Hierarchical Modeling for Large Univariate Areal Data

Sudipto Banerjee

September 03–05, 2017

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles

Areal data

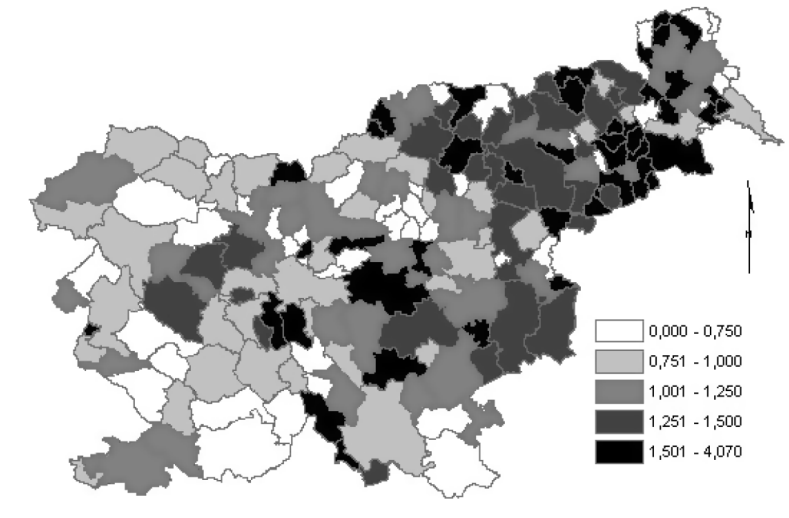
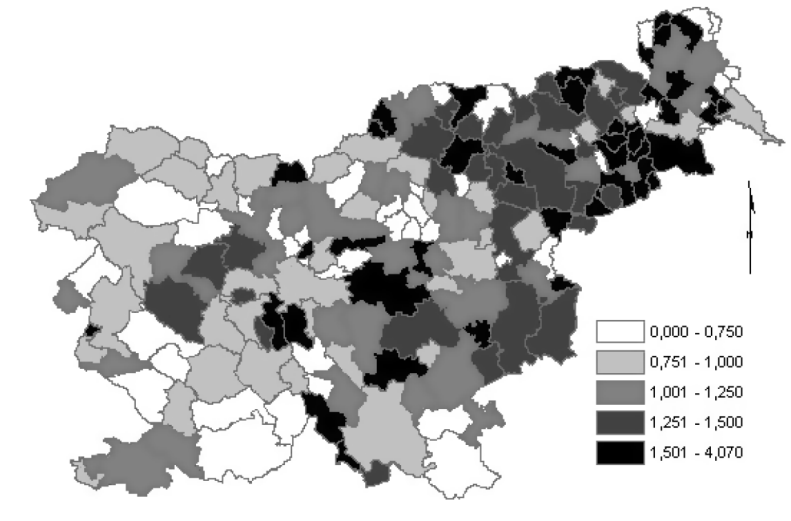


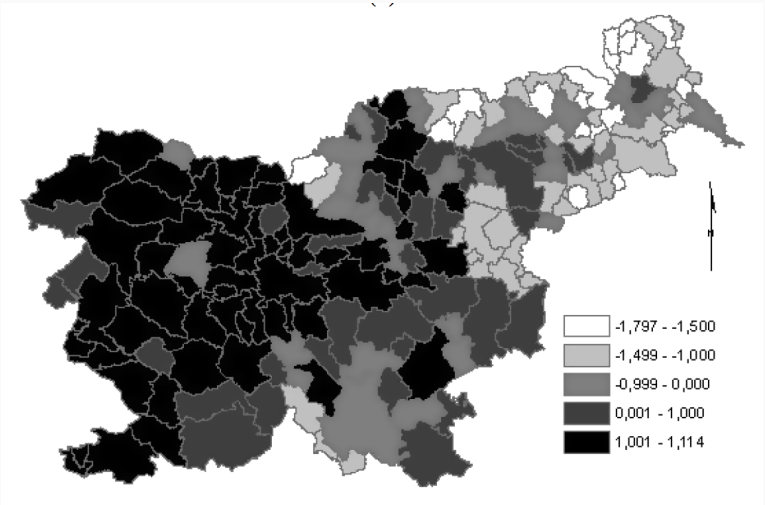
Figure: Standardized stomach cancer incidence in 194 municipalities in Slovenia

- Each datapoint is associated with a region like state, county, municipality etc.
- Usually a result of aggregating point level data

Spatial disease mapping



Standardized cancer incidence



Socio-economic score

Figure: Slovenia stomach cancer data

- Goal: Identify factors (covariates) associated with the disease
- Goal: Identify spatial pattern, if any, and smooth spatially
- Inference is often restricted only to the given set of regions

Key Issues

- Is there *spatial* pattern? *Spatial pattern* implies that observations from units closer to each other are more similar than those recorded in units farther away.
- Do we want to *smooth* the data? Perhaps to adjust for low population sizes (or sample sizes) in certain units? How much do we want to smooth?
- Inference for *new* areal units? Is prediction meaningful here? If we modify the areal units to new units (e.g. from zip codes to county values), what can we say about the new counts we expect for the latter given those for the former? This is the Modifiable Areal Unit Problem (MAUP) or Misalignment.

