Spatio-temporal statistics (MATH4341) Michaelmas term, 2023

Handout 4: Aerial unit data / spatial data on lattices

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce Aerial unit data modeling: the basic building models.

Reading list & references:

[1] Cressie, N. (2015; Part II). Statistics for spatial data. John Wiley & Sons.
[2] Gaetan, C., & Guyon, X. (2010; Ch 3). Spatial statistics and modeling (Vol. 90). New York: Springer.

Specialized reading.

[3] Kent, J. T., & Mardia, K. V. (2022). Spatial analysis (Vol. 72). John Wiley & Sons. (on Spatial analysis)

Part 1. Basic stochastic models & related concepts for model building

Note 1. Recall from Section 2.2 of "Handout 1: Types of spatial data" that modeling aerial unit / lattice data types involves the use of random field models with a discrete index set. Such data are collected over areal units such as pixels, census districts or tomographic bins. Often, there is a natural adjacency relation or neighborhood structure.

Note 2. This means we need to introduce new basic building models able to acceptably represent the characteristics of the underline data generating mechanisms. These as the "Discrete Random Fields".

1. Discrete Random Fields

Note 3. We re-introduce the definition of the random field adjusting it to the aerial unit data framework.

Definition 4. A random field $Z = (Z_s; s \in \mathcal{S})$ on a set of indexes $\mathcal{S}$ taking values in $\mathcal{Z}^{\mathcal{S}}$ is a family of random variables $\{Z_s := Z_s(\omega); s \in \mathcal{S}, \omega \in \Omega\}$ where each $Z_s(\omega)$ is defined on the same probability space $(\Omega, \mathfrak{F}, \mathrm{pr})$ and taking values in $\mathcal{Z}$.

Note 5. In aerial unite data modeling, the (spatial) set of sites $\mathcal{S}$, at which the process is defined, is discrete, it can be finite or infinite (e.g. $\mathcal{S} \subseteq \mathbb{Z}^d$), regular (e.g. pixels of an image) or irregular (states of a country).

Note 6. The general state space $\mathcal{Z}$ of the random field can be quantitative, qualitative or mixed. E.g., $\mathcal{Z} = \mathbb{R}^+$ in a Gamma random field, $\mathcal{Z} = \mathbb{N}$ in a Poisson random field, $\mathcal{Z} = \{0, 1\}$ in a binary random field.

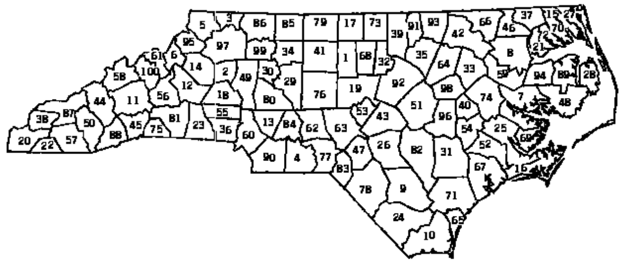


FIGURE 1.1. Lattice of spatial sites for North Carolina SIDS data. Each county is a site. Each site is coded according to its alphabetical order. The collection of sites is the lattice of sites.

Note 7. If $\mathcal{Z}$ is finite or countably infinite, the (joint)distribution of Z has a PMF

$$\mathrm{pr}_Z(z) = \mathrm{pr}(Z = z) = \mathrm{pr}(\{Z_s = z_s;\ s \in \mathcal{S}\}),\ \forall z_s \in \mathcal{Z}^{\mathcal{S}}$$

otherwise if $\mathcal{Z} \subseteq \mathbb{R}^d$ and Z continuous we will use the joint PDF.

Definition 8. The discrete set of sites $\mathcal{S} = \{s_i; i = 1, ..., n\}$ is often called lattice of sites.

Notation 9. We will more often use the notation Z_s instead of $Z(s)$ or Z_i instead of $Z(s_i)$. Hence, since $\mathcal{S} = \{s_i; i = 1, ..., n\}$, we can consider a more convenient notation

$$Z = (Z_s; s \in \mathcal{S})^\top = (Z_i = Z(s_i); i = 1, ...n)^\top .$$

Notation 10. The notation $i \sim j$ between two sites $i, j \in \mathcal{S}$ means that "sites i and j are adjacent".

Example 11. Recall the North Carolina SIDS data Ex 24 in Handout 1. Fig 1.1 presents the sites and the lattice of sites. Each county is a site. Each site is coded according to its alphabetical order. The collection of sites is the lattice of sites coded according to alphabetical order of the county name. One may define the "adjacency between sites $i \sim j$" as the counties that share common boarders. Then for site $i = 43$, $i \sim j$ involves any $j \in \{63, 53, 19, 92, 51, 82, 26, 47\}$ in Fig 1.1.

Example 12. (Logistic/Ising model) Consider Z_i denotes a characteristic presence coded as 1 or absence coded as 0 on a region labeled by $i \in \mathcal{S}$. Then $\mathcal{Z} = \{0, 1\}$. The Ising model is defined by the (joint) PMF

$$\mathrm{pr}_Z(z) \propto \exp\left(\alpha \sum_{i \in \mathcal{S}} z_i + \beta \sum_{\{i,j\}: i \sim j} z_i z_j\right),\ \forall z \in \mathcal{Z}^{\mathcal{S}} \tag{1.1}$$

E.g., it can model a black & white image noisy image, where $\mathcal{S}$ denotes the labels of the image pixels, and Z_i denotes the presence of a black pixel ($Z_i = 1$) or its absence ($Z_i = 0$). Under Ising model, the characteristic is observed with probability $\mathrm{pr}_{Z_i}(z_i = 1) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$

when $\beta = 0$. The characteristic's present is encouraged in neighboring sites when $\beta > 0$, and discouraged when $\beta < 0$.

Notation 13. We use notation, for $\mathcal{A} \subset \mathcal{S}$

$$\mathrm{pr}_{\mathcal{A}} \left(z_{\mathcal{A}} | z_{\mathcal{S} \backslash \mathcal{A}} \right) = \mathrm{pr} \left(Z_{\mathcal{A}} = z_{\mathcal{A}} | Z_{\mathcal{S} \backslash \mathcal{A}} = z_{\mathcal{S} \backslash \mathcal{A}} \right)$$

Definition 14. Local characteristics of a random field Z on $\mathcal{S}$ with values in $\mathcal{Z}$ are the conditionals

$$\mathrm{pr}_i \left(z_i | z_{\mathcal{S}-i} \right) = \mathrm{pr}_{\{i\}} \left(z_{\{i\}} | z_{\mathcal{S} \backslash \{i\}} \right), \, i \in \mathcal{S}, \, z \in \mathcal{Z}$$

Example 15. The characteristics of the Ising model in (1.1) are

$$\mathrm{pr}_i \left(z_i = 1 | z_{\mathcal{S}-i} \right) = \frac{\exp \left(\alpha + \beta \sum_{\{i,j\}: i \sim j} z_j \right)}{1 + \exp \left(\alpha + \beta \sum_{\{i,j\}: i \sim j} z_j \right)}$$

2. Compatibility of conditional distributions

Note 16. Essentially, we attempt to answer the following question. Under what conditions a parameterized family $\{ \pi_i \left(z_i | z_{\mathcal{S}-i} \right); i \in \mathcal{S} \}$ of distributions on $\mathcal{S}$ conditioned on $z_{\mathcal{S}-i}$ can represent conditional distributions of a joint distribution $\mathrm{pr}_Z (\cdot)$?

Note 17. To answer the above we need to be able to specify partially or wholly the joint and conditional distribution of pr_Z. However, an arbitrary chosen set of conditional distributions $\{ \pi_i (\cdot|\cdot) \}$ is not generally compatible, and hence we need to impose conditions.

Proposition 18. [1]*(Compatibility condition) Let F be a joint distribution with $dF(x, y) = f(x, y)\, d(x, y)$ on $\mathcal{S}_x \times \mathcal{S}_y$. Let candidate condition distributions*

$$G \text{ with } dG(x|y) = g(x|y)\, dx, \text{ on } x \in \mathcal{S}_x$$
$$Q \text{ with } dQ(y|x) = q(y|x)\, dy, \text{ on } y \in \mathcal{S}_y$$

and let $N_g = \{(x, y) : g(x|y) > 0\}$ and $N_q = \{(x, y) : q(y|x) > 0\}$. A distribution F with conditionals exists iff

1. *$N_g = N_q = N$*
2. *there exist functions u and $v$where $g(x|y) / q(y|x) = u(x) v(y)$ for all $(x, y) \in N$ and $\int u(x)\, dG(x|y) < \infty$*

[1] Arnold, B. C., & Press, S. J. (1989). Compatible conditional distributions. Journal of the American Statistical Association, 84(405), 152-156.

Note 19. Essentially the above conditions guarantee that

$$k(y)\, g(x|y) = f(x, y) = h(x)\, q(y|x)$$

where k, g, h, q are densities.

Example 20. The conditionals $x|y \sim \mathrm{N}(a + by, \sigma^2 + \tau^2 y^2)$ and $y|x \sim \mathrm{N}(c + dx, \tilde{\sigma}^2 + \tilde{\tau}^2 x^2)$ are compatible if $\tau^2 = \tilde{\tau}^2 = 0$ and $d/\tilde{\sigma}^2 = b/\sigma^2$.

Solution. See Ex 24 in the Exercise sheet.

Note 21. Proposition 18 can be extended to more dimensions.

Note 22. The following theorem shows that local characteristics can determine the entire distribution in certain cases.

Theorem 23. *(Besag's factorization theorem; Brook's Lemma) Let Z be a $\mathcal{Z}$ valued random field taking values in $\mathcal{Z}^{\mathcal{S}}$ where $\mathcal{S} = \{1, ..., n\}$ with $n \in \mathbb{N}$, and such as $pr_Z(z) > 0$, $\forall z \in \mathcal{Z}^{\mathcal{S}}$. Then for all*

$$\frac{pr_Z(z)}{pr_Z(z^*)} = \prod_{i=1}^{n} \frac{pr_i\left(z_i|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}{pr_i\left(z_i^*|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}, \quad \forall z, z^* \in \mathcal{Z}^{\mathcal{S}} \tag{2.1}$$

Proof. I will show that

$$\mathrm{pr}_Z(z) = \prod_{i=1}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)} \mathrm{pr}_Z(z^*)$$

It is

$$\mathrm{pr}_Z(z_1, ..., z_n) = \frac{\mathrm{pr}_n(z_n|z_1, ..., z_{n-2}, z_{n-1})}{\mathrm{pr}_n(z_n^*|z_1, ..., z_{n-2}, z_{n-1})} \mathrm{pr}_Z(z_1, ..., z_{n-1}, z_n^*)$$

Let proposition P_j be

$$\mathrm{pr}_Z(z) = \prod_{i=n-j}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1, ..., z_{i-1}, z_{i+1}^*, ..., z_n^*\right)} \mathrm{pr}_Z\left(z_1, ..., z_{n-j-1}, z_{n-j}^*, ..., z_n^*\right)$$

Proposition P_0 is true

$$\mathrm{pr}_Z(z) = \frac{\mathrm{pr}_n(z_n|z_1, ..., z_{n-1})}{\mathrm{pr}_n(z_n^*|z_1, ..., z_{n-1})} \mathrm{pr}_Z(z_1, ..., z_{n-1}, z_n^*) \tag{2.2}$$

Proposition P_1 is true

$$\mathrm{pr}_Z(z_1, ..., z_{n-1}, z_n^*) = \frac{\mathrm{pr}_{n-1}(z_{n-1}|z_1, ..., z_{n-2}, z_n^*)}{\mathrm{pr}_{n-1}\left(z_{n-1}^*|z_1, ..., z_{n-2}, z_n^*\right)} \mathrm{pr}_Z\left(z_1, ..., z_{n-2}, z_{n-1}^*, z_n^*\right)$$

Assume that P_j is true. Then proposition $\mathrm{P}_{j+1}$is true as well, because

$$\begin{aligned}
\mathrm{pr}_Z(z) &= \prod_{i=n-j}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z_{i+1}^*,...,z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1,...,z_{i-1},z_{i+1}^*,...,z_n^*\right)} \mathrm{pr}_Z\left(z_1,...,z_{n-j-1},z_{n-j}^*,...,z_n^*\right) \\
&= \prod_{i=n-j}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z_{i+1}^*,...,z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1,...,z_{i-1},z_{i+1}^*,...,z_n^*\right)} \\
&\quad \times \frac{\mathrm{pr}_{n-j-1}\left(z_{n-j-1}|z_1,...,z_{n-j-2},z_{n-j}^*,...,z_n^*\right)}{\mathrm{pr}_{n-j-1}\left(z_{n-j-1}^*|z_1,...,z_{n-j-2},z_{n-j}^*,...,z_n^*\right)} \mathrm{pr}_Z\left(z_1,...,z_{n-j-2},z_{n-j-1}^*,...,z_n^*\right) \\
&= \prod_{i=n-(j+1)}^{n} \frac{\mathrm{pr}_i\left(z_i|z_1,...,z_{i-1},z_{i+1}^*,...,z_n^*\right)}{\mathrm{pr}_i\left(z_i^*|z_1,...,z_{i-1},z_{i+1}^*,...,z_n^*\right)} \mathrm{pr}_Z\left(z_1,...,z_{n-(j+1)-1},z_{n-(j+1)}^*,...,z_n^*\right)
\end{aligned}$$

Then (2.1) is correct according to the induction principle. □

Note 24. The theorem shows that the joint $\mathrm{pr}_Z(\cdot)$ can be constructed from its conditionals $\{\mathrm{pr}_i(\cdot|\cdot)\}$ if distributions $\{\mathrm{pr}_i(\cdot|\cdot)\}$ are compatible for $\mathrm{pr}_Z(\cdot)$, as this reconstruction has to be invariant wrt the coordinate permutation $\{1,...,n\}$ and the reference state z^*– these invariances correspond to the conditions in Proposition 18.

3. Gaussian Autoregressive models

Modeling snapshot

Definition 25. Adjacency matrix N is called a matrix N with $[N]_{i,j} = 1(i \sim j)$ for some symmetric neighborhood relation $\sim$ on $\mathcal{S}$. It aims at spatially connecting unites i and j.

