Auto-Modelli

Vers. 1.1.1

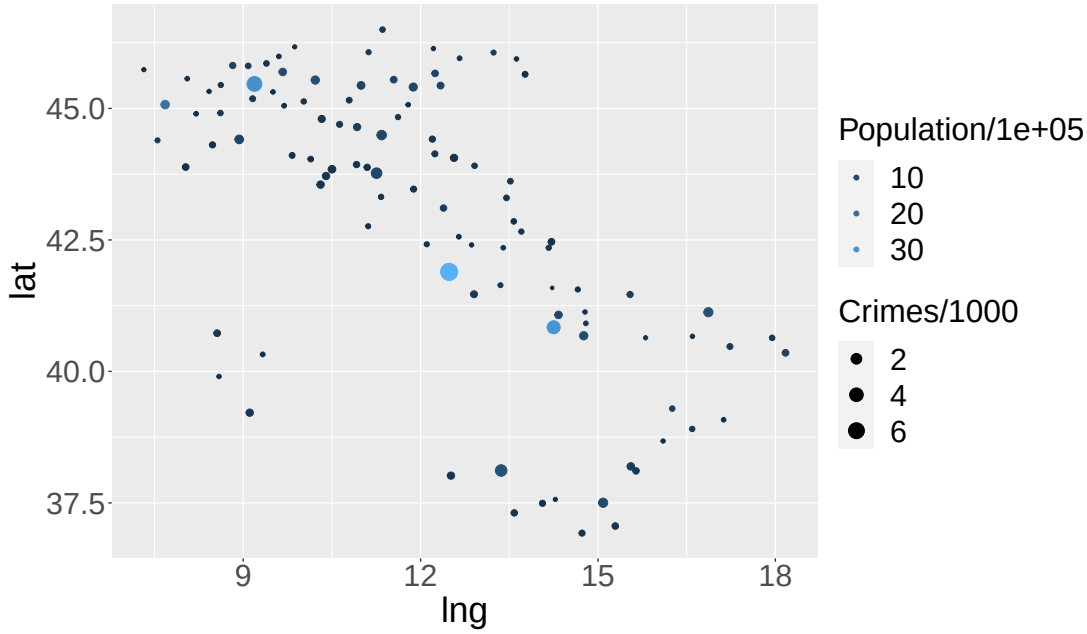
Gianluca Mastrantonio

email: gianluca.mastrantonio@polito.it

Markov Random Field

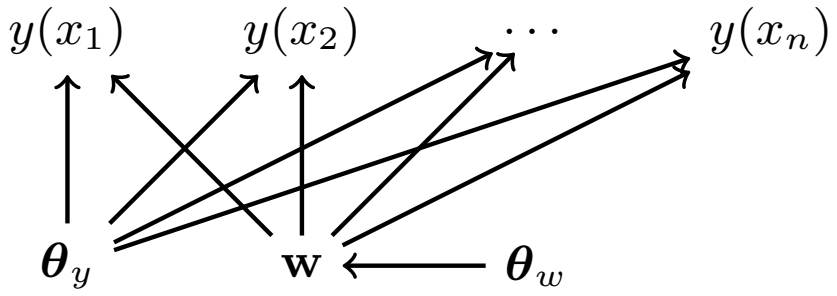
Markov Random Field

Riprendiamo l'esempio in cui avevamo dei crimini divisi per provincie.