Proximity Matrices

- A, entries a_{ij}, $(a_{ii} = 0)$; choices for a_{ij}:
  - $a_{ij} = 1$ if i, j share a common boundary (possibly a common vertex)
  - a_{ij} is an *inverse* distance between units
  - $a_{ij} = 1$ if distance between units is $\leq K$
  - $a_{ij} = 1$ for m nearest neighbors.
- A need not be symmetric.
- $\widetilde{A}$: standardize row i by $A_{i+} = \sum_j a_{ij}$ (row stochastic but need not be symmetric).
- A elements often called "weights"; nicer interpretation?

- Note that proximity matrices are user-defined.

- We can define distance intervals, $(0, d_1]$, $(d_1, d_2]$, and so on.
  - First order neighbours: all units within distance d_1.
  - First order proximity matrix $A^{(1)}$. Analogous to A, $a_{ij}^{(1)} = 1$ if i and j are first order neighbors; 0 otherwise.
  - Second order neighbors: all units within distance d_2, but separated by more than d_1.
  - Second order proximity matrix $A^{(2)}$; $a_{ij}^{(2)} = 1$ if i and j are second order neighbors; 0 otherwise
  - And so on...

- There are analogues for areal data of the empirical correlation function and the variogram.

- Moran's I: analogue of lagged autocorrelation

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

  I is not supported on $[-1, 1]$.

- Geary's C: analogue of Durbin-Watson statistic

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

- Both are asymptotically normal if Y_i are i.i.d., the first with mean $-1/(n-1)$ and the second with mean 1.

- Significance testing using a Monte Carlo test, permutation invariance

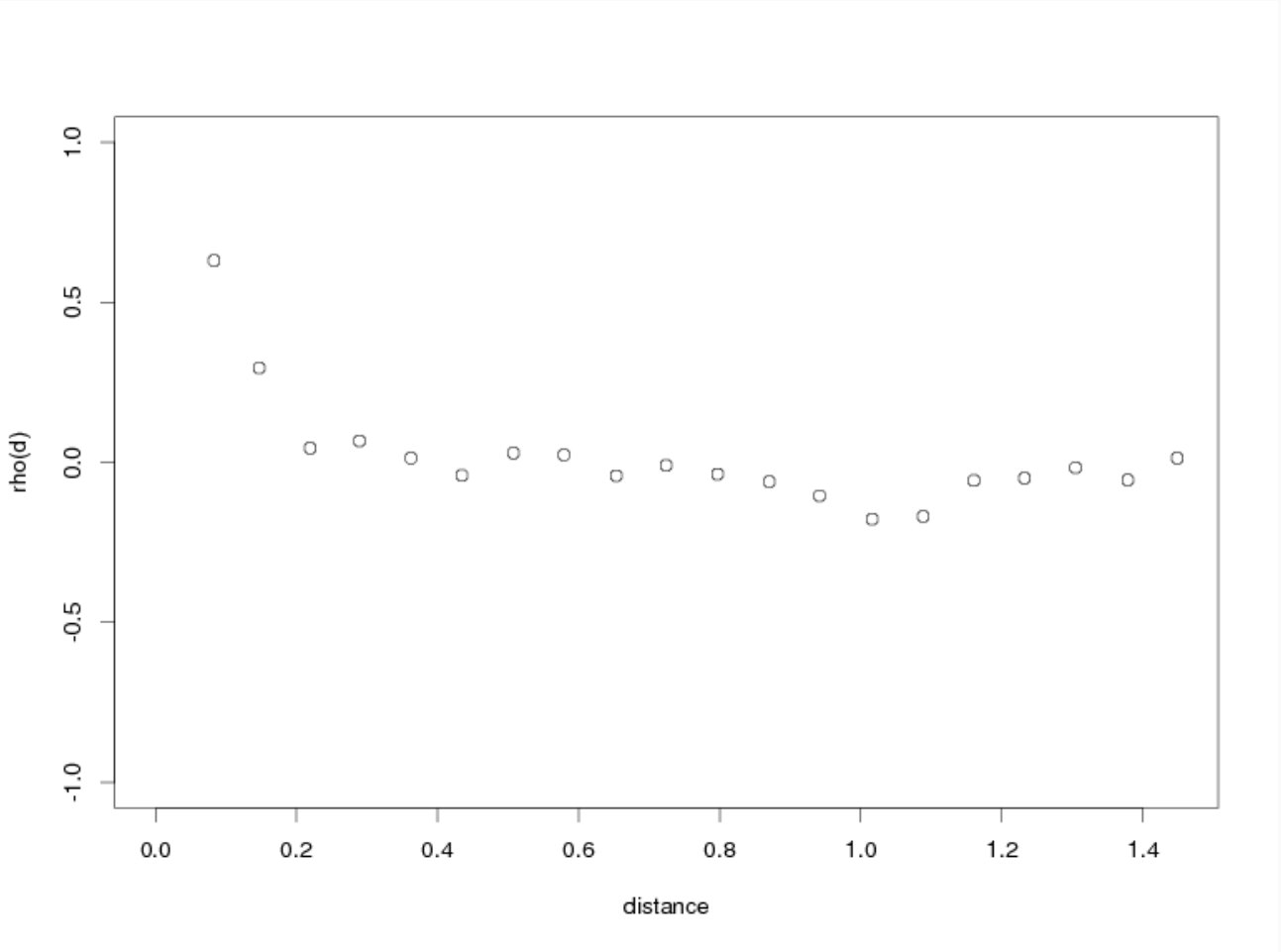
- The *areal correlogram* is a useful tool to study spatial association with areal data.