Definition 26. Proximity matrix W is called a matrix W which aims at spatially connecting unites i and j in some fashion for some symmetric neighbourhood relation $\sim$ on $\mathcal{S}$. Usually $[W]_{i,i} = 0$

3.1. Conditional autoregressive models (CAR).

Definition 27. Assume a random field $Z = (Z_s; s \in \mathcal{S})$ on a set of indexes $\mathcal{S}$ with values in $\mathcal{Z}$. We say that Z follows a conditional autoregressive model (CAR) if the distribution of each element Z_s of the random field Z is specified conditionally on the values at the neighboring sites of s.

3.1.1. *Gaussian CAR.*

Definition 28. Gaussian CAR assumes that the local characteristics $\{\mathrm{pr}_i(z_i|z_{\mathcal{S}-i})\}$ are Gaussian distributions

$$Z_i|z_{\mathcal{S}-i} \sim \mathrm{N}\left(\mu_i + \sum_{j\neq i} b_{i,j}\left(Z_j - \mu_j\right), \kappa_i\right) \tag{3.1}$$

with mean $\mathrm{E}\left(Z_i|Z_{\mathcal{S}-i}\right) = \mu_i + \sum_{j\neq i} b_{i,j}\left(Z_j - \mu_j\right)$ and variance $\mathrm{Var}\left(Z_i|Z_{\mathcal{S}-i}\right) = \kappa_i$ for $i \in \mathcal{S}$.

Proposition 29. *Let $K = diag(\{\kappa_i\})$ with $\kappa_i > 0$, matrix B with $B_{i,i} := [B]_{i,i} = 0$, and real vector μ with suitable dimentions. If Z follows a Gaussian CAR (Def 28), $I - B$ is non-singular, and $(I - B)^{-1} K > 0$, then the joint distribution of Z is*

$$Z \sim N\big(\mu, (I - B)^{-1} K\big) . \tag{3.2}$$

Proof. Without lose of generality, consider zero mean $\mu = 0$ (or equivalently set $Z := Z - \mu$). The full conditionals $Z_i|z_{\mathcal{S}-i}$ in (3.1) are compatible with the joint distribution $\text{pr}_Z(z)$. By using Besag's factorization theorem with reference state $z^* = 0$ we get

$$\begin{aligned}
\text{pr}_Z(z) &= \prod_{i=1}^{n} \frac{\text{pr}_i\left(z_i|z_1, ..., z_{i-1}, z^*_{i+1} = 0, ..., z^*_n = 0\right)}{\text{pr}_i\left(z^*_i = 0|z_1, ..., z_{i-1}, z^*_{i+1} = 0, ..., z^*_n = 0\right)} \text{pr}_Z(z^* = 0) \\
&= \prod_{i=1}^{n} \frac{\text{N}\left(z_i| \sum_{j<i} b_{i,j} z_j + 0, \kappa_i\right)}{\text{N}\left(0| \sum_{j<i} b_{i,j} z_j + 0, \kappa_i\right)} \text{pr}_Z(z^* = 0) \\
&\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\kappa_i}\left(z_i - \sum_{j<i} b_{i,j} z_j\right)^2 + \frac{1}{2\kappa_i}\left(0 - \sum_{j<i} b_{i,j} z_j\right)^2\right) \\
&= \prod_{i=1}^{n} \exp\left(-\frac{1}{2\kappa_i}\left(z_i^2 - 2z_i \sum_{j<i} b_{i,j} z_j\right)\right) \text{pr}_Z(z^* = 0) \\
&= \exp\left(-\sum_i \frac{z_i^2}{2\kappa_i} + \frac{1}{2} 2 \sum_i \sum_{j<i} \frac{b_{i,j}}{\kappa_i} z_i z_j\right) \text{pr}_Z(z^* = 0) \\
&= \exp\left(-\frac{1}{2} z^\top K^{-1} z + \frac{1}{2} z^\top K^{-1} B z\right) \text{pr}_Z(z^* = 0) = \exp\left(-\frac{1}{2} z^\top \left[K^{-1}(I - B)\right] z\right) \text{pr}_Z(z^* = 0)
\end{aligned}$$

$$= \text{N}\left(z|0, (I - B)^{-1} K\right) \tag{3.3}$$

Recovering the mean from (3.3), it is

$$\text{pr}_Z(z) = \text{N}\left(z - \mu|0, (I - B)^{-1} K\right) = \text{N}\left(z|\mu, (I - B)^{-1} K\right)$$

□

Note 30. When CAR is used for modeling, B is often specified to be sparse either due to some natural problem specific property, or for our computational convenience as is may allow the use of sparse solvers. To achieve this, one way is to specify $B = \phi N$ where $\phi > 0$ and N is an adjacency matrix; that is $[B]_{i,j} = \phi 1(i \sim j) 1(i \neq j)$ will be non-zero only for adjacent pairs i and j.

Note 31. The system in (3.2) can be rewritten as

$$Z = \mu + B(Z-\mu) + E \Longleftrightarrow E = (I-B)(Z-\mu) \tag{3.4}$$

by setting $E = (I-B)(Z-\mu)$. The distribution of Z in (3.2) induces a distribution on E as $E \sim \mathrm{N}\left(0, K(I-B)^{\top}\right)$ because

$$\mathrm{E}(E) = \mathrm{E}((I-B)(Z-\mu)) = (I-B)\,\mathrm{E}(Z-\mu) = 0$$
$$\mathrm{Var}(E) = \mathrm{Var}((I-B)Z) = (I-B)\,\mathrm{Var}(Z)\,(I-B)^{\top} = (I-B)(I-B)^{-1}K(I-B)^{\top}$$

3.2. **Simultaneus Autoregressive (SAR) models.**

3.2.1. *Gaussian SAR.*

Note 32. CAR sets the AR relation, and specifies the distribution on Z which induces the distribution on E; see 3.4. SAR does does the reverse; sets the same AR relation but it specifies the distribution on E which induces the distribution on Z –this is more might be more intuitive (?).

Definition 33. Consider discrete set of sites $\mathcal{S} = \{s_i; i = 1, ..., n\}$. Consider a random field $Z = (Z_s; s \in \mathcal{S})^{\top} = (Z_i = Z(s_i); i = 1, ...n)^{\top}$ on the discrete set of indexes $\mathcal{S}$ with values in $\mathcal{Z}$. Define

$$Z = \mu + \tilde{B}(Z-\mu) + E \Longleftrightarrow E = \left(I - \tilde{B}\right)(Z-\mu)$$

Assume that matrix $\tilde{B}$ is such that $\left(I-\tilde{B}\right)^{-1}$ exists, and $\left[\tilde{B}\right]_{i,i} = 0$. Assume that $E = (E_i; i = 1, ...n)$ is an n-dimensional Gaussian random vector $E \sim \mathrm{N}_n(0, \Lambda)$ with $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ whose elements are indexed by $\mathcal{S}$. Then we say that Z follows a Gaussian Simultaneous Autoregressive (SAR) model.

Proposition 34. *The joint distribution of Z following the SAR model in Def 33 is*

$$Z \sim N\left(\mu, \left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}\right) \tag{3.5}$$

Proof. Z is a linear combination of Gaussians, hence if follows a Gaussian distribution. It's mean and variance are

$$\mathrm{E}(Z) = \mathrm{E}\left(\left(I-\tilde{B}\right)^{-1}E + \mu\right) = \mu,$$
$$\mathrm{Var}(Z) = \mathrm{Var}\left(\left(I-\tilde{B}\right)^{-1}E + \mu\right) = \left(I-\tilde{B}\right)^{-1}\mathrm{Var}(E)\left(I-\tilde{B}^{\top}\right)^{-1} = \left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}$$

□

3.3. **CAR vs SAR.**

Remark 35. From (3.2) and (3.5), CAR and SAR are equivalent iff

$$\underbrace{(I-B)^{-1}K}_{\text{CAR}} = \underbrace{\left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}}_{\text{SAR}}$$

Proposition 36. *Any SAR can be written as a CAR model.*

Proof. Let Λ be $n\times n$ positive diagonal matrix. Let $\tilde{B}$ be $n\times n$ positive matrix where $I-\tilde{B}$ is non-singular and $\tilde{B}_{i,i} := \left[\tilde{B}\right]_{i,i} = 0$. Then $\left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1}$ is well defined and I need to solve wrt B and $K=\operatorname{diag}(\kappa_1,...,\kappa_n)$

$$\begin{aligned}
(I-B)^{-1}K = \left(I-\tilde{B}\right)^{-1}\Lambda\left(I-\tilde{B}^{\top}\right)^{-1} &\Leftrightarrow \\
K^{-1}(I-B) = \left(I-\tilde{B}^{\top}\right)\Lambda^{-1}\left(I-\tilde{B}\right) &\Leftrightarrow \\
K^{-1}-K^{-1}B = \Lambda^{-1}-\tilde{B}^{\top}\Lambda^{-1}-\Lambda^{-1}\tilde{B}+\tilde{B}^{\top}\Lambda^{-1}\tilde{B} &
\end{aligned}$$

If I focus of the diagonal part and set $B_{i,i} := [B]_{i,i} = 0$

$$\left[K^{-1}\right]_{i,i} - \cancelto{=0}{\left[K^{-1}B\right]_{i,i}} = \left[\Lambda^{-1}\right]_{i,i} - \cancelto{=0}{\left[\tilde{B}^{\top}\Lambda^{-1}\right]_{i,i}} - \cancelto{=0}{\left[\Lambda^{-1}\tilde{B}\right]_{i,i}} + \left[\tilde{B}^{\top}\Lambda^{-1}\tilde{B}\right]_{i,i}$$

so

$$\kappa_i = \left(\frac{1}{\lambda_i}+\sum_{j=1}^{n}\frac{\tilde{B}_{j,i}^2}{\lambda_j}\right)^{-1} > 0,\ \forall i=1,...,n$$

and hence I can solve with respect to K and Bin a manner that they satisfy the assumptions of CAR. □

Remark 37. The converse of Prop 36 is not true.

Proposition 38. *Any positive-definite covariance matrix Σ can be expressed as the covariance matrix of a CAR model $\Sigma_{CAR} = (I-B)^{-1}K$, for a unique pair of matrices B and K where $(I-B)$ is non-singular and K is diagonal.*

Proof. (This proof can be considered as an exercise for understanding CAR) Express

$$\Sigma^{-1} = D - R$$

for

$$[D]_{i,j} = \begin{cases}\left[\Sigma^{-1}\right]_{i,i} & i=j\\ 0 & i\neq j\end{cases},\ \text{and}\ [R]_{i,j} = \begin{cases}0 & i=j\\ -\left[\Sigma^{-1}\right]_{i,j} & i\neq j\end{cases}$$

then

$$\Sigma = (D-R)^{-1} = \left(D\left(I-D^{-1}R\right)\right)^{-1} = \left(I-D^{-1}R\right)^{-1}D^{-1}$$

Now define $B = D^{-1}R$ and $K = D^{-1}$, and you get $\Sigma = \Sigma_{\text{CAR}}$. Now regarding the uniqueness, assume there is another pair of $\mathring{B}$, and $\mathring{K}$ such that $\Sigma_{\text{CAR}} = \left(I - \mathring{B}\right)^{-1} \mathring{K}$. Then

$$\operatorname{diag}\left(\Sigma^{-1}\right) = \operatorname{diag}\left(\Sigma_{\text{CAR}}^{-1}\right) = \operatorname{diag}\left(\mathring{K}^{-1}\left(I - \mathring{B}\right)\right) = \operatorname{diag}\left(\mathring{K}^{-1}\right)$$

and similarly $\operatorname{diag}\left(\Sigma^{-1}\right) = \operatorname{diag}\left(K^{-1}\right)$. Hence it has to be $\mathring{K} = K$ because both are diagonal matrices. Then it is

$$\left(I - \mathring{B}\right)^{-1} \mathring{K} = (I - B)^{-1} K \overset{\mathring{K}=K}{\iff} \mathring{B} = B.$$

So the representation is unique. □

Proposition 39. *Any positive-definite covariance matrix Σ can be expressed as the covariance matrix of a SAR model $\Sigma_{SAR} = \left(I - \tilde{B}\right)^{-1} \Lambda \left(I - \tilde{B}^{\top}\right)^{-1}$ for a (non-unique) pair of matrices $\tilde{B}$ and Λ where $\left(I - \tilde{B}\right)$ is non-singular, $\left[\tilde{B}\right]_{i,i} = 0$, and Λ is diagonal.*

Proof. (This proof can be considered as an exercise for understanding SAR) Express

$$\Sigma^{-1} = LL^{\top}$$

where L is a lower triangular matrix with $[L]_{i,i} > 0$. Such matrix decomposition can be done by Cholesky decomposition, square-matrix decomposition, etc... and hence it is not always unique. Then

$$\Sigma = \left(LL^{\top}\right)^{-1} = L^{-\top}L^{-1}$$

Now express, $L = D - C$ for

$$[D]_{i,j} = \begin{cases} [L]_{i,i} & i = j \\ 0 & i \neq j \end{cases}, \text{ and } [C]_{i,j} = \begin{cases} 0 & i = j \\ -[L]_{i,j} & i \neq j \end{cases}$$

then

$$\begin{aligned}\Sigma &= (D - C)^{-\top}(D - C)^{-1} = \left(I - D^{-1}C\right)^{-\top} D^{-\top}D^{-1}\left(I - D^{-1}C\right)^{-1} \\ &= \left(I - C^{\top}D^{-\top}\right)^{-1} D^{-\top}D^{-1}\left(I - \left(C^{\top}D^{-\top}\right)^{\top}\right)^{-1}\end{aligned}$$

Set $\tilde{B} = C^{\top}D^{-\top}$ and $\Lambda = D^{-\top}D^{-1}$ and you get $\Sigma_{\text{SAR}} = \Sigma$ for non-unique pairs of $\tilde{B}$ and Λ. □

Proposition 40. *Any SAR model can be written as a unique CAR model.*

Proof. (This proof can be considered as an exercise for understanding CAR/sar) SAR and CAR are both Gaussian's with the same mean. SAR's variance matrix is positive definite, and hence it can be written in a unique manner as a CAR's variance matrix by Prop 38. □

Proposition 41. *Any CAR model can be written as a non-unique SAR model.*

Proof. SAR and CAR are both Gaussian's with the same mean. CAR's variance matrix is positive definite, and hence it can be written in a non-unique manner as a SAR's variance matrix by Prop 39. $\square$

Example 42. Show that

(1) Z_i and E_j are independent for $i \neq j$ in Gaussian CAR
(2) Z_i and E_j are not necessarily independent for $i \neq j$ in Gaussian SAR

Solution.