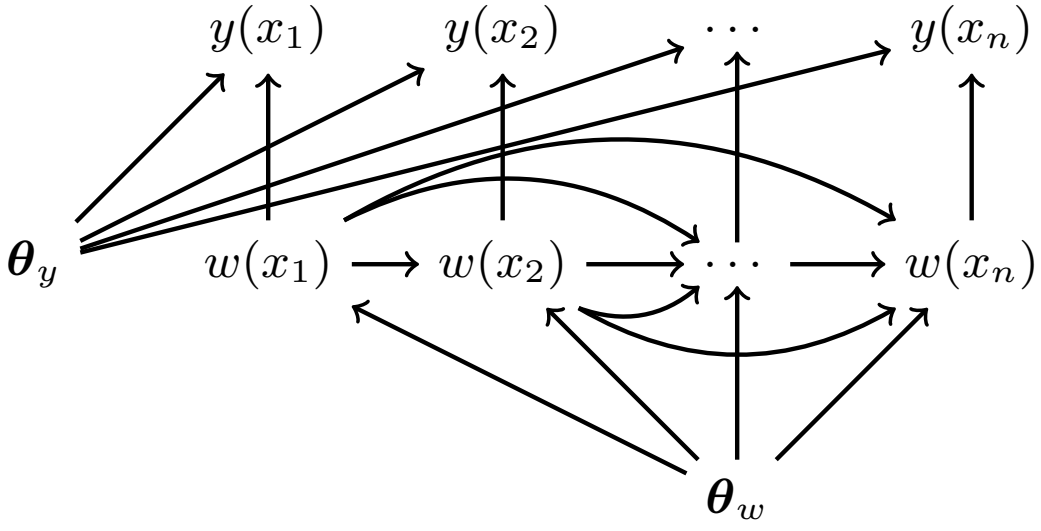


che avevamo modellizzato usando il seguente DAG

Markov Random Field



oppure



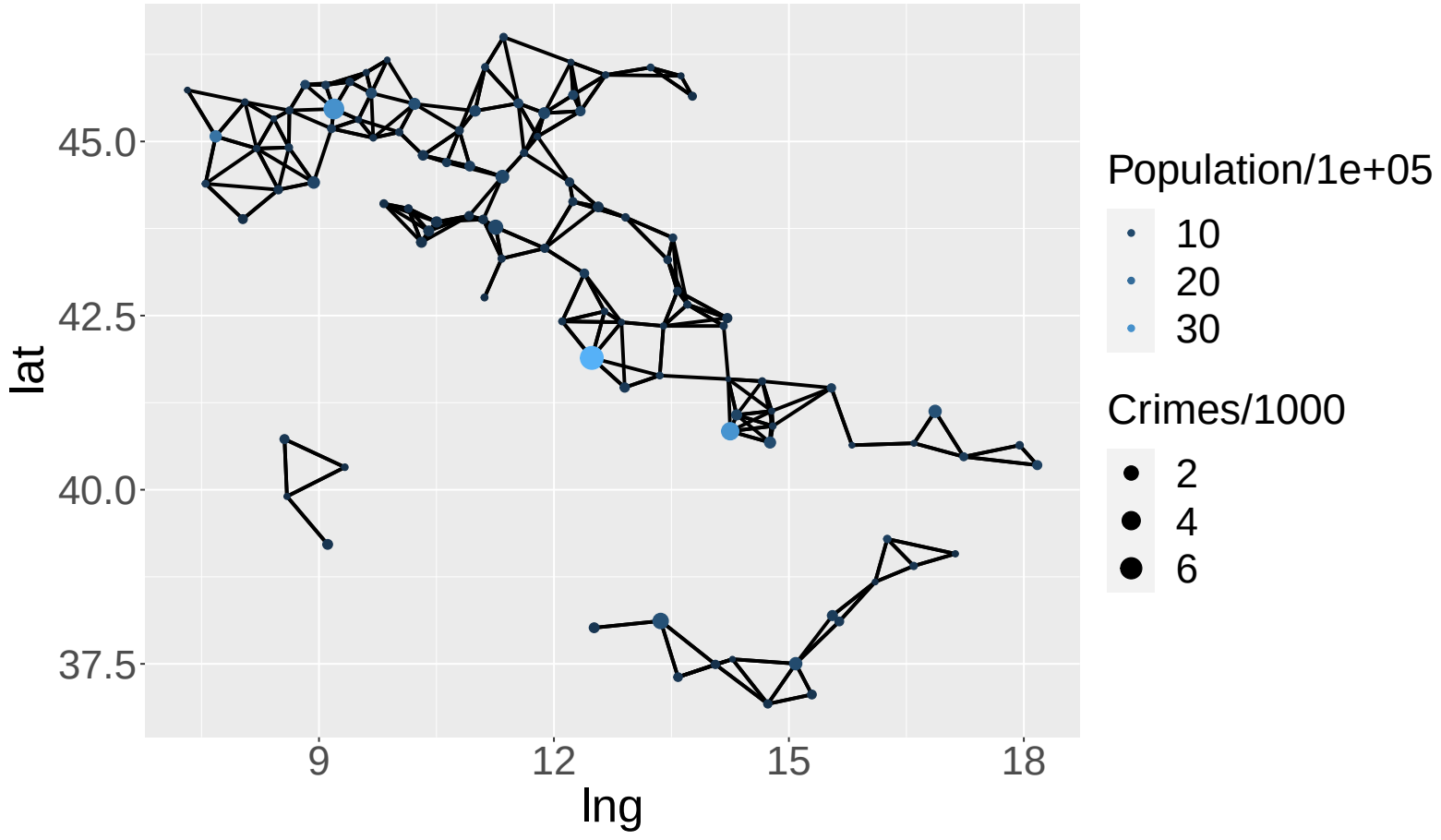
Markov Random Field

Stiamo implicitamente assumendpo che tutto dipende da tutto anche se questo non è necessariamente vero, e la dipendenza può dipendere

- presenza di connessioni, tipo strade
- avere un confine in comune
- l'esistenza di una rete sottostante di criminalità

Creiamo un grafo (non DAG) che rispecchi le dipendenze, assumendo che ogni provincia selezioni al più 4 vicini che sono più vicini spazialmente, e due città sono vicine solo se più vicine di 1 (distanza in latitudine-longitudine)

Markov Random Field



Se vogliamo partire da questo grafo per costruire un DAG abbiamo diversi problemi

- dove iniziare (nelle serie temporali c'è un prima e un dopo che in spaziale non c'è)

Markov Random Field

- se diamo un ordine e definiamo le condizionate incrementalmente, come controlliamo le marginali