- Working with I, we can replace a_{ij} with $a_{ij}^{(1)}$ taken from $A^{(1)}$ and compute $\rightarrow I^{(1)}$

- Next replace a_{ij} with $a_{ij}^{(2)}$ taken from $A^{(2)}$ and compute $\rightarrow I^{(2)}$, etc.

- Plot $I^{(r)}$ vs. r

- If there is spatial pattern, we expect $I^{(r)}$ to decline in r initially and then vary about 0.


rho(d)
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
distance

Spatial smoothers

- To smooth Y_i, replace with $\hat{Y}_i = \frac{\sum_i a_{ij} Y_j}{a_{i+}}$ Note: K-nearest neighbours (KNN) regression falls within this framework.

- More generally,

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

  Linear (convex) combination, shrinkage

- Model-based smoothing, e.g.,
  $E(Y_i | \{Y_j,\ j = 1, 2, ..., n\})$

Markov Random Fields

- First, consider $Y = (y_1, y_2, ..., y_n)$ and consider the set $\{p(y_i \mid y_j,\ j \neq i)\}$
- We know $p(y_1, y_2, ...y_n)$ determines $\{p(y_i \mid y_j, j \neq i)\}$ (full conditional distributions)
- ??? Does $\{p(y_i \mid y_j, j \neq i)\}$ determine $p(y_1, y_2, ...y_n)$? If so, we call the joint distribution a *Markov Random Field*.
- In general we cannot write down an arbitrary set of conditionals and assert that they determine the joint distribution. Example:

$$Y_1 \mid Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$$
$$Y_2 \mid Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2).$$

- The first equation implies that $E[Y_1] = \alpha_0 + \alpha_1 E[Y_2]$, i.e., $E[Y_1]$ is linear in $E[Y_2]$. The second equation implies that $E[Y_2] = \beta_0 + \beta_1 E[Y_1^3]$, i.e. $E[Y_2]$ is linear in $E[Y_1^3]$. Clearly this isn't true in general. Hence no joint distribution.

- Also $p(y_1, \ldots, y_n)$ may be improper even if all the full conditionals are proper.

$$p(y_1, y_2) \propto \exp\{(y_1 - y_2)^2\}$$

But $p(Y_2 \mid Y_1) \propto N(Y_2)$ and $p(Y_1 \mid Y_2) \propto N(Y_2, 1)$. Yet the joint distribution is improper.

- Compatibility: Brook's Lemma. Let $y_0 = (y_{10}, \ldots, y_{n0})$ be any fixed point in the support of $p(\cdot)$.

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

If LHS is proper, the fact that it integrates to 1 determines the normalizing constant!

GLM for Spatial disease mapping

- At unit (region) i, we observe response y_i and covariate x_i
- $g(E(y_i)) = x_i^\top \beta + w_i$ where $g(\cdot)$ denotes a suitable link function

Hierarchical areal model:

$$\prod_{i=1}^{k} p_1(y_i | x_i' \beta + w_i) \times N^{-1}(w \,|\, 0, \tau_w Q(\rho)) \times p_2(\beta, \tau_w, \rho)$$

- Notation: $N^{-1}(m, Q)$ denotes normal distribution with mean m and precision (inverse covariance) Q
- p_1 denotes the functional form of the density corresponding to the link $g(\cdot)$

How to model $Q(\rho)$

- Choice of $Q(\rho)$ should enable spatial smoothing
- One possibility: Represent each region by a single point and use Gaussian Process covariance i.e. $Q(\rho)_{ij}^{-1} = C(m(i), m(j))$
- Many possible choices to map the region i into a Euclidean coordinate $m(i)$
- Is it appropriate to represent a large area with a single point?
- Also GP approach is computationally very expensive
- Alternate approach: Represent spatial information in terms of a graph depicting the relative orientation of the regions

CAR models

- Conditional autoregressive (CAR) model (Besag, 1974; Clayton and Bernardinelli, 1992)
- Areal data modeled as a graph or network: V is the set of vertices (regions)
- $i \sim j$ if regions i and j share a common border
- Adjacency matrix $A = (a_{ij})$ such that $a_{ij} = I(i \sim j)$
- n_i is the number of neighbors of i
- CAR model:

$$w_i \mid w_{-i} \sim N^{-1}\left(\frac{\rho}{n_i} \sum_{j \mid i \sim j} w_j, \tau_w n_i\right)$$

CAR models

- CAR model:

$$w_i \mid w_{-i} \sim N^{-1}\left(\frac{1}{n_i}\sum_{j \mid i \sim j} w_j, \tau_w n_i\right)$$