(1) For Gaussian CAR,

$$\operatorname{Cov}(E, Z) = \operatorname{Cov}((I - B)Z, Z) = (I - B)\operatorname{Var}(Z) = (I - B)(I - B)^{-1} K = K$$

which is a diagonal; hence Z_i and E_j are independent for $i \neq j$.

(2) For Gaussian SAR,

$$\operatorname{Cov}(Z, E) = \operatorname{Cov}\left(\left(I - \tilde{B}\right)^{-1} E, E\right) = \left(I - \tilde{B}\right)^{-1} \operatorname{Var}(E) = \left(I - \tilde{B}\right)^{-1} \Lambda$$

which is not a diagonal matrix in general; hence Z_i and E_j may be dependent for $i \neq j$.

4. Related Random Fields with particular properties

Note 43. We introduce general modeling structures of basic building models which are computationally convenient yet reasonable for use in spatial statistics models. Convenient because they aim to break a high-dimensional problem into smaller ones using conditional independence, and reasonable because they allow representation of spatial dependence as well. We introduce the Gibbs Random Fields and the Markov Random Fields. The Ising model, CAR, and SAR are just particular cases of such models.

4.1. **Gibbs Random Fields.**

Notation 44. Recall notation $z_{\mathcal{A}} = (z_i : i \in \mathcal{A})$ and $\mathcal{Z}^{\mathcal{A}} = \left\{z_{\mathcal{A}} : z \in \mathcal{Z}^{\mathcal{S}}\right\}$ for $\mathcal{A} \subseteq \mathcal{S}$.

Definition 45. Let $\mathcal{S} \neq \emptyset$ be a finite collection of sites. Let $\mathcal{Z} \subset \mathbb{R}$. Interaction potential is a family $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\}$ of potential functions $V_{\mathcal{A}} : \mathcal{Z}^{\mathcal{A}} \to \mathbb{R}$ such that $V_{\emptyset}(\cdot) := 0$ and for every set $\mathcal{A} \subseteq \mathcal{S}$ the sum

$$U_{\mathcal{A}}^{\mathcal{V}}(z) = \sum_{\{\forall \mathcal{B} \subseteq \mathcal{S} : \mathcal{A} \cap \mathcal{B} \neq \emptyset\}} V_{\mathcal{B}}(z_{\mathcal{B}}) \tag{4.1}$$

exists.

Definition 46. The function $V_{\mathcal{A}} : \mathcal{Z}^{\mathcal{A}} \to \mathbb{R}$ in Def 45 is called potential on $\mathcal{A}$.

Definition 47. The function $U_{\mathcal{A}}^{\mathcal{V}}(z)$ in (4.1) in Def 45 is called energy function of interaction potential $\mathcal{V}$ on $\mathcal{A}$ is called.

Definition 48. The interaction potential $\mathcal{V}$ is said to be admissible if for all $\mathcal{B} \subseteq \mathcal{S}$ and $z_{\mathcal{S}\backslash\mathcal{B}} \in \mathcal{Z}^{\mathcal{S}\backslash\mathcal{B}}$

$$C_{\mathcal{A}}^{\mathcal{V}}\left(z_{\mathcal{S}\backslash\mathcal{A}}\right) = \int \exp\left(U_{\mathcal{A}}^{\mathcal{V}}\left(\left(z_{\mathcal{A}}, z_{\mathcal{S}\backslash\mathcal{A}}\right)\right)\right) \mathrm{d}z_{\mathcal{A}} < \infty$$

Note 49. This allow as to define a distribution corresponding to the energy.

Definition 50. Let Z be $\mathcal{Z}$ valued Random Field on a finite collection of sites $\mathcal{S}$ with $\mathcal{S} \neq \emptyset$, and let $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\}$ be an interaction potential of functions $V_{\mathcal{A}} : \mathcal{Z}^{\mathcal{A}} \to \mathbb{R}$. Assume that $\mathcal{V}$ is admissible. Then Z is a Gibbs Random Field with interaction potentials $\mathcal{V} = \{V_{\mathcal{A}} : \mathcal{A} \subseteq \mathcal{S}\}$ if

$$\mathrm{pr}_Z\left(z_{\mathcal{A}} | z_{\mathcal{S}\backslash\mathcal{A}}\right) = \frac{1}{C_{\mathcal{A}}^{\mathcal{V}}\left(z_{\mathcal{S}\backslash\mathcal{A}}\right)} \exp\left(\underbrace{\sum_{\{\mathcal{B}\subseteq\mathcal{S}\,:\,\mathcal{A}\cap\mathcal{B}\neq\emptyset\}} V_{\mathcal{B}}\left(z_{\mathcal{B}}\right)}_{=U_{\mathcal{A}}^{\mathcal{V}}(z)}\right), \quad z \in \mathcal{Z}^{\mathcal{S}} \tag{4.2}$$

Definition 51. The normalizing integral $C_{\mathcal{A}}^{\mathcal{V}}$ in (4.2) is called partition function.

Notation 52. Obviously for the marginal $\mathrm{pr}_Z\left(z_{\mathcal{S}}\right)$ we will denote for $z \in \mathcal{Z}^{\mathcal{S}}$

$$\mathrm{pr}_Z\left(z_{\mathcal{S}}\right) = \frac{1}{C_{\mathcal{S}}^{\mathcal{V}}} \exp\left(U_{\mathcal{S}}^{\mathcal{V}}(z)\right) = \frac{1}{C_{\mathcal{S}}^{\mathcal{V}}} \exp\left(\sum_{\mathcal{B}\subseteq\mathcal{S}} V_{\mathcal{B}}\left(z_{\mathcal{B}}\right)\right)$$

where $C_{\mathcal{S}}^{\mathcal{V}} < \infty$ is the constant. For easy of the notation, in this case, we can omit $\cdot_{\mathcal{S}}^{\mathcal{V}}$ and just write

$$\mathrm{pr}_Z\left(z_{\mathcal{S}}\right) = \frac{1}{C} \exp\left(\sum_{\mathcal{B}\subseteq\mathcal{S}} V_{\mathcal{B}}\left(z_{\mathcal{B}}\right)\right), \ z \in \mathcal{Z}^{\mathcal{S}}$$

Example 53. (Ising model) In Ex 12, the Ising model has non-zero potentials

$$\begin{aligned}
V_{\emptyset}(z), &= 0 \\
V_{\{i\}}(z), &= \alpha z_i \,\forall i \in \mathcal{S} \\
V_{\{i,j\}}(z) &= \begin{cases} \beta z_i z_j & \text{if } i \sim j \\ 0 & \text{if } i \not\sim j \end{cases} \\
V_{\mathcal{A}}(z) &= 0, \text{ if } \operatorname{card}(\mathcal{A}) > 2
\end{aligned}$$

it has energy function

$$U(z) := U_{\mathcal{S}}^{\mathcal{V}}(z_{\mathcal{S}}) = \alpha \sum_{i\in\mathcal{S}} z_i + \beta \sum_{\{i\in\mathcal{S}, j\in\mathcal{S}: i\sim j\}} z_i z_j$$

and energy function conditional on $\mathcal{S}\backslash\mathcal{B}$

$$U_{\mathcal{B}}^{\mathcal{V}}\left(z_{\mathcal{B}}|z_{\mathcal{S}\backslash\mathcal{B}}\right) = \alpha \sum_{i\in\mathcal{B}} z_i + \beta \sum_{\{i\in\mathcal{B}, j\in\mathcal{S}: i\sim j\}} z_i z_j$$

Identifiability of the potential.

Definition 54. The interaction potential $\mathcal{V}$ is said to be normalized with respect to $\zeta \in \mathcal{Z}$ if there is $i \in \mathcal{S}$ which for any for any $z \in \mathcal{Z}^{\mathcal{S}}$with $z_i = \zeta$ implies that $V_{\mathcal{B}}(z) = 0$.

Note 55. The mapping $\mathcal{V} \to \mathrm{pr}_Z$ in (4.2) is non-identifiable as 4.2 can be constructed from a different interaction potential $\tilde{\mathcal{V}} = \{V_{\mathcal{B}} + c : \mathcal{B} \subseteq \mathcal{S}\}$ for any constant c. I.e. $U_{\mathcal{S}}^{\mathcal{V}}(z) = U_{\mathcal{S}}^{\tilde{\mathcal{V}}}(z)$.

Note 56. One way to make $\mathcal{V}$ identifiable is to impose restriction

$$\forall \mathcal{A} \neq \emptyset,\ V_{\mathcal{A}}(z) = 0,\ \text{if for some}\ \ i \in \mathcal{A},\ z_i = \zeta \tag{4.3}$$

This follows from the following theorem which uniquely associates potentials satisfying (4.3) with (4.2).

Notation 57. For convenience, consider notation related to $z^{[\mathcal{B},\zeta]}$ such as

$$\left[z^{[\mathcal{B},\zeta]}\right]_i = \begin{cases} \zeta, & \text{if } i \notin \mathcal{B} \\ z_i, & \text{if } i \in \mathcal{B} \end{cases}$$

and $z_{\mathcal{A}}^{[\mathcal{B},\zeta]} = \left(z_s^{[\mathcal{B},\zeta]}; s \in \mathcal{A}\right)$, and $z_s^{[\mathcal{B},\zeta]} = z_{\{s\}}^{[\mathcal{B},\zeta]}$ for some fixed ζ.

Example 58. For instance if $z \in \mathcal{Z}^{\mathcal{S}}$ where $\mathcal{S} = \{1, ..., n\}$ then

$$z^{[\emptyset,\zeta]} = \Big(\overbrace{\zeta, ..., \zeta}^{n\text{ times}}\Big)^{\top}; \qquad z^{[\{i\},\zeta]} = \Big(\zeta, ..., \zeta, \overbrace{z_i}^{\substack{i\text{th location}\\ \downarrow}}, \zeta, ..., \zeta\Big)^{\top};$$

$$z^{[\{i,j\},\zeta]} = \Big(\zeta, ...\zeta, \overbrace{z_i}^{\substack{i\text{th location}\\ \downarrow}}, \zeta, ..., \zeta, \overbrace{z_j}^{\substack{j\text{th location}\\ \downarrow}}, ..., \zeta\Big)^{\top}; \qquad z^{[\mathcal{S},\zeta]} = (z_1, ..., z_n)^{\top};$$

Theorem 59. *Let Z be an $\mathcal{Z}$-valued random field on a finite collection $\mathcal{S} \neq \emptyset$ of sites such that $pr_Z(z) > 0$ for all $z \in \mathcal{Z}^{\mathcal{S}}$. Then Z is a Gibbs Random Field with respect to the* *canonical potential*

$$V_{\mathcal{A}}\left(z_{\mathcal{A}}\right)=\sum_{\mathcal{B}\subseteq\mathcal{A}}\left(-1\right)^{Card(\mathcal{A}\backslash\mathcal{B})}U_{\mathcal{B}}^{\mathcal{V}}\left(z^{[\mathcal{B},\zeta]}\right),\ z\in\mathcal{Z}^{\mathcal{S}}$$
$$=\sum_{\mathcal{B}\subseteq\mathcal{A}}\left(-1\right)^{Card(\mathcal{A}\backslash\mathcal{B})}\log\left(pr_{Z}\left(z^{[\mathcal{B},\zeta]}\right)\right),\ z\in\mathcal{Z}^{\mathcal{S}} \tag{4.4}$$

where $\zeta\in\mathcal{Z}$ is a fixed value and notation $z^{[\mathcal{B},\zeta]}$ denotes the vector based on $z\in\mathcal{Z}^{\mathcal{S}}$ but modified such that its i-th element is $\left[z^{[\mathcal{B},\zeta]}\right]_i=z_i$ if $i\in\mathcal{B}$ and $\left[z^{[\mathcal{B},\zeta]}\right]_i=\zeta$ if $i\notin\mathcal{B}$. This is the unique normalized potential w.r.t $\zeta\in\mathcal{Z}$.

Proof. The proof is is based on Möbius inversion formula[2], and hence out of scope. □

Corollary 60. *From Thm 59, for all $i\in\mathcal{A}$ it is*

$$V_{\mathcal{A}}\left(z_{\mathcal{A}}\right)\sum_{\mathcal{B}\subseteq\mathcal{A}}\left(-1\right)^{Card(\mathcal{A}\backslash\mathcal{B})}\log\left(pr_{Z}\left(z_i^{[\mathcal{B},\zeta]}|z_{\mathcal{S}\backslash\{i\}}^{[\mathcal{B},\zeta]}\right)\right),\ z\in\mathcal{Z}^{\mathcal{S}} \tag{4.5}$$

Note 61. The following example explains the use of Thm 59 regarding the Def 45.

Example 62. Consider $\mathcal{S}=\{1,2\}$. Let $z=(z_1,z_2)^\top$. Consider a fixed $\zeta\in\mathcal{Z}$. Then $\mathcal{V}=\{V_{\mathcal{A}}:\mathcal{A}\subseteq\mathcal{S}\}=\left\{V_{\{1\}},V_{\{2\}},V_{\{1,2\}}\right\}$. The decomposition of the energy $U\left(z=(z_1,z_2)^\top\right):=U_{\mathcal{S}}^{\mathcal{V}}\left(z\right)$ is written as (from (4.1))

$$U\left(z_1,z_2\right)-U\left(\zeta,\zeta\right)=V_{\{1\}}\left(z_1\right)+V_{\{2\}}\left(z_2\right)+V_{\{1,2\}}\left(z_1,z_2\right)$$

where (from (4.4)) it is

$$V_{\{1\}}\left(z_1\right)=U\left(z_1,\zeta\right)-U\left(\zeta,\zeta\right)$$
$$V_{\{2\}}\left(z_2\right)=U\left(\zeta,z_2\right)-U\left(\zeta,\zeta\right)$$
$$V_{\{1,2\}}\left(z_1,z_2\right)=U\left(z_1,z_2\right)-U\left(z_1,\zeta\right)-U\left(\zeta,z_2\right)+U\left(\zeta,\zeta\right)$$

[2]Rota, G. C. (1964). On the foundations of combinatorial theory: I. Theory of Möbius functions. In Classic Papers in Combinatorics (pp. 332-360). Boston, MA: Birkhäuser Boston.