Sembra più naturale lavorare con le full conditionals.
Notazione

- $i \sim j$ significa che y_i e y_j sono vicini
- ∂y_i è il set delle osservazioni che sono vicine a y_i (i.e., connesse da un segmento), ...
- $\mathbf{y}_{-i}$ è vettore di tutte le osservazioni meno la $i-$esima

Con la struttura scelta precedentemente, indipendentemente dal DAG, abbiamo che

$$f(y_i|\mathbf{y}_{-i}) = f(y_i|\partial y_i)$$

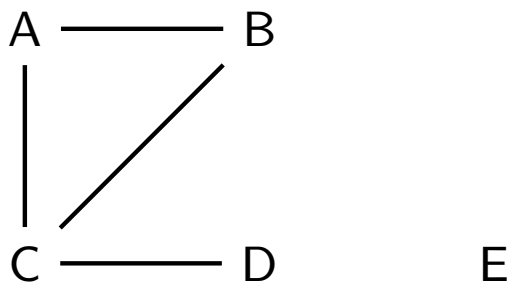
e questo vale per ogni i. Questo somiglia molto alla proprietà di Markov, ma non richiede un ordinamento particolare. Quando un processo ha questa proprietà per ogni i, allora si dice che è un **Markov Random Field** $\rightarrow$ Naturalmente, per avere "senso", la cardinalità di ∂y_i deve essere molto più piccola di $\mathbf{y}_{-i}$

Markov Random Field

Clique

Una **Clique** sono un set di nodi un grafo che sono tutti connessi tra di loro. Per definizione, ogni nodo è un clique