- $w = (w_1, w_2, \ldots, w_k)^\top \sim N^{-1}(0, \tau_w(D - \rho A))$ where $D = diag(n_1, n_2, \ldots, n_k)$
- $\rho = 1 \Rightarrow$ Improper distribution as $(D - A)1 = 0$ (ICAR)
  - Can be still used as a prior for random effects
  - Cannot be used directly as a data generating model

CAR models

- CAR model:
$$w_i \mid w_{-i} \sim N^{-1}\left(\frac{1}{n_i}\sum_{j \mid i \sim j} w_j, \tau_w n_i\right)$$

- $w = (w_1, w_2, \ldots, w_k)^\top \sim N^{-1}(0, \tau_w(D - \rho A))$ where $D = diag(n_1, n_2, \ldots, n_k)$

- $\rho = 1 \Rightarrow$ Improper distribution as $(D - A)1 = 0$ (ICAR)
  - Can be still used as a prior for random effects
  - Cannot be used directly as a data generating model

- $\rho < 1 \Rightarrow$ Proper distribution with added parameter flexibility

SAR models

- Simultaneous Autoregressive (SAR) model (Whittle, 1954)
- Instead of taking the conditional route, SAR model proceeds by simultaneously modeling the random effects

$$w_i = \rho \sum_{i \neq j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k$$

- $\epsilon_i \overset{ind}{\sim} N^{-1}(0, \tau_i)$ are errors independent of w
- A common choice is to define $b_{ij} = I(i \sim j)/n_i$
- Joint distribution: $w \sim N^{-1}(0, (I - \rho B)^\top F(I - \rho B))$, $B = (b_{ij})$ and $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$
- $\rho = 1 \Rightarrow$ Improper distribution

Interpretation of ρ in proper CAR and SAR models

- Calibration of ρ as a correlation, e.g., (as reported in Banerjee et al. 2014)

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

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

Interpretation of ρ in proper CAR and SAR models

- ρ cannot be interpreted as correlation between neighboring w_i's (Wall, 2004; Assuncao and Krainski, 2009)

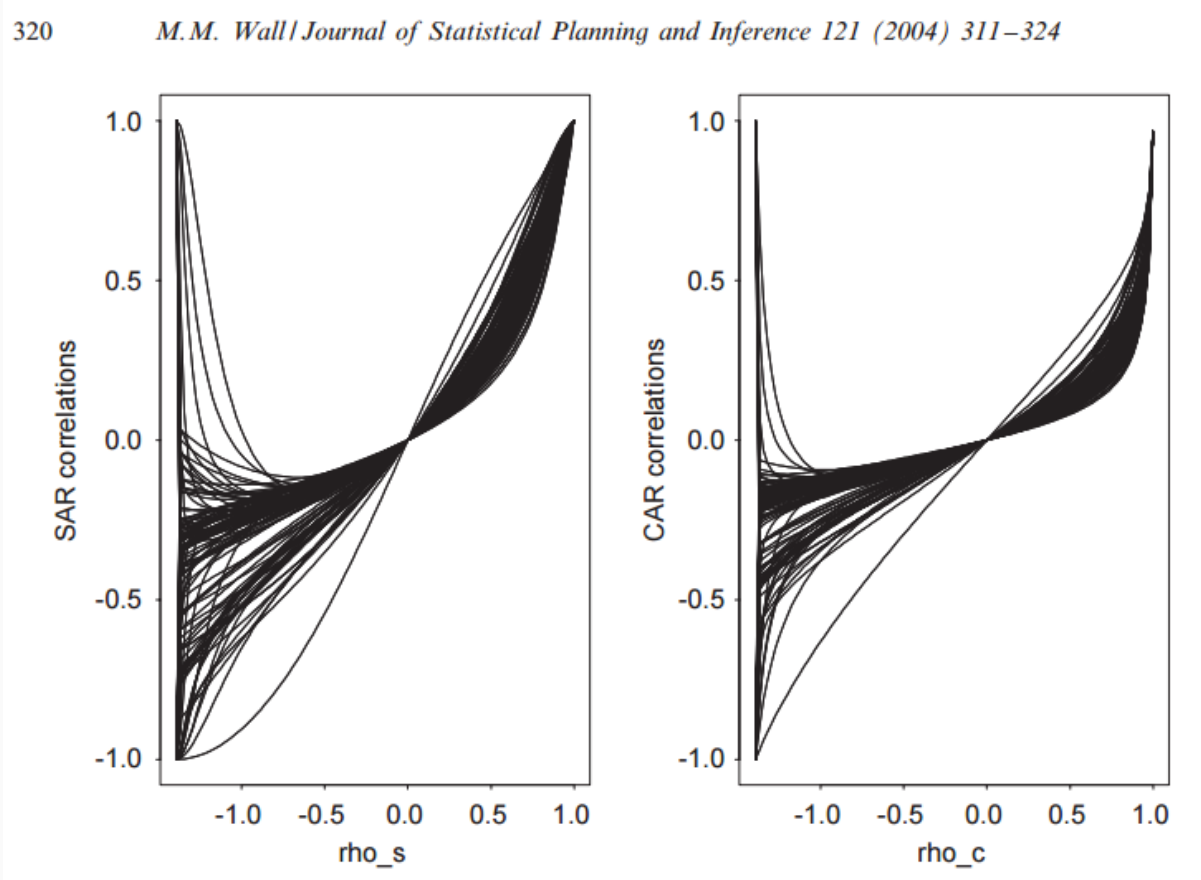


Figure: Neighbor pair correlations as a function of ρ for proper CAR and SAR models over the graph of US states