Example 63. (Ising model) Revisiting Ex 12, w.r.t Theorem 59. Consider $\zeta = 0$. Note that we use Notation 57, for instance,

$$z^{[\emptyset,\zeta]} = \left(\overbrace{\zeta, ..., \zeta}^{n\,\text{times}} \right)^{\top};$$

$$z^{[\{i\},\zeta]} = \left(\zeta, ..., \zeta, \overbrace{z_i}^{\overset{i\text{th location}}{\downarrow}}, \zeta, ..., \zeta \right)^{\top};$$

$$z^{[\{i,j\},\zeta]} = \left(\zeta, ...\zeta, \overbrace{z_i}^{\overset{i\text{th location}}{\downarrow}}, \zeta, ..., \zeta, \overbrace{z_j}^{\overset{j\text{th location}}{\downarrow}}, ..., \zeta \right)^{\top}$$

By using Thm 59, $V_\emptyset = 0$, and for any $i \in \mathcal{S}$, it is

$$V_{\{i\}}(z), = (-1)^{1-1} U\left(z^{[\{i\},\zeta]}\right) + (-1)^{1-0} U\left(z^{[\emptyset,\zeta]}\right) = a z_i$$

for any $i, j \in \mathcal{S}$, with $i \sim j$ it is

$$\begin{aligned} V_{\{i,j\}}(z) =& \left[(-1)^{2-2} U\left(z^{[\{i,j\},\zeta]}\right)\right] + \left[(-1)^{2-1} U\left(z^{[\{i\},\zeta]}\right)\right] \\ &+ \left[(-1)^{2-1} U\left(z^{[\{j\},\zeta]}\right)\right] + \left[(-1)^{2-0} U\left(z^{[\emptyset,\zeta]}\right)\right] \\ =& [\alpha z_i + \alpha z_j + \beta z_i z_j] + [-\alpha z_i] + [-\alpha z_j] + [0] = \beta z_i z_j \,. \end{aligned}$$

Obviously, for any $i, j \in \mathcal{S}$, with $i \not\sim j$ it is $V_{\{i,j\}}(z) = 0$, and for $\operatorname{card}(\mathcal{A}) > 2$ it is $V_{\mathcal{A}}(z) = 0$.

4.2. **Markov Random Fields.**

Note 64. Recall the Ising model whose sites are are equipped with a symmetric relation "$\sim$". It's potentials $V_{\mathcal{A}}$ are non-zero only when $\mathcal{A}$ is a pair of sites $\{i, j\}$ satisfying the relation $\sim$ or when $\mathcal{A}$ a singleton. Consequently, its local characteristics $\text{pr}_i\left(z_i | z_{\mathcal{S}\setminus\{i\}}\right)$ depend only on the values of the sites $j \in \mathcal{S}\setminus\{i\}$ that satisfy $\sim$.

Note 65. Regarding spatial modeling, $\sim$ can describe adjacent sites which is in accordance to "dogma" that *near things are more related than distant things.* Also it is computationally convenient for big data problems (large number of sites) as it introduces sparsity and allows specialized numerical algorithms to be implemented.

Note 66. Markov Random Fields constrain the problem such that the conditional distribution of the label at some site i given those at all other sites $j \in \mathcal{S} - \{i\}$ depends only on the labels at neighbors of site i.

Definition 67. We define as the boundary of $\mathcal{A}$, $\mathcal{A} \subseteq \mathcal{S}$, for a given relation $\sim$ the set

$$\partial\mathcal{A} = \{s \in \mathcal{S}\backslash\mathcal{A} : \exists t \in \mathcal{A} \,\text{s.t.}\, s \sim t\}$$

Definition 68. Let $\partial\mathcal{A}$ be the boundary of $\mathcal{A} \subseteq \mathcal{S}$ for a symmetric relation $\sim$ the finite set $\mathcal{S} \neq \emptyset$. Z is a random field on $\mathcal{S}$ taking values in $\mathcal{Z}$ with respect to the symmetric relation $\sim$ if for each $\mathcal{A} \subset \mathcal{S}$ and $Z_{\mathcal{A}\backslash\mathcal{S}} \in \mathcal{Z}_{\mathcal{A}\backslash\mathcal{S}}$the distribution of Z on $\mathcal{A}$ conditional on $Z_{\mathcal{A}\backslash\mathcal{S}}$ only depends on $Z_{\partial\mathcal{A}}$ (i.e. the configuration of Z on the neighborhood boundary of $\mathcal{A}$) i.e.

$$\mathrm{pr}_Z\left(z_{\mathcal{A}}|z_{\mathcal{S}\backslash\mathcal{A}}\right) = \mathrm{pr}_Z\left(z_{\mathcal{A}}|z_{\partial\mathcal{A}}\right) \tag{4.6}$$

when $\mathrm{pr}_Z\left(z_{\mathcal{S}\backslash\mathcal{A}}\right) > 0$

Note 69. Def 68 implies that (4.6) becomes

$$\mathrm{pr}_Z\left(z_i|z_{-i}\right) = \mathrm{pr}_Z\left(z_i|z_{\partial\{i\}}\right), \quad \forall i \in \mathcal{S} \tag{4.7}$$

when $\mathrm{pr}_Z\left(z_{\mathcal{S}\backslash\{i\}}\right) > 0$

Definition 70. A non-empty subset $\mathcal{C}$, $\mathcal{C} \subset \mathcal{S}$, is a clique in $\mathcal{S}$ with respect to $\sim$ if for all $s, t \in \mathcal{C}$ with $s \neq t$ it is $s \sim t$ or if $\mathcal{C}$ is a singleton set.

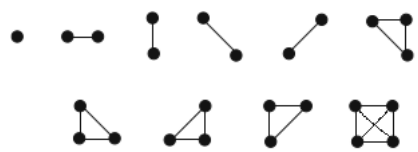


FIGURE 4.1. Examples of cliques

Note 71. The following theorem shows that the distribution of any Markov random field such that $\mathrm{pr}_Z(z)$ is positive can be expressed in terms of interactions between neighbors.

Theorem 72. *(Hammersley–Clifford) Let Z be an $\mathcal{Z}$-valued random field on a finite collection $\mathcal{S} \neq \emptyset$ of sites such that $pr_Z\left(z_{\mathcal{A}}|z_{\mathcal{C}\backslash\mathcal{A}}\right) > 0$ for all $\mathcal{A} \subset \mathcal{S}$ and $z \in \mathcal{Z}^{\mathcal{S}}$. Let $\sim$ be a symmetric relation on $\mathcal{S}$. Then Z is a Markov Random Field with respect to $\sim$ if and only if*

$$pr_Z(z) = \prod_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} \varphi_{\mathcal{C}}(z_{\mathcal{C}}) \tag{4.8}$$

for some interaction functions $\varphi_{\mathcal{C}} : \mathcal{Z}^{\mathcal{C}} \to \mathbb{R}^{+}$ defined on cliques $\mathcal{C} \in \boldsymbol{\mathcal{C}}$.

Proof. □

For convenience, let $\left[z^{\mathcal{B},\delta}\right]_i = \begin{cases} \delta, & \text{if } i \notin \mathcal{B} \\ z_i, & \text{if } i \in \mathcal{B} \end{cases}$, *and* $z_{\mathcal{A}}^{\mathcal{B},\delta} = \left(z_s^{\mathcal{B},\delta}; s \in \mathcal{A}\right)$, *and* $z_s^{\mathcal{B},\delta} = z_{\{s\}}^{\mathcal{B},\delta}$.

for $\Longrightarrow$**:** *By Thm 59, Z is Gibbs with a canonical potential (4.4)*

$$V_{\mathcal{A}}\left(z_{\mathcal{A}}\right)=\sum_{\mathcal{A}\subseteq\mathcal{B}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\mathcal{B}\right)}\log\left(pr_{Z}\left(z^{[\mathcal{B},\zeta]}\right)\right),$$

for $z\in\mathcal{Z}^{\mathcal{S}}$. We need to show that for all $\mathcal{A}$ which are not a cliques, $\mathcal{A}\notin\boldsymbol{\mathcal{C}}$.

Assume a set $\mathcal{A}$ with $\mathcal{A}\subseteq\mathcal{S}$ which is not a clique, $\mathcal{A}\notin\boldsymbol{\mathcal{C}}$, there are two distinct sites $s,t\in\mathcal{A}$ with $s\not\sim t$. Then,

$$\begin{aligned}V_{\mathcal{A}}\left(z\right)=&\sum_{\mathcal{B}\subseteq\mathcal{A}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\mathcal{B}\right)}\log\left(pr_{Z}\left(z_{s}^{\mathcal{B},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B},\delta}\right)\right)\\
=&\sum_{\mathcal{B}\subseteq\mathcal{A}\backslash\{s,t\}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\mathcal{B}\right)}\log\left(pr_{Z}\left(z_{s}^{\mathcal{B},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B},\delta}\right)\right)\\
&+\sum_{\mathcal{B}\subseteq\mathcal{A}\backslash\{s,t\}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\left(\mathcal{B}\cup\{s\}\right)\right)}\log\left(pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{s\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{s\},\delta}\right)\right)\\
&+\sum_{\mathcal{B}\subseteq\mathcal{A}\backslash\{s,t\}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\left(\mathcal{B}\cup\{t\}\right)\right)}\log\left(pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{t\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{t\},\delta}\right)\right)\\
&+\sum_{\mathcal{B}\subseteq\mathcal{A}\backslash\{s,t\}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\left(\mathcal{B}\cup\{s,t\}\right)\right)}\log\left(pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{s,t\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{s,t\},\delta}\right)\right)\end{aligned}$$

Rearranging I get

$$V_{\mathcal{A}}\left(z\right)=\sum_{\mathcal{B}\subseteq\mathcal{A}}\left(-1\right)^{Card\left(\mathcal{A}\backslash\mathcal{B}\right)}\log\left(\frac{pr_{Z}\left(z_{s}^{\mathcal{B},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B},\delta}\right)}{pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{t\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{t\},\delta}\right)}\frac{pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{s,t\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{s,t\},\delta}\right)}{pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{s\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{s\},\delta}\right)}\right)$$

Because $s\not\sim t$, it is $pr_{Z}\left(z_{s}^{\mathcal{B},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B},\delta}\right)=pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{t\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{t\},\delta}\right)$ and $pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{s,t\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{s,t\},\delta}\right)=pr_{Z}\left(z_{s}^{\mathcal{B}\cup\{s\},\delta}|z_{\mathcal{S}\backslash s}^{\mathcal{B}\cup\{s\},\delta}\right)$. This implies $V_{\mathcal{A}}\left(z\right)=0$ for any subset $\mathcal{A}$ with $\mathcal{A}\subseteq\mathcal{S}$ which is not a clique. Hence (4.8) holds.

for $\Longleftarrow$**:** *By using (4.2), I can write*

$$pr_{Z}\left(z_{\mathcal{A}}|z_{\mathcal{S}\backslash\mathcal{A}}\right)=\frac{1}{C_{\mathcal{A}}\left(z_{\mathcal{S}\backslash\mathcal{A}}\right)}\exp\left(U_{\mathcal{A}}\left(z\right)\right)$$

where

$$U_{\mathcal{A}}\left(z\right)=\sum_{\{\mathcal{C}\subseteq\mathcal{S}\,:\,\mathcal{A}\cap\mathcal{C}\neq\emptyset\}}V_{\mathcal{C}}\left(z_{\mathcal{C}}\right)$$

depends only on $\{z_i:i\in\mathcal{A}\cup\partial\mathcal{A}\}$ as $pr_{Z}\left(\cdot\right)$ is a Markov Random Field.

Note 73. Because $\mathrm{pr}_{Z}\left(z\right)>0$, the Markov Random Field in (4.8) is a Gibbs Random Field as

$$\mathrm{pr}_{Z}\left(z\right)=\exp\left(\sum_{\mathcal{C}\in\boldsymbol{\mathcal{C}}}\log\left(\varphi_{\mathcal{C}}\left(z_{\mathcal{C}}\right)\right)\right)$$

with non-zero interaction potentials restricted to cliques $\mathcal{C} \in \boldsymbol{\mathcal{C}}$.

Note 74. Essentially Thm 72, says that:

for $\Longrightarrow$: we need to show that there exists an interaction potential $\boldsymbol{\varphi} = \{\varphi_{\mathcal{C}} : \mathcal{C} \in \boldsymbol{\mathcal{C}}\}$ defined on the cliques $\boldsymbol{\mathcal{C}}$ such that $\mathrm{pr}_Z(\cdot)$ is a Gibbs Random Field with iteration potential $\boldsymbol{\varphi}$.

for $\Longleftarrow$: a Gibbs Random Field with potentials $\{\varphi_{\mathcal{C}} : \mathcal{C} \in \boldsymbol{\mathcal{C}}\}$ defined on the cliques $\boldsymbol{\mathcal{C}}$ is a Markov Random Field.

Example 75. (Ising model) Revisiting Ex 12... The joint PMF is

$$\begin{aligned}\mathrm{pr}(z) =& \frac{\exp\left(\alpha \sum_{i\in\mathcal{S}} z_i + \beta \sum_{\{i,j\}:i\sim j} z_i z_j\right)}{\sum_{z\in\mathcal{Z}^{\mathcal{S}}} \exp\left(\alpha \sum_{i\in\mathcal{S}} z_i + \beta \sum_{\{i,j\}:i\sim j} z_i z_j\right)} \\ =& \frac{1}{\sum_{z\in\mathcal{Z}^{\mathcal{S}}} \exp\left(\alpha \sum_{i\in\mathcal{S}} z_i + \beta \sum_{\{i,j\}:i\sim j} z_i z_j\right)} \prod_{i\in\mathcal{S}} \exp(\alpha z_i) \prod_{i\in\mathcal{S}} \prod_{j:j\sim i} \exp(\beta z_i z_j)\end{aligned}$$

I can find that

$$\begin{aligned}\varphi_{\emptyset} &= 1/\sum_{z\in\mathcal{Z}^{\mathcal{S}}} \exp\left(\alpha \sum_{i\in\mathcal{S}} z_i + \beta \sum_{\{i,j\}:i\sim j} z_i z_j\right) \\ \varphi_{\{i\}}\left(z_{\{i\}}\right) &= \exp(\alpha z_i), \qquad \forall i \in \mathcal{S} \\ \varphi_{\{i,j\}}\left(z_{\{i,j\}}\right) &= \exp(\beta z_i z_j), \qquad \forall i,j \in \mathcal{S} \quad \text{s.t.} \quad i \sim j \\ \varphi_{\{i,j\}}\left(z_{\{i,j\}}\right) &= 1, \qquad \forall i,j \in \mathcal{S} \quad \text{s.t.} \quad i \nsim j \\ \varphi_{\mathcal{A}}(z_{\mathcal{A}}) &= 1, \qquad \forall \mathcal{A} \subset \mathcal{S} \quad \text{s.t.} \quad \mathrm{card}(\mathcal{A}) > 2\end{aligned}$$

where $\{i\}$ and $\{i,j\}$ satisfying $i \sim j$ are cliques. Notice that $\varphi_{\emptyset}$ is just the constant term that can be absorbed to the other $\varphi's$ correspond to cliques.