In questo caso



Markov Random Field

le Clique sono $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$, $\{E\}$, $\{A, B\}$, $\{A, C\}$, $\{B, C\}$, $\{C, D\}$, $\{A, B, C\}$. Indichiamo con $\mathcal{C}$ il set di tutte le clique di un grafo, e con $\mathbf{y}_C$ come gli elementi della clique $C \in \mathcal{C}$, allora

Hammersley-Cliffort Theorem

Se $f(\mathbf{y}) > 0$ per tutte le possibili configurazioni $\mathbf{y} \in \mathcal{Y}$ (positivity condition), allora

1. Se $\mathbf{Y}$ è un MRF, allora

$$f(\mathbf{y}) \propto \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{y}_C)$$

   dove $0 < \Psi_C < \infty$ sono funzioni appropriate

2. se la densità di una variabile aleatoria $\mathbf{Y}$ si può scrivere come

$$f(\mathbf{y}) \propto \prod_{C \in \mathcal{C}} \Psi_C(\mathbf{y}_C)$$

   allora $\mathbf{Y}$ è un MRF

Markov Random Field

Possiamo anche scrivere

$$f(\mathbf{y}) = \frac{1}{Z} \exp\left(- \sum_{C \in \mathcal{C}} V_C(\mathbf{y}_C) \right)$$

con $V_C(\mathbf{y}_C) = \log \Psi_C(\mathbf{y}_C)$, e si dice che $f(\mathbf{y})$ è un distribuzione di Gibbs (usata molto in fisica) e $V_C(\mathbf{y}_C)$ sono **potential-function**

Date delle full conditional, possiamo usare il teorema per trovare la congiunta (e poi verificare se è' valida). Un'altro teorema che ci può essere utile allo stesso scopo è il brook's lemma, famoso tra chi fa statistica spaziale, ma generalmente poco noto.

Markov Random Field

Brook's Lemma

Se $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{Y}$, $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{Y}$ sono due campioni da una distribuzione $f()$, allora

$$\begin{aligned}\frac{f(\mathbf{y})}{f(\mathbf{x})} =& \frac{f(y_1|y_2, y_2, \ldots y_{n-1}, y_n)}{f(x_1|y_2, y_2, \ldots y_{n-1}, y_n)} \frac{f(y_2|x_1, y_3, \ldots y_{n-1}, y_n)}{f(x_2|x_1, y_3, \ldots y_{n-1}, y_n)} \times \\ & \ldots \\ & \frac{f(y_{n-1}|x_1, x_2, \ldots x_{n-2}, y_n)}{f(x_{n-1}|x_1, x_2, \ldots x_{n-2}, y_n)} \frac{f(y_n|x_1, x_2, \ldots x_{n-2}, x_{n-1})}{f(x_n|x_1, x_2, \ldots x_{n-2}, x_{n-1})} \Rightarrow \\ \frac{f(\mathbf{y})}{f(\mathbf{x})} =& \prod_{i=1}^{n} \frac{f(y_i|x_1, x_2, \ldots, x_{i-1}, y_{i+1}, y_{n-2}, y_{n-1}, y_n)}{f(x_i|x_1, x_2, \ldots, x_{i-1}, y_{i+1}, y_{n-2}, y_{n-1}, y_n)}\end{aligned}$$

La dimostrazione è facile, visto che utilizza il teorema di Bayes iterativamente, ma molto lunga. Facciamolo nel caso di $n = 2$.

$$f(y_1, y_2) = f(y_1|y_2)f(y_2) = f(y_1|y_2)\frac{f(y_2|x_1)f(x_1)}{f(x_1|y_2)} = \frac{f(y_1|y_2)}{f(x_1|y_2)} f(y_2|x_1)f(x_1) =$$

