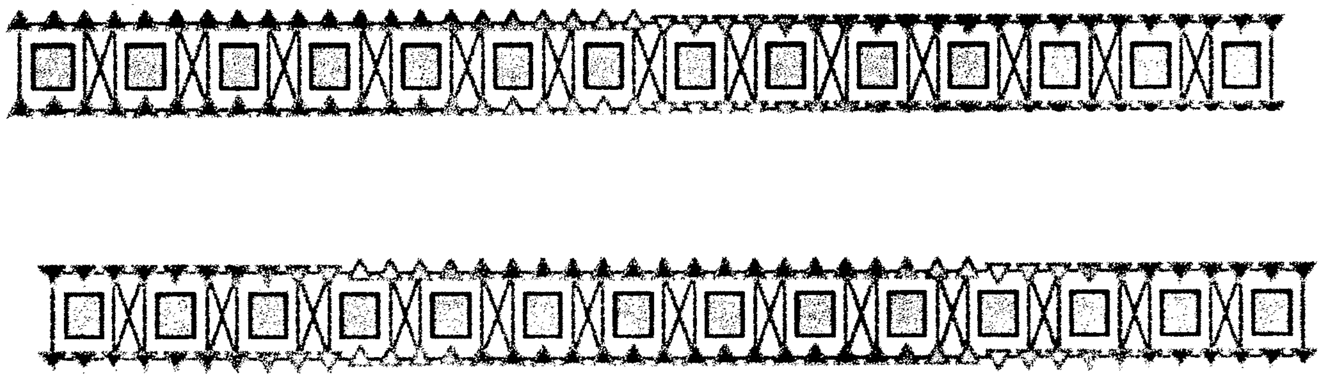
Column of V_1 corresponding to smallest $d_i > 0$, i.e., weakest prior constraint



Column of V_1 corresponding to largest $d_i > 0$, i.e., strongest prior constraint.

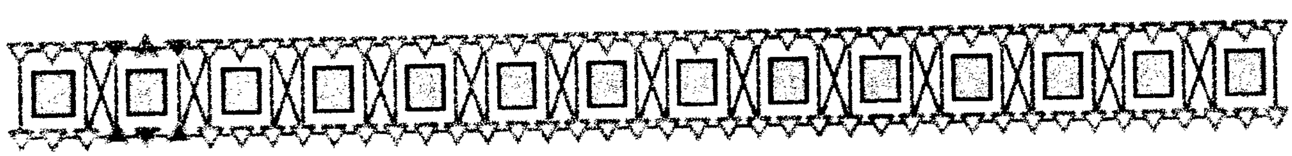
Periodontal Data

Smallest $d_i > 0$ (weakest prior constraint)



middling $d_i > 0$ (moderate prior constraint)

largest $d_i > 0$ (strongest prior constraint)



IE 1/12 2/13/08

Tentative generalization:

- low-frequency (large-scale) contrasts are smoothed least

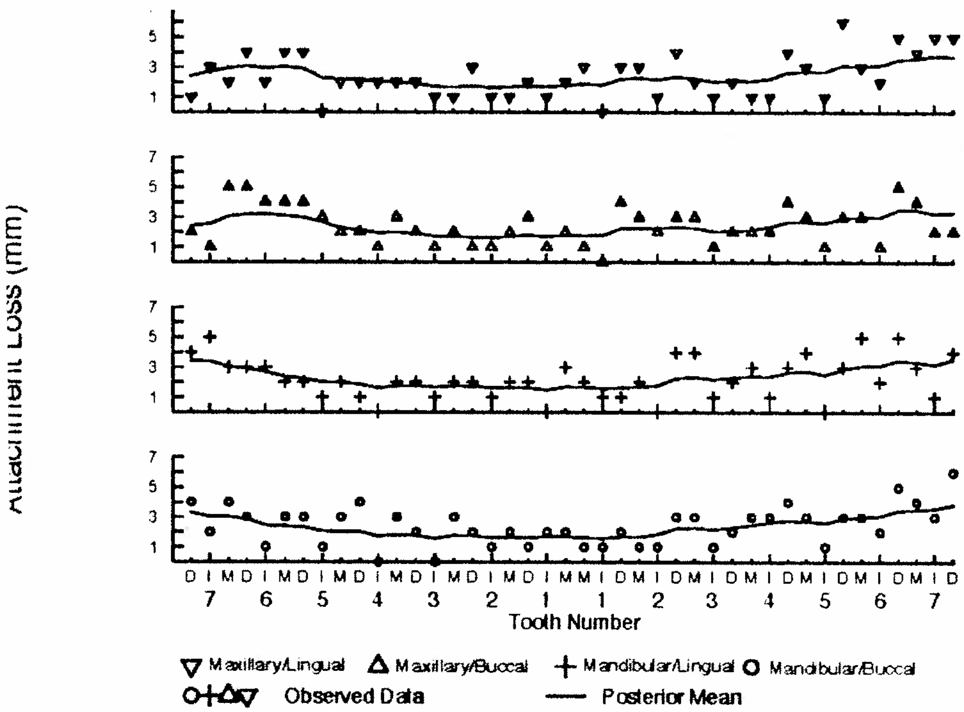
- high-frequency (small-scale) contrasts are smoothed most, i.e. are largely pushed into error.

Thesis topic: Demonstrate this in some generality.