Part 2. Model building for aerial data & related inference

5. Automodels

Note 76. We introduce a general class of models introducing spatial dependence which are associated to the exponential family of distributions: the automodels, and their special case Besag's automodels.

Recall

Definition 77. A random variable $X \in \mathcal{X}$ follows an exponential family (of distributions) labeled by parameter $\theta \in \Theta$ if the associated PMF/PDF $\mathrm{pr}_X(x|\theta)$ can be expressed in the form

$$\mathrm{pr}_X\left(x|\theta\right)=\exp\Big(A\left(\theta\right)^\top B\left(x\right)+C\left(x\right)+D\left(\theta\right)\Big)$$

where $A(\cdot)$, $B(\cdot)$, $C(\cdot)$, and $D(\cdot)$ are known functions.

5.1. **General multi-parameter automodels.**

Theorem 78. *Consider Markov random field Z that takes values in $\mathcal{Z}$ on a finite set of points $\mathcal{S}$ and has marginal probability*

$$pr_Z\left(z\right)=\frac{1}{C}\exp\left(U\left(z\right)\right),\quad z\in\mathcal{Z}^{\mathcal{S}}. \tag{5.1}$$

Consider some fixed normalization $\zeta=\left(\zeta,...,\zeta\right)^\top\in\mathcal{Z}^{\mathcal{S}}$. Assume that the following assumptions are satisfied:

(1) *In the energy function $U(\cdot)$ the dependence between the sites is pairwise only, i.e.*

$$U\left(z\right)=\sum_{i\in\mathcal{S}}V_i\left(z_i\right)+\sum_{\{\{i,j\}\in\mathcal{S}^2:i\sim j\}}V_{i,j}\left(z_i,z_j\right),\ z\in\mathcal{Z}_{\mathcal{S}}$$

with $V_i\left(\zeta\right)=V_{i,j}\left(z_i,\zeta\right)=V_{i,j}\left(\zeta,z_j\right)=0$.

(2) *For all $i\in\mathcal{S}$, the conditional distributions (characteristics) are exponential family distributions*

$$\log\left(pr_i\left(z_i|z_{-i}\right)\right)=\left(A_i\left(z_{-i}\right)\right)^\top B_i\left(z_i\right)+C_i\left(z_i\right)+D_i\left(z_{-i}\right), \tag{5.2}$$

where $A_i\left(z_{-i}\right)\in\mathbb{R}^\ell$, $B_i\left(z_i\right)\in\mathbb{R}^\ell$, for $\ell\geq 1$ and $C_i\left(z_i\right)\in\mathbb{R}$, and $D_i\left(z_{-i}\right)\in\mathbb{R}$ with $C_i\left(\zeta\right)=0$ and $D_i\left(\zeta,...,\zeta\right)=0$.

(3) *For all $i\in\mathcal{S}$, span$\left\{B_i\left(z_i\right);z_i\in\mathcal{Z}\right\}=\mathbb{R}^\ell$,* for $\ell\geq 1$.

Then, necessarily,

(1) *the functions $A_i\left(z_{-i}\right)\in\mathbb{R}^\ell$ take the form*

$$A_i\left(z_{-i}\right)=\alpha_i+\sum_{i\neq j}\beta_{i,j}B_j\left(z_j\right),\ i\in\mathcal{S}$$

where $\{\alpha_i;i\in\mathcal{S}\}$ is a family of ℓ-dimensional vectors, and $\{\beta_{i,j};i,j\in\mathcal{S},i\neq j\}$ is a family of $\ell\times\ell$ symmetric matrices, and

(2) *the potentials are given by*

$$V_i\left(z_i\right)=\left(\alpha_i\right)^\top B_i\left(z_i\right)+C_i\left(z_i\right) \tag{5.3}$$

$$V_{i,j}\left(z_i,z_j\right)=\left(B_i\left(z_i\right)\right)^\top\beta_{i,j}B_j\left(z_j\right) \tag{5.4}$$

Proof. Omitted, but can be found in

(1) Hardouin, C., & Yao, J. F. (2008). Multi-parameter automodels and their applications. Biometrika, 95(2), 335-349.

(2) Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 192-225.

$\square$

Definition 79. Automodel is called the model satisfying the assumptions of Theorem 78. It is called univariate automodel when $\ell = 1$, and multivariate automodel when $\ell \geq 1$.

Remark 80. In the univariate automodel, $\ell = 1$, assumption 3 in Theorem 78 is not needed; it is automatically satisfied as B_i's are not identically zero. Yet, (5.3) and (5.4) become

$$V_i\left(z_i\right) = \alpha_i B_i\left(z_i\right) + C_i\left(z_i\right) \tag{5.5}$$

$$V_{i,j}\left(z_i, z_j\right) = \beta_{i,j} B_i\left(z_i\right) B_j\left(z_j\right) \tag{5.6}$$

5.2. **Besag auto-models.**

Definition. Z follows a Besag's auto-model if Z is real-valued and its joint distribution $\mathrm{pr}_Z\left(z\right)$ is given by

$$\mathrm{pr}_Z\left(z\right) = \frac{1}{C}\exp\left(\sum_{i\in\mathcal{S}} V_i\left(z_i\right) + \sum_{\{i,j\}\in\mathcal{S}^2 : i\sim j} \beta_{i,j} z_i z_j\right),\ z \in \mathcal{Z}_{\mathcal{S}} \tag{5.7}$$

with $\beta_{i,j} = \beta_{j,i}$ for all $i, j \in \mathcal{S}$.

Note 81. The following allows us to define a Markov Random Field model from a a set of conditional distributions (characteristics) whose compatibility is automatically satisfied.

Proposition 82. *If each of the*

$$pr_i\left(z_i | z_{-i}\right), \quad for \quad i \in \mathcal{S}$$

is a family of real-valued $z_i \in \mathbb{R}$ conditional distributions which are members of the exponential family of distributions (5.2) with $B_i\left(z_i\right) = z_i$ for $i \in \mathcal{S}$, then they are compatible a Besag's auto-model with distribution (5.7) if $\beta_{i,j} = \beta_{j,i}$ for all $i, j \in \mathcal{S}$.

Proof. For

$$\mathrm{pr}_i\left(z_i | z_{-i}\right) = \exp\left(A_i\left(z_{-i}\right) z_i + C_i\left(z_i\right) + D_i\left(z_{-i}\right)\right)$$

it is

$$V_i\left(z_i\right) = \alpha_i B_i\left(z_i\right) + C_i\left(z_i\right) = \alpha_i z_i + C_i\left(z_i\right)$$

$$V_{i,j}\left(z_i, z_j\right) = \beta_{i,j} B_i\left(z_i\right) B_j\left(z_j\right) = \beta_{i,j} z_i z_j$$

so

$$\mathrm{pr}_Z\left(z\right) \propto \exp\left(\sum_i \left[\alpha_i z_i + C_i\left(z_i\right)\right] + \sum_{i\sim j} \beta_{i,j} z_i z_j\right),\ z \in \mathcal{Z}_{\mathcal{S}}$$

$\square$

Example 83. (Logistic auto-model / Ising model) Consider that $Z(s)$ represents success or failure at location $s \in \mathcal{S}$. Mathematically, assume random field Z taking values on a set of indices $\mathcal{S}$ in $\mathcal{Z} = \{0,1\}$ on $\mathcal{S} = \{1,...,n\}$, $n \in \mathbb{N} - \{0\}$.

Consider that for a given z_{-i} it is

$$z_i | z_{-i} \sim \text{Logistic}\left(\theta_i\left(z_{-i}\right)\right), \quad i \in \mathcal{S}.$$

Hint:: The PMF of Logistic distribution $x|\theta \sim \text{Logistic}(\theta)$ can be written as $\text{pr}(x|\theta) = (1 - \exp(x\theta))^{-1} 1(x \in \{0,1\})$.

Then the characteristics are

$$\text{pr}_i\left(z_i | z_{-i}\right) = \frac{\exp\left(z_i \theta_i\left(z_{-i}\right)\right)}{1 + \exp\left(\theta_i\left(z_{-i}\right)\right)} 1\left(z_i \in \{0,1\}\right) \tag{5.8}$$

Now, let's parameterize $\{\theta_i(\cdot)\}$ as

$$\theta_i\left(z_{-i}\right) = \alpha_i + \sum_{j:j\sim i} \beta_{i,j} z_j \tag{5.9}$$

for $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ with $\beta_{i,j} = \beta_{j,i}$. Then (5.8) becomes

$$\log\left(\text{pr}_i\left(z_i|z_{-i}\right)\right) = \underbrace{z_i}_{B_i(z_i)} \underbrace{\left(\alpha_i + \sum_{j\sim i} \beta_{i,j} \overbrace{z_j}^{B_i(z_j)}\right)}_{A_i(z_{-i})} + \underbrace{0}_{C_i(z_i)} + \underbrace{\left(-\log\left(1 + \exp\left(\alpha_i + \beta_{i,j} \sum_{j:j\sim i} \beta_{i,j} z_j\right)\right)\right)}_{D_i(z_{-i})}$$

Notice that all the conditionals $z_i | z_{-i}$ are Exponential family of distributions with

$$\begin{aligned}
A_i\left(z_{-i}\right) =& z_i \\
B_i\left(z_i\right) =& \alpha_i + \sum_{j\sim i} \beta_{i,j} B_i\left(z_j\right) \\
C_i\left(z_i\right) =& 0 \\
D_i\left(z_{-i}\right) =& -\log\left(1 + \exp\left(\alpha_i + \beta_{i,j} \sum_{j:j\sim i} \beta_{i,j} z_j\right)\right)
\end{aligned}$$

Also, I can get $C_i(\zeta) = 0$ and $D_i(\zeta, ..., \zeta) = 0$ by considering a reference point $\zeta = 0$. From Theorem 82, (5.8) with (5.9), the conditionals $z_i | z_{-i}$ are compatible as a Besag auto-model with marginal distribution

$$\mathrm{pr}_Z(z) \propto \exp\left(\overbrace{\underbrace{\sum_i \alpha_i \underbrace{z_i}_{B_i(z_i)}}_{V_i(z_i)} + \underbrace{\sum_i \sum_{j:j\sim i}}_{\sum_{\{i,j\}:j\sim i}} \beta_{i,j} z_i z_j}^{U(z)=}\right) \tag{5.10}$$

I observe that:

- Here the Ising model has spatially dependent coefficients $\{\alpha_i\}$ and $\{\beta_{i,j}\}$, unlike the Ising model in Example 12 where we considered $\{\alpha_i = \alpha\}$ and $\{\beta_{i,j} = \beta\}$.
- When $\beta_{i,j} = 0$, for all j such as $j \sim i$, it is $\mathrm{pr}_i(z_i|z_{-i}) = \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}$.
- Characteristic's present at site i is encouraged in neighboring site j when $\beta_{i,j} > 0$, and discouraged when $\beta_{i,j} < 0$.

Example 84. (Poisson auto-model) Consider that $Z(s)$ represents counts at location $s \in \mathcal{S}$. Mathematically we can consider Z taking values in $\mathcal{Z} = \mathbb{N}$ on a set of sites $\mathcal{S} = \{1, ..., n\}$, where $n \in \mathbb{N} - \{0\}$.

Consider that for a given z_{-i} it is

$$z_i|z_{-i} \sim \text{Poisson}(\lambda_i(z_{-i}))$$

Hint:: The PMF of Logistic distribution $x|\lambda \sim \text{Poisson}(\lambda)$ can be written as

$$\mathrm{pr}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)\, 1(x \in \mathbb{N})$$

with mean $\mathrm{E}(x|\lambda) = \lambda$.

Then the full conditionals (characteristics) are

$$\mathrm{pr}_i(z_i|z_{-i}) = \frac{1}{z_i!}(\lambda_i(z_{-i}))^{z_i} \exp(-\lambda)\, 1(z_i \in \mathbb{N}) \tag{5.11}$$

Now, let's parameterize $\{\lambda_i(\cdot)\}$ as

$$\log(\lambda_i(z_{-i})) = \alpha_i + \sum_{j:j\sim i} \beta_{i,j} z_j \tag{5.12}$$

for $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ with $\beta_{i,j} = \beta_{j,i}$. So (5.11) becomes

$$\log(\mathrm{pr}_i(z_i|z_{-i})) = \underbrace{\underbrace{z_i}_{B_i(z_i)}\left(\alpha_i + \sum_{j\sim i} \beta_{i,j} \overbrace{z_j}^{B_i(z_j)}\right)}_{A_i(z_{-i})} + \underbrace{\log(z_i!)}_{C_i(z_i)} + \underbrace{0}_{D_i(z_{-i})}$$

with

$$\begin{aligned}A_i\left(z_{-i}\right)=&z_i\\B_i\left(z_i\right)=&\alpha_i+\sum_{j\sim i}\beta_{i,j}B_i\left(z_j\right)\\C_i\left(z_i\right)=&\log\left(z_i!\right)\\D_i\left(z_{-i}\right)=&0\end{aligned}$$

I can notice that all the conditionals $z_i|z_{-i}$ are exponential of exponential distributions. Also, I can get $C_i(\zeta)=0$ and $D_i(\zeta,...,\zeta)=0$ by considering a reference point $\zeta=0$. From Theorem 82, (5.11) with (5.12), the the conditionals $z_i|z_{-i}$ are compatible as a Besag auto-model with marginal distribution

$$\mathrm{pr}_Z\left(z\right)\propto\exp\left(\overbrace{\sum_i\left(\underbrace{\alpha_i\overbrace{z_i}^{B_i(z_j)}+\overbrace{\log\left(z_i!\right)}^{C_i(z_i)}}_{V_i(z_i)}\right)+\sum_i\sum_{j:j\sim i}\beta_{i,j}z_iz_j}^{U(z)=}\right)$$

or otherwise the energy function is

$$U\left(z\right)=\sum_i\left(\alpha_iz_i+\log\left(z_i!\right)\right)+\sum_{j\sim i}\beta_{i,j}z_iz_j$$

Furthermore, to ensure that $U(z)$ is admissible, we need to consider additional conditions. I observe that

$$\sum_{z\in\mathbb{N}^{\mathcal{S}}}\exp\left(U\left(z\right)\right)=\sum_{z\in\mathbb{N}^{\mathcal{S}}}\prod_i\left(\exp\left(\alpha_iz_i+\log\left(z_i!\right)\right)+\sum_{j\sim i}\beta_{i,j}z_iz_j\right)$$

- If we use additional condition $\beta_{i,j}\le 0$ then

  $$\sum_{z\in\mathbb{N}^{\mathcal{S}}}\exp\left(U\left(z\right)\right)\le\sum_{z\in\mathbb{N}^{\mathcal{S}}}\prod_i\left(\exp\left(\alpha_iz_i+\log\left(z_i!\right)\right)\right)=\sum_{z\in\mathbb{N}^{\mathcal{S}}}\prod_i\frac{1}{z_i!}\exp\left(\alpha_iz_i\right)<\infty$$

  which converges. Modeling-wise, $\beta_{i,j}<0$ introduces competition among the neighbors similar to the Ising model. So by introducing a competition such as $\beta_{i,j}\le 0$ in the model I prevent the count z_i at i to explode.
- If $\beta_{i,j}>0$, I discourage competition among neighboring sites. Admissibility can be satisfied if we truncate the state space as $z_i<M$ for some fixed upper bound M. For instance, the characteristics $z_i|z_{-i}$ can follow a Poisson distribution truncated at M.

  $$\mathrm{pr}_i\left(z_i|z_{-i}\right)=\frac{1}{z_i!}\left(\lambda_i\left(z_{-i}\right)\right)^{z_i}\exp\left(-\lambda\right)1\left(z_i\in\{0,1,...,M\}\right)$$

So I can prevent z_i at i to explode by forcefully bounding it $z_i < M$ with a big enough value $M > 0$.