Markov Random Field

$$\frac{f(y_1|y_2)}{f(x_1|y_2)} f(y_2|x_1) \frac{f(x_1,x_2)}{f(x_2|x_1)} = \frac{f(y_1|y_2)}{f(x_1|y_2)} \frac{f(y_2|x_1)}{f(x_2|x_1)} f(x_1,x_2)$$

Questo lemma ci dice che

- c'è una relazione 1 a 1 tra full conditional e congiunta
- con le full conditional posso conoscere solo il kernel e non la costante di normalizzazione, visto che mi danno il rapporto $\frac{f(\mathbf{y})}{f(\mathbf{x})}$

Dal punto di vista pratico devo "solo" dimostrare che il kernel integri/sommi a una costante

Possiamo mettere insieme i due teoremi, nel caso in cui

$$f(\mathbf{0}) > 0$$

e definendo

$$G(\mathbf{y}) = log \frac{f(\mathbf{y})}{f(\mathbf{0})}$$

Markov Random Field

In questo caso esiste una rappresentazione di $G(\mathbf{x})$ basata solo sulle clique

$$G(\mathbf{y}) = \sum_{i=1} y_i Q_i(y_i) + \sum_{i<j} y_i y_j Q_{i,j}(y_i, y_j) + \sum_{i<j<k} y_i y_j y_k Q_{i,j,k}(y_i, y_j, y_k) + \cdots +$$

$$y_1 y_2 \ldots y_n Q_{1,2,\ldots,n}(y_1, y_2, \ldots, y_n)$$

dove $Q_{.,.,.,}() = 0$ a meno che le variabili non formino un clique, e le funzioni Q sono arbitrarie e finite.

Auto-Poisson model

Torniamo ai dati del crimine. Potremmo definire la full conditional

$$f(y_i|\mathbf{y}_{-i})$$

in modo tale che se i vicini hanno valori elevati allora y_i tende ad avere valori elevati. Per esempio

$$y_i|\mathbf{y}_{-i} \sim Pois\left(\exp(\mu + \sum_{j\sim i} \beta y_j)\right)$$

sebbene questa abbia senso, intuitivamente, dobbiamo assicurarci che la congiunta esista, perchè definire le full conditional non ci assicura niente sull'esistenza della congiunta. Definendo

$$\lambda_i = \exp(\mu + \sum_{j\sim i} \beta y_j)$$

Abbiamo che

$$f(y_i|\partial y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

Auto-Poisson model

e visto che

$$f(0|\partial y_i) = e^{-\lambda_i}$$

allora

$$\frac{f(y_i|\partial y_i)}{f(0|\partial y_i)} = {\lambda_i}^{y_i}$$

Utilizzando Brook's lemma abbiamo

$$log\left(\frac{f(\mathbf{y})}{f(\mathbf{0})}\right) = \exp\left(\mu y_1 + \sum_{\substack{j\sim 1\\ j>1}} \beta y_1 y_j\right) \exp\left(\mu y_2 + \sum_{\substack{j\sim 2\\ j>2}} \beta y_2 y_j\right) \times$$

$$\ldots \exp\left(\mu y_{n-1} + \sum_{\substack{j\sim n-1\\ j>(n-1)}} \beta y_{n-1} y_j\right) \times \exp\left(\mu y_n\right) \exp\left(-\log(y_1!) - \cdots - \log(y_n!)\right) =$$

$$\exp\left(\sum_i (\mu y_i - \log(y_i!)) + \sum_{\substack{j\sim i\\ i<j}} \beta y_i y_j\right)$$

Auto-Poisson model

dovremmo determinare se somma a 1 ma è' molto complesso. SI dimostra che lo fa solo per $\beta < 0$, il che lo rende un modello poco interessante

Auto-Normal model (conditional autoregressive (CAR) model)