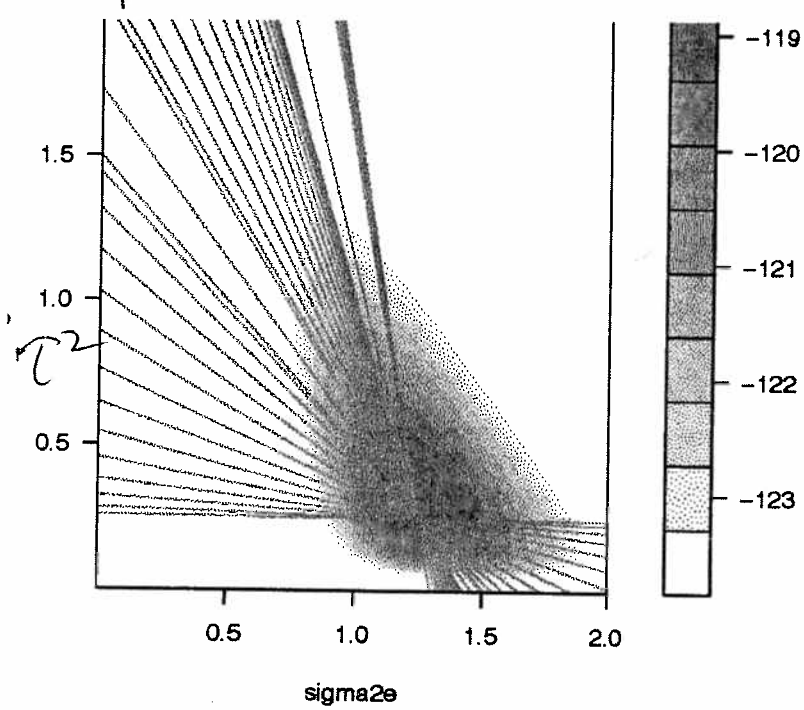
IE 1/13 2/13/08

Example: Attachment loss data (Reich & Hodges 200

Data and simplest CAR fit



Marginal posterior of σ_ε^2 and σ_u^2



The lines correspond to distinct positive eigenvalues of $Q(d_i)$ — more later!

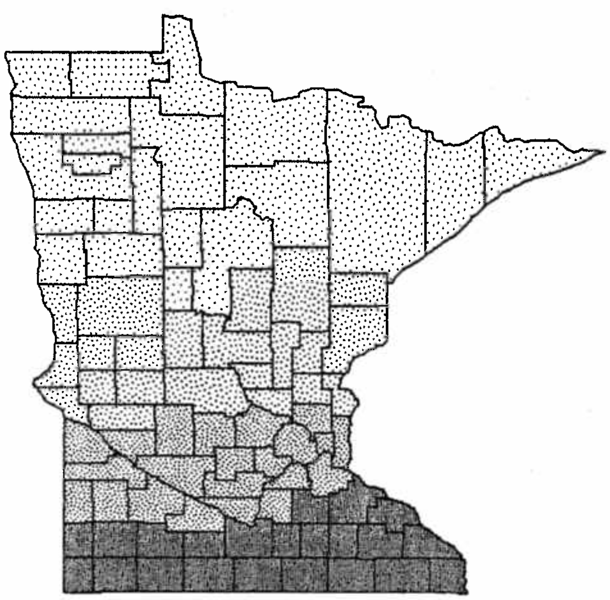
IE 1/14 2/13/08

CAR Models/Priors are modular also.

Example: Using CAR in SANOVA to smooth a spatial effect (Zhang, Hodges, Banerjee '07)

3-cancer dataset: For each of MN's 87 counties, we have incidence of 3 cancers: lung, larynx, esphagus

- This can be viewed as a 3 X 87 ANOVA, in which we smooth counties using a CAR model.



Note: only one island, so only one zero eigenvalue in Q, corresponding to $V_2 = \frac{1}{\sqrt{87}} 1_{87}$

To fit this into SANOVA, we need:

- The same re-parametrization trick for the CAR model
  - Here $V_2 = V^{(-)}$
- a 3×2 matrix H_{CA} with orthonormal columns, describing 2 contrasts in the cancers.

In the MLM Framework $y = X\beta + Zu + \varepsilon$:

Fixed effects:

Grand mean & Cancer Main effect

$$X = \left[\frac{1}{\sqrt{3 \times 87}} 1_{3 \times 87} \;\middle|\; \frac{1}{\sqrt{87}} 1_{87} \otimes H_{CA} \right]$$

where $A \otimes B = [a_{ij} B]$

Random effects

	county main effect	county-by-cancer interaction: 1st column	county-by-cancer interaction: 2nd column
$Z =$	$V^{(-)} \otimes \frac{1}{\sqrt{3}} 1_3$	$V^{(-)} \otimes H_{CA}^{(1)}$	$V^{(-)} \otimes H_{CA}^{(2)}$
	u_0 86x1	u_1 86x1	u_2 86x1
$\mathrm{cov}(u_j)$	$\tau_0^2 D_1^{-1}$	$\tau_1^2 D_1^{-1}$	$\tau_2^2 D_1^{-1}$

IE 1/17 2/13/08

- This is a restriction of the multivariate CAR (MCAR) model

- MCAR is specified by the same Q describing spatial neighbor pairs

  AND $\Sigma_{3\times 3}$ = covariance of cancers in a county

- If $\Sigma = \Phi \Delta \Phi'$ is the spectral decomposition of Σ

  the SANOVA presumes $\Phi = \left[\frac{1}{\sqrt{3}} 1_3 \,\middle|\, H_{CA}\right]$ is known.

- Z, H, & B shows that this is competitive with MCAR even when $\left[\frac{1}{\sqrt{3}} 1_3 \,\middle|\, H_{CA}\right]$ is "wrong"

JE 1/18 2/13/08