SAR model and Cholesky factors

- General SAR model:

$$w_i = \sum_{i \neq j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k$$

- $w \sim N^{-1}(0, (I - B)^\top F(I - B))$ where $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$
- Only proper when $I - B$ is invertible which is not guaranteed for arbitrary B
- SAR is essentially modeling the precision matrix through the Cholesky factor $I - B$

SAR model and Cholesky factors

- General SAR model:

$$w_i = \sum_{i \neq j} b_{ij} w_j + \epsilon_i \text{ for } i = 1, 2, \ldots, k$$

- $w \sim N^{-1}(0, (I - B)^\top F(I - B))$ where $F = diag(\tau_1, \tau_2, \ldots, \tau_k)$
- Only proper when $I - B$ is invertible which is not guaranteed for arbitrary B
- SAR is essentially modeling the precision matrix through the Cholesky factor $I - B$
- Cholesky factors are not unique
- We can always choose a lower triangular Cholesky factor

New model

$$
\begin{aligned}
w_1 &= \epsilon_1 \\
w_2 &= b_{21} w_1 + \epsilon_2 \\
w_3 &= b_{31} w_1 + b_{32} w_2 + \epsilon_3 \\
&\vdots \\
w_k &= b_{k1} w_1 + b_{k1} w_2 + \ldots + b_{k,k-1} w_{k-1} + \epsilon_k
\end{aligned}
$$

- $B = (b_{ij})$ is now a strictly lower triangular matrix.

New model

- Advantages of lower triangular B:
  - $w \sim N^{-1}(0, (I-B)^\top F(I-B))$ is a proper distribution for any choice of lower triangular B
  - $\det(L^\top FL) = \prod_{i=1}^{n} \tau_i$ where $F = \mathrm{diag}(\tau_1, \ldots, \tau_k)$ and $L = I - B$
  - $w^\top L^\top FLw = \tau_1 w_1^2 \; + \sum_{i=2}^{k} \tau_i (w_i - \sum_{\{j<i\}} w_j b_{ij})^2$
  - Likelihood $N^{-1}(w \,|\, 0, (I-B)^\top F(I-B))$ can be computed using $O(k+s)$ flops where s denotes the sparsity (number of non-zero entries) of B.
  - Even if k is large, evaluation of likelihood is fast if each region only shares border with a few others

Choice of B and F

- How to specify B and F?
- Sparsity of B is desirable
- If data had replicates for each region, there is large literature on fully data driven estimation of sparse Cholesky factors (Wu and Pourahmadi, 2003; Huang et al., 2006; Rothman et al., 2008; Levina et al., 2008; Wagaman and Levina, 2009; Lam and Fan, 2009)
- Unfortunately many areal datasets lack replication

Choice of B and F

- How to specify B and F?
- Sparsity of B is desirable
- Like in NNGP set $b_{ij} = 0$ for j outside neighbor sets $N(i)$
  - Pros: For graphs neighbor sets are naturally chosen: $N(i) = \{j \,|\, j \sim i, j < i\}$
  - Cons: There is no covariance function on arbitrary graphs from which we can obtain non-zero b_{ij}'s and F

Autoregressive models on trees

- $D = (d_{ij})$ is the shortest distance matrix on the graph
- If the graph was a tree (no loops), then $\rho^D = (\rho^{d_{ij}}$ is then a valid *autoregressive* correlation matrix (AR(1) model on a tree, Basseville et al., 2006).
- Areal graphs are loopy and are not usually trees

Local embedded spanning trees

- Embedded spanning trees (EST) of a graph G is a subgraph of G which is a tree and spans all the vertices of G
- Note that to specify $w_i = \sum_{j \in N(i)} b_{ij} w_j + \epsilon_i$ we only need a joint distribution on $\{i\} \cup N(i)$
- Let G_i denote the subgraph of G which includes vertices $\{i\} \cup N(i)$ and the edges among them
- The subgraph T_i of G_i which only contains the edges $\{i \sim j \,|\, j \in N(i)\}$ is an embedded spanning tree of G_i
- Use the local embedded spanning trees T_i to specify the b_{ij}'s and τ_i

Directed acyclic graph autoregressive (DAGAR) model

- AR_i denotes the $AR(1)$ distribution on T_i
- Solve for b_{ij} and τ_i such that $E_{AR_i}(w_i \mid w_{N(i)}) = \sum_{j \in N(i)} b_{ij} w_j$ and $\tau_i = 1/Var_{AR_i}(w_i \mid w_{N(i)})$
- No edge is left out !