Invece di lavorare con la poisson, potremmo indurre dipendenza nel secondo livello della gerarchia

$$y_i|w_i \sim P(exp(\mu + w_i))$$
$$\mathbf{w} \sim MRF()$$

La cosa più semplice è un modello Gaussiano con full conditional (tra elementi del processo)

$$w_i|\partial w_i \sim N(\sum_{j\sim i} \beta_{i,j} w_j, \tau_i^2)$$

Dovremmo mettere dei vincoli ai parametri, ma per adesso lasciamoli liberi. sappiamo che

$$f(w_i|\partial w_i) \propto \exp\left(-\frac{(w_i - \sum_{j\sim i} \beta_{i,j} w_j)^2}{2\tau_i^2}\right) =$$
$$\exp\left(-\frac{w_i^2 + (\sum_{j\sim i} \beta_{i,j} w_j)^2 - 2w_i \sum_{j\sim i} \beta_{i,j} w_j}{2\tau_i^2}\right)$$

Auto-Normal model (conditional autoregressive (CAR) model)

e

$$f(0|\partial w_i) \propto \exp\left(-\frac{(\sum_{j\sim i}\beta_{i,j}w_j)^2}{2\tau_i^2}\right)$$

e

$$\frac{f(w_i|\partial w_i)}{f(0|\partial w_i)} = \exp\left(-\frac{w_i^2 - 2w_i\sum_{j\sim i}\beta_{i,j}w_j}{2\tau_i}\right)$$

e quindi, usando Brook's lemma otteniamo la seguente congiunta (se esiste)

$$f(w_i|\partial w_i) \propto \exp\left(-\sum_i \frac{w_i^2 - 2\sum_{\substack{j\sim i\\ i<j}}\beta_{i,j}w_iw_j}{2\tau_i}\right)$$

Il processo ha una congiunta Gaussiana, $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Lambda}^{-1})$. Se indichiamo con $\lambda_{j,k} = [\mathbf{\Lambda}]_{j,k}$ in questo caso la densità è

$$f(\mathbf{w}) \propto \exp\left(-\sum_i \frac{\lambda_{i,i}w_i^2 - 2\sum_{i<j}\lambda_{i,j}w_iw_j}{2}\right)$$

Auto-Normal model (conditional autoregressive (CAR) model)

che assomiglia a quella trovata come congiunta di un MRF (**NOTATE:** gli zeri nella matrice di precisione ci dicono tutto ciò che dobbiamo sapere sulla dipendenza condizionata). Quindi, la congiunta del MRF può essere scritta come

$$f(\mathbf{w}) \propto \exp\left(-\frac{\mathbf{w}'\mathbf{D}^{-1}(\mathbf{I}-\mathbf{B})\mathbf{w}}{2}\right)$$

dove

$$[\mathbf{B}]_{j,k} = \beta_{j,k}$$

con diagonale pari a 1 e $\mathbf{D}$ è diagonale con elementi τ_i^2. Dobbiamo però richiedere che $\mathbf{D}^{-1}(\mathbf{I}-\mathbf{B})$ sia simmetrica, e lo è se

$$\frac{\beta_{i,j}}{\tau_i^2} = \frac{\beta_{j,i}}{\tau_j^2}$$

Auto-Normal model (conditional autoregressive (CAR) model)

Il modello ha dei problemi, e per vederlo introduciamo una particolare formalizzazione molto utile. Definiamo $\mathbf{H}$ come la matrice di prossimità tra gli elementi di $\mathbf{w}$, tale per cui $[\mathbf{H}]_{i,j} = 1$ se e solo se $i \sim j$, altrimenti è zero. Definiamo $h_{i,+} = \sum_{j=1}^{n} [\mathbf{H}]_{i,j}$ e

$$\beta_{i,j} = \frac{h_{i,j}}{h_{i,+}} \qquad \tau_i^2 = \frac{\tau^2}{h_{i,+}}$$