Note 85. CAR models is an auto-model as well. Consider that CAR model is defined as the one whose characteristics (full conditional distributions) are Normal distributions which is an exponential family of distributions. Hence the joint distribution of this Markov Random Field could have been derived from Theorem 78 as well.

5.3. Parametric estimation via MLE.

Maximum likelihood estimation.

Remark 86. Given a dataset $\{(s_i, Z_i = Z(s_i)); i = 1, ..., n\}$,estimation of the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ of a Besag auto-model, in the MLE framework, can be performed by maximizing the likelihood, as

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\mathrm{pr}_Z\left(Z|\{\alpha_i\}, \{\beta_{i,j}\}\right)\right) \tag{5.13}$$

$$\text{subject to } \beta_{i,j} = \beta_{j,i},\ \forall i, j \in \mathcal{S}$$

$$\text{...and any other problem specific restrictions}$$

where $\mathrm{pr}_Z(Z|\{\alpha_i\}, \{\beta_{i,j}\})$ is the joint distribution (5.7) given the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$.

Example 87. (Logistic auto-model / Ising model) Computing MLE $\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}$ of $\{\alpha_i\}, \{\beta_{i,j}\}$ requires

$$\begin{aligned}\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\log\left(\mathrm{pr}_Z\left(Z|\{\alpha_i\}, \{\beta_{i,j}\}\right)\right)\right) \\ &= \underset{\{\alpha_i\},\{\beta_{i,j}\}}{\arg\max} \left(\sum_i \alpha_i z_i + \sum_{\{i,j\}:j\sim i} \beta_{i,j} z_i z_j - \log\left(C\left(\{\alpha_i\}, \{\beta_{i,j}\}\right)\right)\right) \quad (5.14)\end{aligned}$$

where

$$C\left(\{\alpha_i\}, \{\beta_{i,j}\}\right) = \sum_{\forall z \in \mathcal{Z}} \exp\left(\sum_i \alpha_i z_i + \sum_{\{i,j\}:j\sim i} \beta_{i,j} z_i z_j\right) \tag{5.15}$$

is the normalizing constant.

Note 88. The optimization problem (5.13) can be too computationally expensive. For instance, in Example 87, a recursive optimization algorithm, like Newton-Raphson, requires several iterations. At each iteration the evaluation of the (parameter dependent) constant (5.15) has to be evaluated. A computation of that constant can be too expensive when the set of sites $i \in \mathcal{S}$ is large because the sum $\sum_{\forall z \in \mathcal{Z}}$ in (5.15) scans all the possible configurations of $z \in \mathcal{Z}$. A way to mitigate this is to use instead an "approximation" of the likelihood, such as the Pseudo-likelihood.

Pseudo maximum likelihood estimation.

Definition. The pseudo likelihood pseudo$L\left(Z;\theta\right)$ of observables $Z = \left(Z_1, ..., Z_n\right)^\top$ given parameters θ is an approximation of the (exact) likelihood $L\left(Z;\theta\right)$ of observables $Z = \left(Z_1, ..., Z_n\right)^\top$ given parameters θ which is equal to

$$\text{pseudo}L\left(Z;\theta\right) = \prod_i \text{pr}\left(Z_i | Z_{-i}, \theta\right)$$

where $\text{pr}\left(Z_i | Z_{-i}, \theta\right)$ are the conditionals of the joint pdf/pmf of the sampling distribution $\text{pr}\left(Z|\theta\right)$ of Z given parameter θ.

Definition. (Maximum PseudoLikelihood Estimator) The Maximum Pseudo-Likelihood Estimator (MPLE) $\tilde{\theta}$ of θ is the maximizer of the pseudo likelihood function pseudo$L\left(Z;\theta\right)$ where the parameter θ is the argument and the observables $Z = \left(Z_1, ..., Z_n\right)^\top$ are fixed values.

$$\tilde{\theta} = \arg\max_{\theta}\left(\text{pseudo}L\left(Z;\theta\right)\right)$$

Remark 89. Then (5.13) becomes: Given a dataset $\left\{\left(s_i, Z_i = Z\left(s_i\right)\right); i = 1, ..., n\right\}$,estimation of the unknown parameters $\{\alpha_i\}$ and $\{\beta_{i,j}\}$ of a Besag auto-model, in the MPLE framework, can be performed by maximizing the pseudo-likelihood, as

$$\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) = \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}}\left(\prod_{i\in\mathcal{S}} \text{pr}_Z\left(Z_i | Z_{-i}, \theta\right)\right) \tag{5.16}$$

$$= \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}}\left(\sum_{i\in\mathcal{S}} \log\left(\text{pr}_Z\left(Z_i | Z_{-i}, \theta\right)\right)\right) \tag{5.17}$$

$$\text{subject to } \beta_{i,j} = \beta_{j,i},\ \forall i, j \in \mathcal{S}$$

$$\text{...and any other problem specific restrictions}$$

Example 90. (Logistic auto-model / Ising model) For the Ising model it is

$$\log\left(\text{pr}_i\left(z_i | z_{-i}\right)\right) = z_i\left(\alpha_i + \sum_{j\sim i}\beta_{i,j} z_j\right) - \log\left(1 + \exp\left(\alpha_i + \beta_{i,j}\sum_{j:j\sim i}\beta_{i,j} z_j\right)\right)$$

and hence

$$\begin{aligned}\left(\{\hat{\alpha}_i\}, \left\{\hat{\beta}_{i,j}\right\}\right) &= \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}}\left(\log\left(\text{pr}_i\left(z_i | z_{-i}\right)\right)\right)\\ &= \arg\max_{\{\alpha_i\},\{\beta_{i,j}\}}\left(\sum_{i\in\mathcal{S}} z_i\left(\alpha_i + \sum_{j\sim i}\beta_{i,j} z_j\right) - \sum_{i\in\mathcal{S}}\log\left(1 + \exp\left(\alpha_i + \beta_{i,j}\sum_{j:j\sim i}\beta_{i,j} z_j\right)\right)\right)\end{aligned}$$

which does not depend on the normalizing constant (5.15) and it is less computationally demanding.

Parameterization remarks.

Remark 91. The unknown parameter vector $\theta = ((\alpha_i; i \in \mathcal{S}), (\beta_{i,j}; i, j \in \mathcal{S}))$ in the Besag's auto model (5.7) or even more general in the automodel (5.1) can be further parameterised to have a particular structure without the need to consider any additional constrains in Theorem 78.

Remark 92. The dimensionality of θ may be computationally prohibitively large when the size of the set of sites $\mathcal{S}$ is large (a usual case). To mitigate this issue, a way (alternative to pseudoLikelihood) is to set a structure on $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S},}$ reducing its dimentionality.

- For instance, one may specify

$$\alpha_i = aw_i, \quad \text{and} \quad \beta_{i,j} = b_i c_j; \text{ for } i, j \in \mathcal{S},$$

  with some known weights $\{w_i;\ i \in \mathcal{S},\}$ and unknown $\{a, b_i, c_j;\ i, j \in \mathcal{S}\}$. Then learning $\text{Card}(\mathcal{S})(1 + \text{Card}(\mathcal{S}))$ unknown parameter $\{\alpha_i, \beta_{i,j};\ i, j \in \mathcal{S},\}$ reduces to learning just $1 + 2\text{Card}(\mathcal{S})$ unknown parameters $\{a, b_i, c_j;\ i, j \in \mathcal{S}\}$. Note, that $\beta_{i,j} = b_i c_j$ restricts the interaction between i, j.

Remark 93. When observable covariates $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$ are available, one could "link" them to the model via the parameters $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S},}$.

- For instance by setting

$$\alpha_i = a_i + \sum_{k=1}^{p} d_k x_{i,k}, \quad \text{and} \quad \beta_{i,j} = \beta_{i,j}; \text{ for } i, j \in \mathcal{S}, \tag{5.18}$$

  where $\{a_i;\ i \in \mathcal{S}\}$, $\{d_k; k = 1, ..., k\}$ and $\{\beta_{i,j};\ i, j \in \mathcal{S}\}$ are unknown parameters. d_k represents the influence of k-th covariate $x_{i,k}$, for all $i \in \mathcal{S}$. $\beta_{i,j}$ represents the influence of the $z_{\partial i}$ at the neighboring sites of Z_i. Examination of the sign of $\beta_{i,j}$, and d_k or whether $\beta_{i,j} \neq 0$, $d_k \neq 0$ facilitates the discovery of pasterns and conditional dependencies.

Remark 94. Perhaps, in Remark 93, one (or many) of the observable covariates in vector x_i for $i \in \mathcal{S}$ can be "time" t dependent. One could "link" them to the model via the parameters $\{\alpha_i, \beta_{i,j}\}_{i,j\in\mathcal{S},}$, and make the automodel dynamical (aka spatio-temporal).

- For instance, if one consider $x_i = (t_i, t_i^2)^\top$ for $i \in \mathcal{S}$ and set

$$\alpha_i = a_i + d_1 t_i + d_2 (t_i)^2, \quad \text{and} \quad \beta_{i,j} = \beta_{i,j}; \text{ for } i, j \in \mathcal{S},$$

  essentially he/she makes the model dynamic (or space-time, or spatio-temporal)

Of course, how to parameterize the covariates in (5.18) is problem dependent, (similar to the linear regression), and hence model comparison methods, such as AIC, BIC, cross-validation, LASSO, etc... can be utilised.

Example 95. In Example 84, given observable covariates $x_i = (x_{i,1}, ..., x_{i,p})^\top$ for $i \in \mathcal{S}$, one may set (5.12) as

$$\log\left(\lambda_i\left(z_{-i}\right)\right) = \left[a_i + \sum_{k=1}^{p} d_k x_{i,k}\right] + \left[\sum_{j:j\sim i} \beta_{i,j} z_j\right] \tag{5.19}$$

Then d_k represent the influence of k-th covariate $x_{i,k}$, for all $i \in \mathcal{S}$, and $\beta_{i,j}$ represents the influence of the $z_{\partial i}$ at the neighboring sites of Z_i. For admissibility, a condition such as $\beta_{i,j} \leq 0$ should be specified (see Example 84). Further restrictions on the unknown parameters, or dimension reduction techniques, should be used because the number of unknowns is greater than the number of observations in (5.19).

Example 96. In Example 84, if the dataset is $\{(t_i, s_i, Z_i)\,; i \in \mathcal{S}\}$ where Z_i is the measurement (e.g. counts of a characteristic), at time t_i, at location $s_i \in \mathbb{R}^2$ of the i-th observation, a researcher may consider a parametrization

$$\log\left(\lambda_i\left(z_{-i}, t_i\right)\right) = \left[a_i + d_1 t_i\right] + \left[\sum_{j:j\sim i} \beta_{i,j} z_j\right] \tag{5.20}$$

and be interested in learning the unknown parameters $\{a_i\}$, d_1, and $\{\beta_{i,j}\}$. Obviously, the resulted model is space-time.

6. Hierarchical modeling (Bayesian modeling)

6.1. Revision: A general framework for the hierarchical modeling.

Note 97. Uncertainty can be decomposed according to the Hierarchical spatial model

$$\begin{cases} Z|Y,\vartheta & \text{data model} \\ Y|\vartheta & \text{spatial process model} \end{cases} \tag{6.1}$$

with

$$\mathrm{pr}\left(Z, Y|\vartheta\right) = \mathrm{pr}\left(Z|Y,\vartheta\right)\mathrm{pr}\left(Y|\vartheta\right)$$

Data model: expresses the measurement uncertainty as it is quantified via the distribution $\mathrm{pr}\left(Z|Y,\vartheta\right)$ possibly labeled by some parameter ϑ. It is often specified/modeled so that it can measure the goodness of fit between Z and Y.