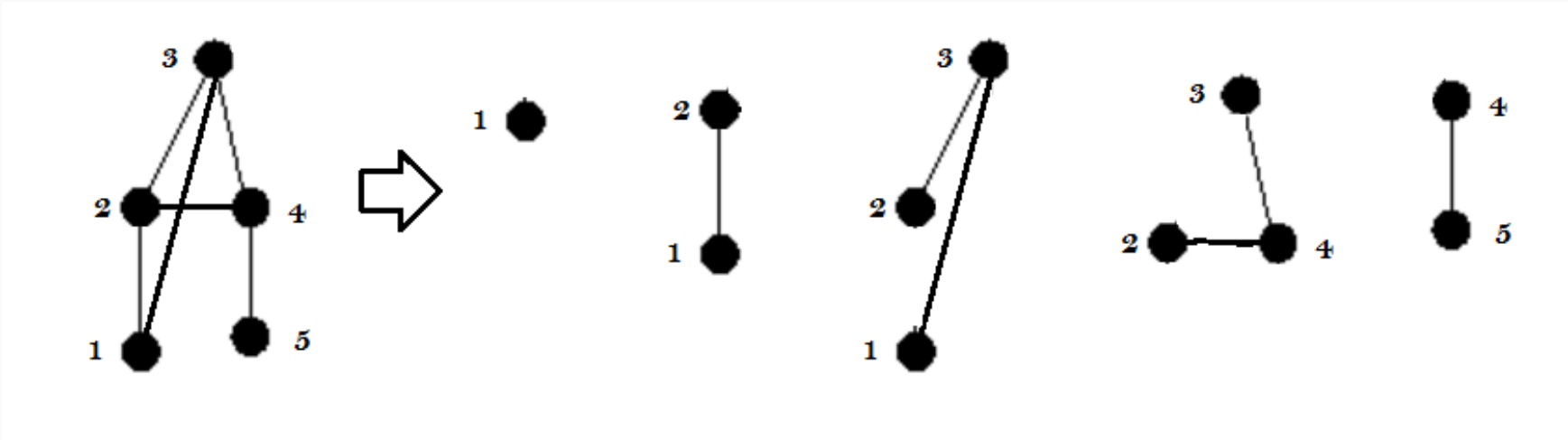


Figure: Decomposing a graph into a sequence of embedded spanning trees

Properties of DAGAR models

- $b_{ij} = b_i = \rho/(1 + (|N(i)| - 1)\rho^2)$
- $\tau_i = (1 + (|N(i)| - 1)\rho^2)/(1 - \rho^2)$
- $\det(Q_{DAGAR}) = \prod_{i=1}^{k} \tau_i$
- Positive definite for any $0 \leq \rho \leq 1$
- Interpretability of ρ:
  - If the graph is a tree, then DAGAR model is same as the AR(1) model on the tree i.e. correlation between d^{th} order neighbors is ρ^d for $d = 1, 2, \ldots$
  - If the graph is a closed two-dimensional grid, then each neighbor pair correlation is ρ
- $p_{DAGAR}(w)$ can be stored and evaluated using $O(e + k)$ flops where e is the total number of neighbor pairs

Dependence on ordering

- DAGAR model depends on the ordering of the regions when decomposing into local trees
- We can define a DAGAR model for every ordering
- Spatial regions do not have natural ordering
- How to choose the ordering?
- Coordinate based orderings were used in Datta et al., 2016; Stein, 2004; Vecchia, 1988
- Model averaging over orderings ? Too many possibilities ($k!$)

Order-free model

- Let Q be the average over DAGAR precision matrices corresponding to all $k!$ possible orderings

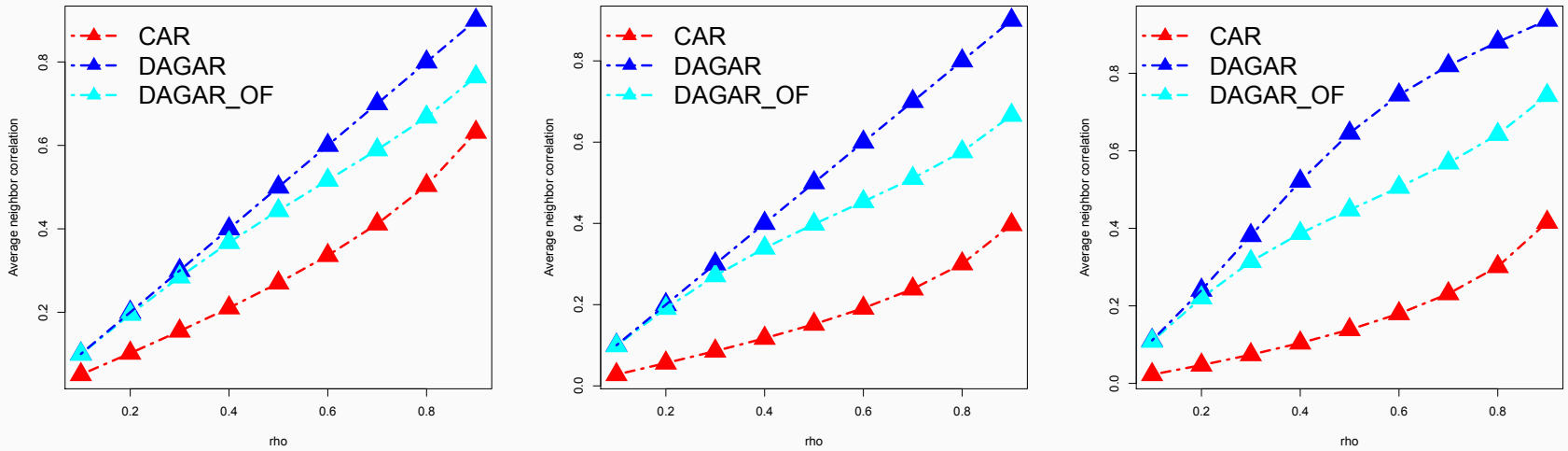
Order-free model

- Let Q be the average over DAGAR precision matrices corresponding to all $k!$ possible orderings
- Q is is free of ordering and available in closed form
- $Q(i, j)$ is non-zero if and only if either $i \sim j$ or $i \approx j$

Order-free model

- Sparsity of Q is e_2 where e_2 is the number of edges in the second order graph (moral graph) created from G
- As $e_2 > e$, Q is less sparse than the CAR model or the ordered DAGAR model precision matrix and has higher flop count
- Total computational total cost for evaluating Q is $O(e_2 n_{\max})$
- $e_2 < k n_{\max}(n_{\max} + 1)/2$ where $n_{\max} = \max(n_i)$
- If $n_{\max}$ is small, i.e., as long as each region only shares border with a few others (which is often the case), Q is still quite sparse even for large k

Interpretation of ρ



path graph grid graph USA state map

Figure: Average neighbor pair correlations as a funcion of ρ for proper CAR and DAGAR models

Simulated data analysis

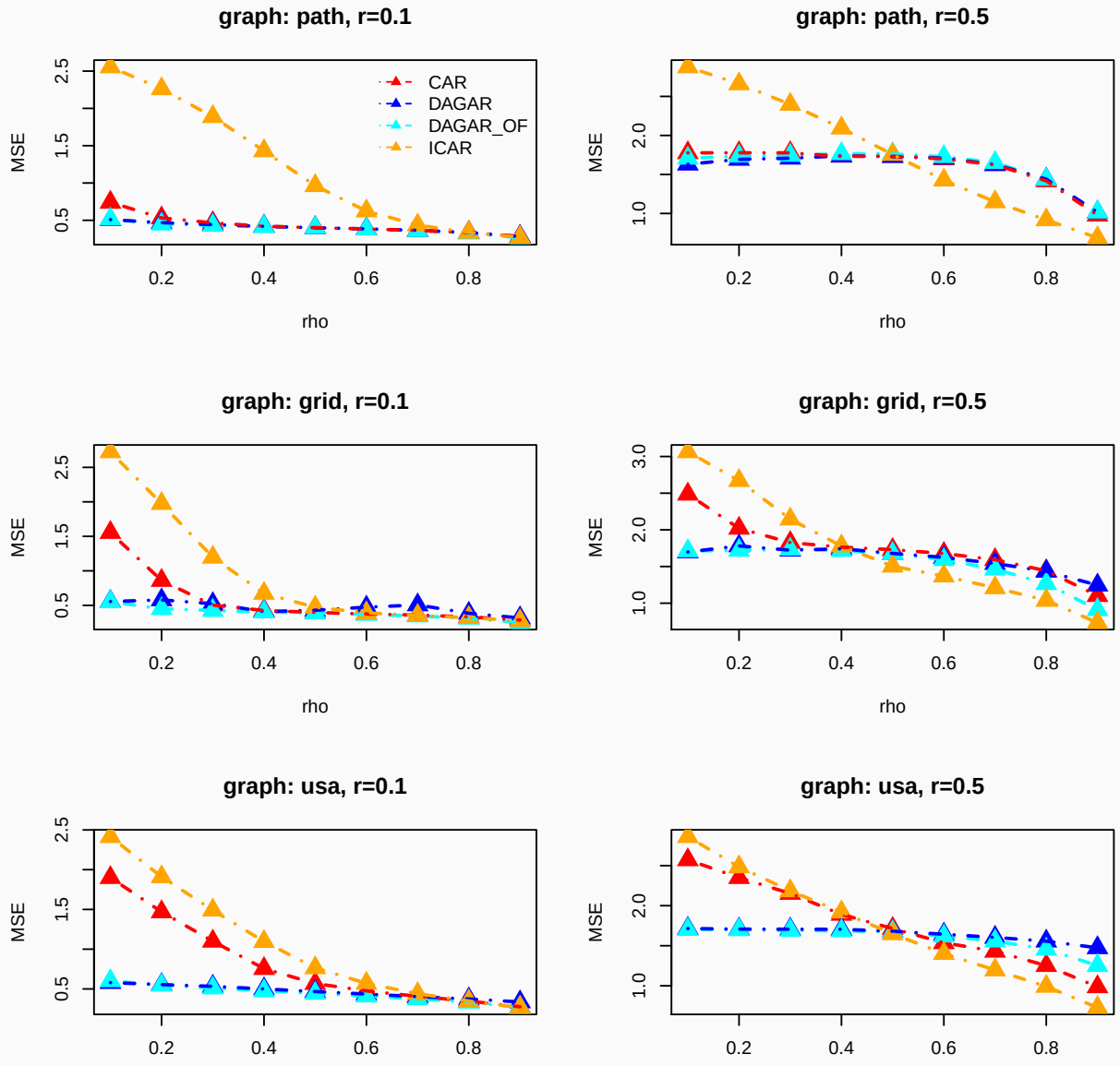
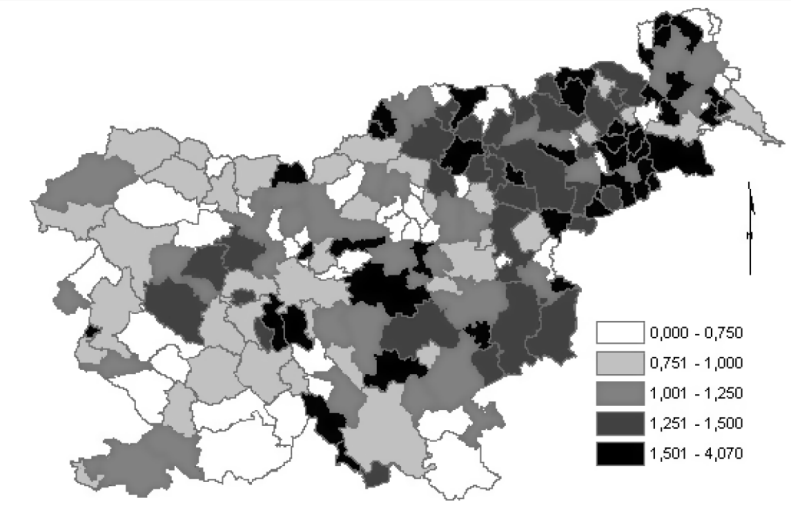
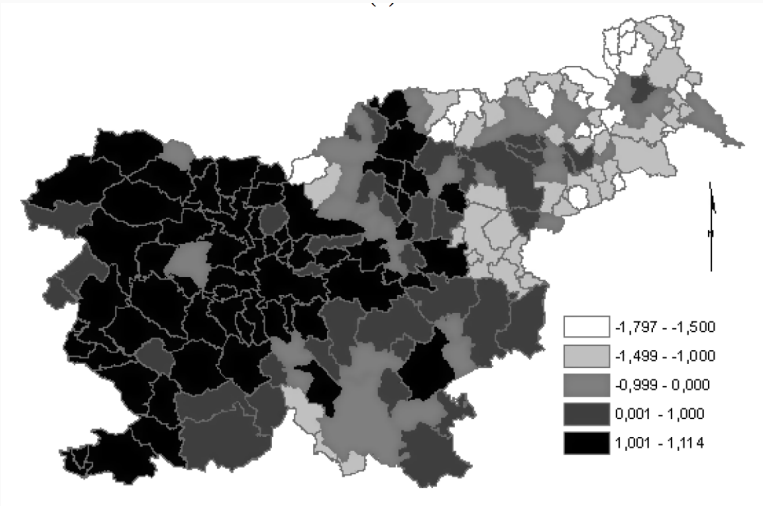