Spatial process model: expresses the scientific uncertainty (e.g., that coming from (Y_s)) as it is quantified via the specified distribution $\mathrm{pr}\left(Y|\vartheta\right)$ possibly labeled by some

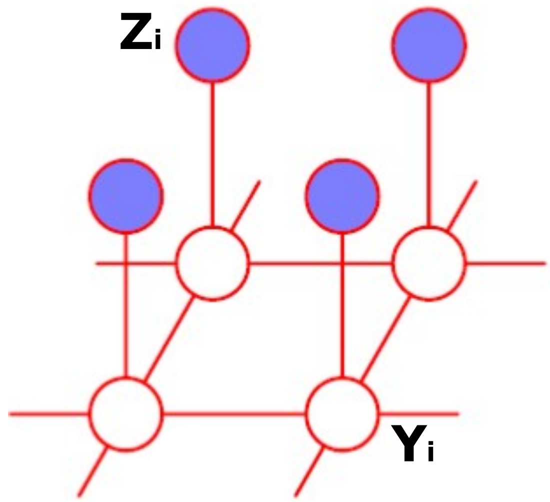


FIGURE 6.1. Hierarchical spatial model structure. $\{Y_i\}$ is the spatial process model which is hidden. $\{Z_i\}$ is the data model. The cartoon depicts a hierarchical spatial model with the special conditional independence structure $Z_i|\{Y_i\},\vartheta \sim \prod_i \mathrm{pr}(Z_i|Y_i,\vartheta)$ and $Y|\vartheta \sim \mathrm{pr}(Y|\vartheta)$

parameter ϑ. It is often specified/modeled with purpose (among others) to encourage spatial coherence and represente spatial dependence.

See for example Fig 6.1

Note 98. Let the unknown parameter vector be $\vartheta = (\vartheta_1, \vartheta_2)^\top$. Assume that a prior is specified for the unknown ϑ_1 as $\vartheta_1|\vartheta_2 \sim \mathrm{pr}(\vartheta_1|\vartheta_2)$ i.e. ϑ_1 is unknown and random. Assume ϑ_2 is a fixed parameter without a specified prior; it can be considered sometimes as known and sometimes as unknown in what follows. (!)

Note 99. Then the Bayesian spatial hierarchical model becomes

$$\begin{cases} Z|Y,\vartheta_1,\vartheta_2 & \text{data model} \\ Y|\vartheta_1,\vartheta_2 & \text{spatial process model} \\ \vartheta_1|\vartheta_2 & \text{hyper-parameter prior model} \end{cases} \tag{6.2}$$

where uncertainty is described by

$$\mathrm{pr}(Z,Y,\vartheta_1|\vartheta_2) = \mathrm{pr}(Z|Y,\vartheta_1|\vartheta_2)\,\mathrm{pr}(Y|\vartheta_1,\vartheta_2)\,\mathrm{pr}(\vartheta_1|\vartheta_2)$$

Note 100. Under Bayesian model (6.2), when ϑ_2 is considered as unknown (but fixed), ϑ_2 can be learned pointwise by computing a point estimator $\hat{\vartheta}_2$ as MLE i.e.

$$\hat{\vartheta}_2 = \arg\min_{\vartheta_2}\left(-2\log\left(\mathrm{pr}(Z|\vartheta_2)\right)\right)$$

by maximizing the marginal likelihood

$$\mathrm{pr}\left(Z|\vartheta_2\right)=\int \mathrm{pr}\left(Z,Y,\vartheta_1|\vartheta_2\right)\mathrm{d}Y\mathrm{d}\vartheta_1$$

or as a MPLE

$$\tilde{\vartheta}_2=\arg\min_{\vartheta_2}\left(-2\log\left(\prod_i \mathrm{pr}\left(Z_i|Z_{-i},\vartheta_2\right)\right)\right)$$

by maximizing the pseudo marginal Likelihood

$$\mathrm{pseudo}L\left(Z|\vartheta_2\right)=\prod_i \mathrm{pr}\left(Z_i|Z_{-i},\vartheta_2\right)$$

as a computational cheap approximation of the MLE.

Note 101. Under Bayesian model (6.2), when ϑ_1 is considered as unknown (but random), namely, the a prior $\vartheta_1\sim\mathrm{pr}\left(\vartheta_1|\vartheta_2\right)$ has been specified, uncertainty about unknown ϑ_1 given Y and ϑ_2 can be represented by the posterior distribution

$$\mathrm{pr}\left(\vartheta_1|Z,\vartheta_2=\hat{\vartheta}_2\right)=\frac{\mathrm{pr}\left(Z|\vartheta_1,\vartheta_2\right)\mathrm{pr}\left(\vartheta_1|\vartheta_2=\hat{\vartheta}_2\right)}{\mathrm{pr}\left(Z|\vartheta_2=\hat{\vartheta}_2\right)}$$

where the value $\hat{\vartheta}_2$ is plugged in. (Alternatively, we can plug-in $\tilde{\vartheta}_2$).

Note 102. General interest lies in computing the posterior distributions of the spatial process model $(Y_i)_{i\in\mathcal{S}}$, (or latent process, or noiseless process) given the data Z

$$\mathrm{pr}\left(Y|Z,\vartheta_2=\hat{\vartheta}_2\right)=\int \mathrm{pr}\left(Y|Z,\vartheta_2=\hat{\vartheta}_2\right)\mathrm{pr}\left(\vartheta_1|Z,\vartheta_2=\hat{\vartheta}_2\right)\mathrm{d}\vartheta_1$$

Note 103. The above statistical problem is naturally addressed in the (either full or empirical) Bayesian statistical framework.

Note 104. Below we give two examples in aerial data.

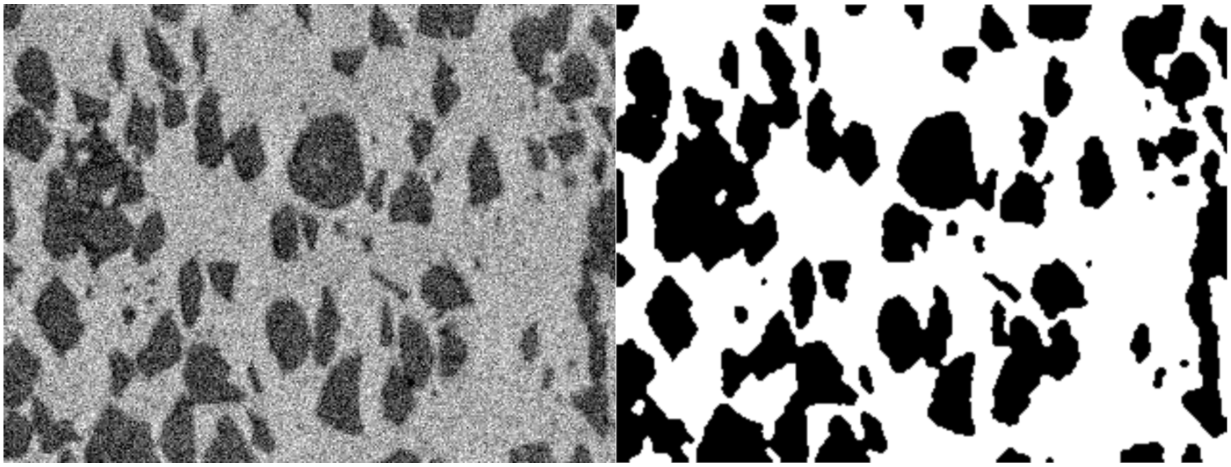
6.2. **Examples.**

6.2.1. *A simple spatial model for binary data (e.g. Image denoising).*

Example 105. (Image denoising) A central aim in image processing is to reconstruct an object (e.g. image) $Y=(Y_i;i\in\mathcal{S})$ based on a measurement (observation) $Z=(Z;i\in\mathcal{S})$ which is contaminated by errors $\varepsilon=(\varepsilon_i;i\in\mathcal{S})$. The framework of hierarchical modeling for aerial spatial data is suitable to address such cases.

Consider the image restoration dataset in Example 23 in Handout 1: Types of spatial data. (Fig 6.2a) We have a black and white noisy image with size 240×320 pixels.

Mathematically, denote $(Z_i)_{i\in\mathcal{S}}$ as the error contaminated (observed) image. The observables are coded as $Z_i=1$ for black and $Z_i=0$ for white at site $i\in\mathcal{S}=\{1,...,240\times 320\}$.



(A) Noised image (B) Reconstructed image

FIGURE 6.2. Ferrite-Pearlite steel image (Image restoration)

Let $n = \mathrm{Card}\left(\mathcal{S}\right)$. Hence $\left(Z_i\right)_{i\in\mathcal{S}}$ is a realization from the data model. The aim is to recover/learn the unknown real (error free) image $\left(Y_i\right)_{i\in\mathcal{S}}$ given the measurement/observation $\left(Z_i\right)_{i\in\mathcal{S}}$.

The data model can be specified (for instance) by "assuming" that the observation Z_i has been contaminated by iid noise with some "probability" p for all pixels $i \in \mathcal{S}$; i.e. $p = \mathrm{pr}\left(\left\{Z_i \neq Y_i\right\} | p\right) = 1 - \mathrm{pr}\left(\left\{Z_i = Y_i\right\}\right)$ for all $i \in \mathcal{S}$. Hence

$$\mathrm{pr}\left(Z_i | Y_i, p\right) = p^{1-1(\{Z_i = Y_i\})} \left(1-p\right)^{1(\{Z_i = Y_i\})},\ i \in \mathcal{S}$$

Consequently, the data model is

$$\begin{aligned}\mathrm{pr}\left(Z | Y, p\right) &= \prod_{i=1}^{n} p^{1-1_{\{Y_i\}}(Z_i)} \left(1-p\right)^{1_{\{Y_i\}}(Z_i)} = p^{n_{(Z,Y)}} \left(1-p\right)^{n-n_{(Z,Y)}} \\ &= \exp\left(n_{(Z,Y)} \log\left(\frac{p}{1-p}\right) + \left(1-p\right)^{n}\right)\end{aligned}$$

where $n_{(Z,Y)} = \sum_{i\in\mathcal{S}} 1\left(\left\{Z_i = Y_i\right\}\right)$.

The spatial process $\left(Y_i\right)_{i\in\mathcal{S}}$ is unknown (unobserved), and, according the Bayesian paradigm, we need to specify a prior process on $\left(Y_i\right)_{i\in\mathcal{S}}$ account for the uncertainty. To introduce spatial dependence, the researcher may judge to specify (for example) an Ising process prior such as

$$\mathrm{pr}\left(Y | \alpha, \beta\right) \propto \exp\left(\alpha \sum_{i\in\mathcal{S}} Y_i + \beta \sum_{\{i,j\}: i\sim j} Y_i Y_j\right)$$

with symmetric relation $i \sim j$ considering only the adjacent pixels.

The researcher may be uncertain about the "real" value of p and hence he/she may want to specify a conjugate Beta prior[3] $p \sim \mathrm{Be}(g,h)$ with known g and h to account for the uncertainty. The researcher may set certain fixed values on α and β; hence consider that g, h α, and β are known values.

The Hierarchical Bayesian model is

$$
\begin{cases}
Z|Y,p \sim \mathrm{pr}(Z|Y,p) & \text{data model} \\
Y \sim \mathrm{pr}(Y|\alpha,\beta) & \text{spatial process model} \\
p \sim \mathrm{Be}(g,h) & \text{hyper-parameter prior model}
\end{cases} \tag{6.3}
$$

To learn $Y|Z$, one can compute the Bayesian MAP estimator of Y, i.e.

$$
\begin{aligned}
\hat{Y} &= \arg\max_{Y} \left(\log\left(\mathrm{pr}(Z|Y)\,\mathrm{pr}(Y)/\mathrm{pr}(Z)\right)\right) \\
&= \arg\min_{Y} \left(-\log\left(\mathrm{pr}(Z|Y)\right) - \log\left(\mathrm{pr}(Y)\right)\right)
\end{aligned}
$$

via an optimization numerical algorithm, or perhaps the posterior expectation, i.e.

$$
\hat{Y} = \mathrm{E}(Z|Y) = \int Z \mathrm{pr}(Z|Y)\,\mathrm{d}Z
$$

via MCMC, INLA, etc... where

$$
\begin{aligned}
\mathrm{pr}(Z|Y) &= \int p^{n(Z,Y)} (1-p)^{n-n(Z,Y)}\, \mathrm{Be}(p|g,h)\,\mathrm{d}p \\
&= \int p^{n(Z,Y)} (1-p)^{n-n(Z,Y)} \frac{p^{g-1}(1-p)^{h-1}}{\mathrm{B}(g,h)}\mathrm{d}p \\
&= \frac{1}{\mathrm{B}(g,h)} \int p^{n(Z,Y)+g-1} (1-p)^{n-n(Z,Y)+h-1}\,\mathrm{d}p \\
&= \frac{1}{\mathrm{B}(g,h)} \mathrm{B}\left(n(Z,Y)+g, n-n(Z,Y)+h\right)
\end{aligned}
$$

[3] $\mathrm{Be}(p|g,h) = p^{g-1}(1-p)^{h-1}\, 1_{(0,1)}(p)/\mathrm{B}(g,h)$

Then the marginal posterior can be computed analytically as

$$\mathrm{pr}(Y|Z) = \int \mathrm{pr}(Y,p|Z)\,\mathrm{d}p = \int \frac{\mathrm{pr}(Z|Y,p)\,\mathrm{pr}(Y)\,\mathrm{pr}(p)}{\int \mathrm{pr}(Z|Y,p)\,\mathrm{pr}(Y)\,\mathrm{pr}(p)\,\mathrm{d}p\mathrm{d}Z}\mathrm{d}p$$

$$\propto \underbrace{\int \mathrm{pr}(Z|Y,p)\,\mathrm{pr}(p)\,\mathrm{d}p}_{=\mathrm{pr}(Z|Y)}\mathrm{pr}(Y) = \mathrm{pr}(Z|Y)\,\mathrm{pr}(Y)$$

$$\propto \underbrace{\frac{1}{\mathrm{B}(g,h)}\mathrm{B}\left(n_{(Z,Y)}+g, n-n(Z,Y)+h\right)}_{=\mathrm{pr}(Z|Y)}$$

$$\times \underbrace{\frac{\exp\left(\alpha\sum_i Y_i + \beta\sum_{j\sim i} Y_iY_j\right)}{\sum_{Y\in\{0,1\}^n}\exp\left(\alpha\sum_i Y_i + \beta\sum_{j\sim i} Y_iY_j\right)}}_{=\mathrm{pr}(Y)}$$

$$\propto \mathrm{B}\left(n(Z,Y)+g, n-n(Z,Y)+h\right)\exp\left(\alpha\sum_i Y_i + \beta\sum_{j\sim i} Y_iY_j\right) \tag{6.4}$$

Note that the only reason that we ignored the constant from the Ising prior in (6.4) was because, in this particular example, the researcher considered α and β as known constants. If α and β where considered unknown, and hence we had to learn them, then that constant should not be ignored of course.

Fig. 6.2b shows the restored image as the Bayesian MAP estimator of $Y|Z$ by using an R optimization function.