Figure: Mean square error as a function of ρ and $r = \tau^2/\sigma^2$ for DAGAR and CAR models

Slovenia stomach cancer data



Standardized cancer incidence



Socio-economic score

Figure: Slovenia stomach cancer data

- Observed (O_i) and expected (E_i) number of cancer counts for each of the 194 municipalities of the country
- $O_i \sim Poisson(E_i \exp(\alpha + \beta SE_i + w_i))$ where $w \sim N^{-1}(0, \tau_w Q(\rho))$

Slovenia stomach cancer data

Table: Parameter estimates with confidence intervals and model comparison metrics

	α	β	ρ	DIC	LPPD_{LOOCV}[1]
CAR	0.09 (0.02, 0.16)	-0.12 (-0.19, -0.04)	0.33 (0.02, 0.86)	1097	1170
DAGAR	0.11 (0.03, 0.18)	-0.12 (-0.19, -0.06)	0.08 (0.004, 0.24)	1091	1127
DAGAR_{OF}	0.11 (0.05, 0.17)	-0.12 (-0.18, -0.06)	0.06 (0.003, 0.2)	1090	1133

- Zadnik and Reich (2006) observed spatial confounding with ICAR model ($\hat{\beta}_{ICAR} = -0.02(-0.10, 0.06)$)
- Here for all three models the CIs for β lie outside zero
- Estimates of ρ are much smaller than 1
- Estimates of β here are closer to those obtained in the non-spatial (NS) analysis ($\hat{\beta}_{NS} = -1.4(-0.17, -0.10)$)

[1] Log-predictive posterior density using Leave one out cross validation

Summary

- DAGAR models for areal data constructed from sparse Cholesky factors
- Scalability for large areal data
- Ordered vs order-free DAGAR
  - For all analysis, ordered model performed very similar to the order-free model
  - Ordered model is faster with theoretical results about interpretability of ρ
- DAGAR models are positive definite and can be directly used to model or simulate any multivariate data on graphs (like imaging or social network data)
- Better performance than CAR modes for many scenarios
- DAGAR available at https://arxiv.org/pdf/1704.07848.pdf