6.2.2. *A simple spatial model for count data (e.g. Disease mapping).*

Note 106. The following model is computationally intractable as far as posterior inference concerns.

Example 107. Consider there is available a dataset $\{(x_i, s_i, Z_i); i = 1, ..., n\}$, where $Z_i \in \mathbb{N}$ is the count of the occurrence of an event in a particular time interval, at a location with coordinates s_i, and associated with a vector of covariates (other measurements) $X_i = (X_{i,1}, ..., X_{i,k})^\top$, for $i \in \mathcal{S}$, $\mathcal{S} = \{1, ..., n\}$, and $n \in \mathbb{N} - \{0\}$ fixed. So, denote $(Z_i)_{i\in\mathcal{S}}$ as the obeserved vector. Assume that $Z_i \in \mathcal{Z}^{\mathcal{S}}$, with $\mathcal{Z} = \mathbb{N}$ and $\mathcal{S} = \{1, ..., n\}$.

For count observations Z_i, it is natural to assume a Poisson distribution with some rate λ_i. The researcher may wish to consider that Z_i are The researcher may specify the data model as

$$Z_i|Y_i \sim \mathrm{Poisson}\left(\lambda_i(Y_i)\right), \text{ for } i = 1, ..., n \tag{6.5}$$

$$\text{where} \quad \log(\lambda_i) = Y_i$$

The spatial process Y is unknown. To specify the uncertainty on Y, the researcher may judge to assign a CAR model prior on $(Y_i)_{i\in\mathcal{S}}$, for instance

$$(6.6)\qquad \begin{aligned} Y_i|Y_{-i} &\sim \mathrm{N}\left(\mu_i+\beta_{i,j}\left(Y_j-\mu_j\right),\kappa_i\right), \text{ for } i=1,...,n\\ \mu_i &= X_i^\top\alpha. \end{aligned}$$

To reduce parametric dimensionality, we impose a more restrictive structure such that $\kappa_i=\kappa$ for all $i=1,...,n$, and

$$\beta_{i,j}=\begin{cases}\phi & \text{if } i\sim j\\ 0 & \text{if } i\nsim j \text{ or } i=j\end{cases}$$

This essentially produces a joint model (see Sec 3.1 or just use Theorem 78)

$$Y\sim\mathrm{N}\left(X\alpha,\left(I-\phi N\right)^{-1}\kappa\right)$$

where N is an $n\times n$ matrix with $[N]_{i,j}=1\left(\{i\sim j\},i\neq j\right)$, where $\sim$ is defined to denote adjacent sites (or otherwise spatial locations sharing same boarders).

For the unknown hyper-parameters α, ϕ and κ, the researcher may consider hyper-priors $\alpha\sim\mathrm{N}\left(0,\Sigma_\alpha\right)$, $\phi\sim\mathrm{U}\left(0,\phi_{\max}\right)$, and $\kappa\sim\mathrm{IG}\left(g,h\right)$; the prior distributions here are chosen for demonstration. The rest hyper-parameters $\Sigma_\alpha>0$ $\phi_{\max}\in\{\phi>0:I-\phi N \text{ is non singular}\}$, $g>0$, and $h>0$ are considered as unknown fixed constants set by the researcher based on his/her subjective believes.

The Bayesian spatial hierarchical model becomes

$$(6.7)\qquad \begin{cases} Z_i|Y_i\overset{\text{ind}}{\sim}\text{Poisson}\left(\exp\left(Y_i\right)\right),\ \forall i & \text{data model}\\ Y|\alpha,B,K\sim\mathrm{N}\left(X\alpha,\left(I-\phi N\right)^{-1}\kappa\right) & \text{spatial process model}\\ \alpha\sim\mathrm{N}\left(\mu_\alpha,\Sigma_\alpha\right) & \text{hyper-prior model}\\ \phi\sim\mathrm{U}\left(0,\phi_{\max}\right) & \text{hyper-prior model}\\ \kappa\sim\mathrm{IG}\left(g,h\right) & \text{hyper-prior model}\end{cases}$$

The joint probability model becomes

$$\begin{aligned}\mathrm{pr}\left(Z,Y,\alpha,\beta,\kappa\right) &= \prod_{i\in\mathcal{S}}\mathrm{pr}\left(Z_i|Y_i\right)\mathrm{pr}\left(Y|\alpha,\phi,\kappa\right)\mathrm{pr}\left(\alpha,\phi,\kappa\right)\\ &= \prod_{i\in\mathcal{S}}\text{Poisson}\left(Z_i|\exp\left(Y_i\right)\right)\mathrm{N}\left(Y|X\alpha,\left(I-\phi\right)^{-1}\kappa\right)\\ &\times\mathrm{N}\left(\alpha|\mu_\alpha,\Sigma_\alpha\right)\mathrm{U}\left(\phi|0,\phi_{\max}\right)\mathrm{IG}\left(\kappa|g,h\right)\end{aligned}$$

Interest lies in learning $Y|Z$which can be addressed for instance by the Bayesian MAP estimator

$$\hat{\lambda}=\arg\max_Y\left(\mathrm{pr}\left(\lambda|Z\right)\right)$$

or the posterior expectation estimator

$$\hat{\lambda} = \mathrm{E}_{\mathrm{pr}}\left(\lambda|Z\right) = \mathrm{E}_{\mathrm{pr}}\left(\exp\left(Y\right)|Z\right)$$

$\mathrm{pr}\left(\lambda|Z\right)$ can be computed via random variable transformation from $\mathrm{pr}\left(Y|Z\right)$ which is given by the Bayesian theorem as

$$\mathrm{pr}\left(Y|Z\right) = \int \frac{\mathrm{pr}\left(Z, Y, \alpha, B, \kappa\right)}{\int \mathrm{pr}\left(Z, Y, \alpha, B, \kappa\right) \mathrm{d}Y \mathrm{d}\alpha \mathrm{d}B \mathrm{d}\kappa} \mathrm{d}\alpha \mathrm{d}B \mathrm{d}\kappa$$

The above integration is analytically intractable, and hence its numerical computation can be performed by methods such as MCMC, INLA, etc...

Example 108. Numerical example in spirit of Example 107.

Consider the (Columbus Columbus OH data set) which concerns spatially correlated count data arising from small area sampling of some underlying process.

This is the R dataset `columbus{spdep}`. Briefly, the columbus data frame has 49 rows and 22 columns. Unit of analysis is 49 neighbourhoods in Columbus, OH, 1980 data. The date frame has the form $(\text{CRIME}_i, \text{HOVAL}_i, \text{INC}_i)^\top$ for each neiborhood $i = 1, .., n = 49$, where

CRIME: residential burglaries and vehicle thefts per thousand households in the neighborhood

HOVAL: housing value (in 1,000 USD)

INC: household income (in 1,000 USD)

Figure 6.3a shows the Property crime (number per thousand households) in 49 districts in Columbus in 1980, as well as the average value of the house in USD. Figure 6.3b presents the corresponding average house value. For privacy reasons, these are typically aggregated over areas that are large enough to ensure that the counts cannot be traced back to individuals. Interest may lie to find whether high rates of crime are clustered in a particular areas, and if yes, perhaps what is the association of it with the value of the houses in the area.

Regarding the data model $(Z_i)_{i\in\mathcal{S}} \mid (Y_i)_{i\in\mathcal{S}}$, it is reasonable to specify a Poisson distribution as the statistical model representing the data model as in (6.5). This is because, for each county $i \in \mathcal{S}$ the observation Z_i represents count namely number of event in a specific time and space. We can impose the (rather strong) assumption that Z_i and Z_j are conditionally independent given the spatial process $(Y_i)_{i\in\mathcal{S}}$. That is

$$\begin{aligned} Z_i|Y_i &\sim \text{Poisson}\left(\lambda_i\left(Y_i\right)\right), \text{ for } i = 1, ..., n \\ \text{where} \quad & \log\left(\lambda_i\right) = Y_i \end{aligned}$$

Regarding the spatial process, we can consider the CAR model as in (6.6)

$$\begin{aligned} Y_i|Y_{-i} &\sim \text{N}\left(\mu_i + \sum_{i\sim j} \phi\left(Y_j - \mu_j\right), \kappa_i\right), \text{ for } i = 1, ..., n \\ \mu_i &= X_i^\top \alpha. \end{aligned}$$

with regressors $X_i = (1, \text{HOVAL}_i, \text{INC}_i)^\top$; that is CRIME ~ HOVAL + INC. This is because we are interested in investigating "whether high rates of crime (CRIME) are clustered in a particular areas ($i \in \mathcal{S}$), eg areas with expensive houses (HOVAL), and if yes, perhaps what is the association of it with the value of the houses in the area (INC)". Hence we wise to use our model in order to see the association of CRIME with SPACE (i.e. $i \in \mathcal{S}$), HOVAL (i.e. house value), and INC (i.e. income). Note that unlike the usual linear model, here I have managed to introduce spatial dependence in the model as well by

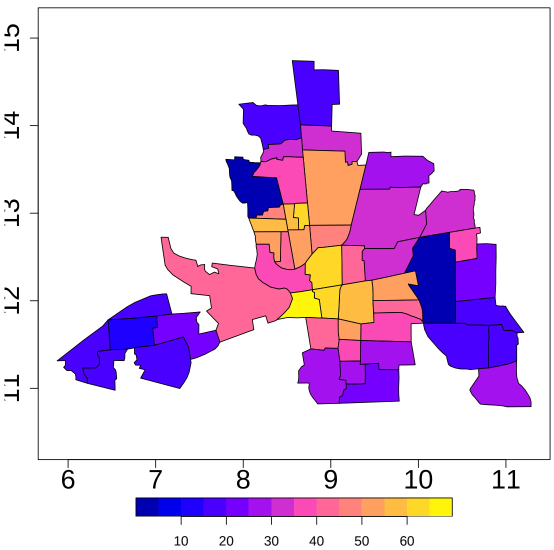
$$\mathrm{E}\left(Y_i|Y_{-i}\right) = \underbrace{\alpha_0 + \alpha_1 \mathrm{HOVAL}_i + \alpha_2 \mathrm{INC}_i}_{\text{cocariate dependence}} + \underbrace{\sum_{i\sim j} \phi\left(Y_j - \mu_j\right)}_{\text{spatial dependence}}$$

For the hyper-parameters, we considered $\mu_\alpha = 0$, $(\Sigma_\alpha)^{-1} = 0$, $g = 1$, $h = 1$, and $\phi_{\max} = 0.5$.

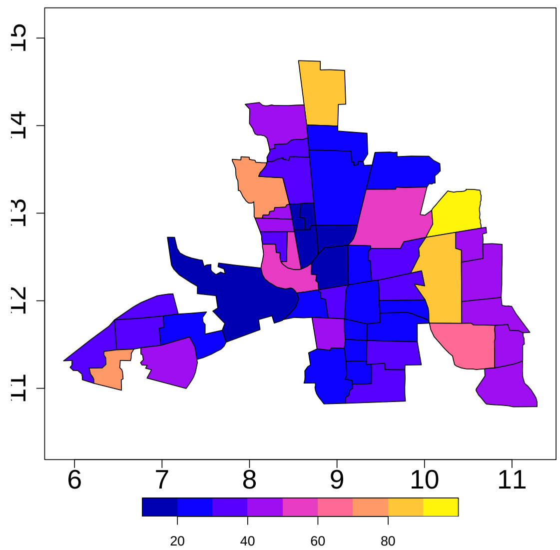
The resulted Bayesian hierarchical model is as in (6.7). Estimation was facilitated via INLA. The estimates are computed as the posterior expected values given the data Z as

$$\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_{\text{const}} \\ \hat{\alpha}_{\text{HOVAL}} \\ \hat{\alpha}_{\text{INC}} \end{pmatrix} = \begin{pmatrix} 54.3139189 \\ -0.2821969 \\ -0.9882862 \end{pmatrix}$$

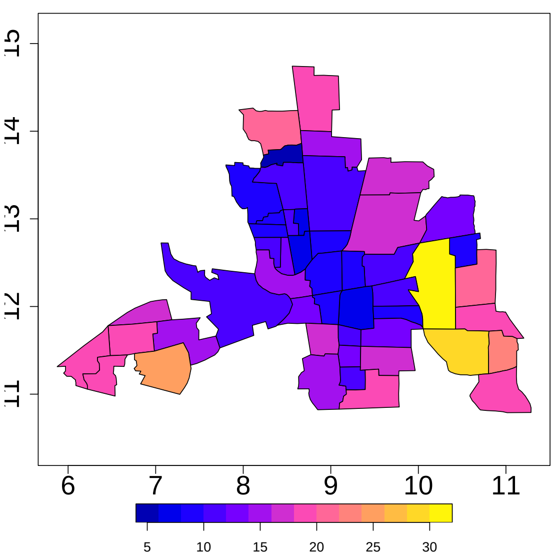
$\hat{\phi} = 0.1589004$, and $\hat{\kappa} = 87.65$. The fitted counts $\hat{Y}$ are presented in Figure 6.3d.



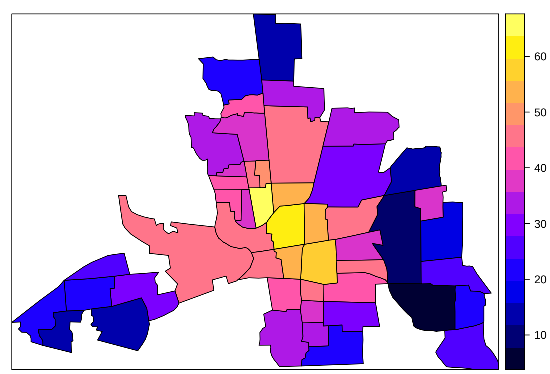
(A) CRIME: residential burglaries and vehicle thefts per thousand households in the neighborhood



(B) HOVAL: housing value (in 1,000 USD)



(C) INC: household income (in 1,000 USD)



(D) Fitted CRIME ~ HOVAL+INC in a Poisson model with rate modeled as CAR with mean CRIME ~ HOVAL+INC

FIGURE 6.3. Columbus Columbus OH spatial analysis dataset

The produced Log likelihood is -182.2198.

By comparing Fig 6.3a and Fig 6.3d, we see that there are certain locations where the fitted counts $\hat{Y}$ and Z are substantially different. Perhaps, we could improve our parameterization in (6.6) by considering a less restrictive $\beta_{i,j}$ or by including more covariates in the mean μ_i.