Adesso possiamo scrivere

$$f(\mathbf{w}) \propto \exp\left(-\frac{\mathbf{w}'(\mathbf{D}_h - \mathbf{H})\mathbf{w}}{2\tau^2}\right)$$

dove $\mathbf{D}_h$ è una matrice diagonale con elementi sulla diagonale pari a $h_{i,+}$, che può anche essere scritta

$$f(\mathbf{w}) \propto \exp\left(-\frac{\sum_{j=1}^{n}\sum_{i=1}^{n} h_{i,j}(w_i - w_j)^2}{2\tau^2}\right)$$

Si vede facilmente che

$$(\mathbf{D}_h - \mathbf{H})\mathbf{1} = \mathbf{0} \Rightarrow \text{ la precisione è singolare } \Rightarrow \text{ la varianza non esiste}$$

Auto-Normal model (conditional autoregressive (CAR) model)

Questo è un problema se viene usata come verosimiglianza, ma se la si usa come "prior", allora si può fare fintanto che la a posteriori è ben definita $\Rightarrow$ la distribuzione si definisce **impropria**.

LA ragione per cui questo succede si può vedere dalla formalizzazione

$$f(\mathbf{w}) \propto \exp\left(-\frac{\sum_{i\neq j} h_{i,j}(w_i - w_j)^2}{2\tau^2}\right)$$

dato che possiamo aggiungere un valore c a tutte le variabili, e la densità rimane la stessa.

Questo si può risolvere (almeno in STAN) anche mettendo un vincolo del tipo che

$$\sum_i w_i \sim N(0, 0.0000001)$$

Il modello si chiama spesso con **ICAR (intrinsic CAR)**. Una soluzione al problema è' di ridefinire la precisione

$$(\mathbf{D}_h - \mathbf{H})$$

Auto-Normal model (conditional autoregressive (CAR) model)

come

$$(\mathbf{D}_h - \rho\mathbf{H})$$

e scegliendo ρ in modo tale che sia non singolare. Si può dimostrare che questo succedere se

$$\frac{1}{l_n} < \rho < \frac{1}{l_1}$$

dove l_1 e l_n sono il più piccolo e più grande autovalore di $\mathbf{D}_h^{-\frac{1}{2}}\mathbf{H}\mathbf{D}_h^{-\frac{1}{2}}$. Esistono anche altre formulazioni, ma non le vediamo

L'ICAR ha diversi vantaggi

- Non necessita il calcolo della precisione, ma è gia disponibile
- non ha parametri nella precisione

Prima di passare a qualche esempio, vediamo l'ultimo modello

Autologistico e Potts Model

Nel modello di Potts abbiamo che $y_i \in \{1, 2, \ldots, K\}$. L'idea alla base del modello è che la probabilità che $y_i = k$ cresce se i vicini sono uguali a k. Per questo modello partiamo direttamente dalla scrittura della congiunta in forma di distribuzione di Gibbs

$$f(\mathbf{y}) \propto \exp\left(\rho \sum_{\substack{i<j \\ i\tilde{j}}} \mathbf{1}(y_i = y_j)\right)$$

Questo modello si chiama **Potts Models**. IN questo modello, la costante di normalizzazioni è particolarmente complicata visto che è uguale a

$$\sum_{l_1=1}^{K} \sum_{l_2=1}^{K} \cdots \sum_{l_n=1}^{K} \exp\left(\rho \sum_{\substack{i<j \\ i\tilde{j}}} \mathbf{1}(l_i = l_j)\right)$$

Il caso in cui $K = 2$ è abbastanza famoso in fisica e si chiama **Ising Model**

Crime Data

Adesso stimiamo i seguenti modelli

- Modello 1

$$y_i|w_i \sim P(exp(\mu + w_i))$$
$$\mathbf{w} \sim ICAR()$$

- Modello 2

$$y_i|w_i \sim P(p_i exp(\mu + w_i))$$
$$\mathbf{w} \sim ICAR()$$

  dove p_i è la popolazione

- Modello 3

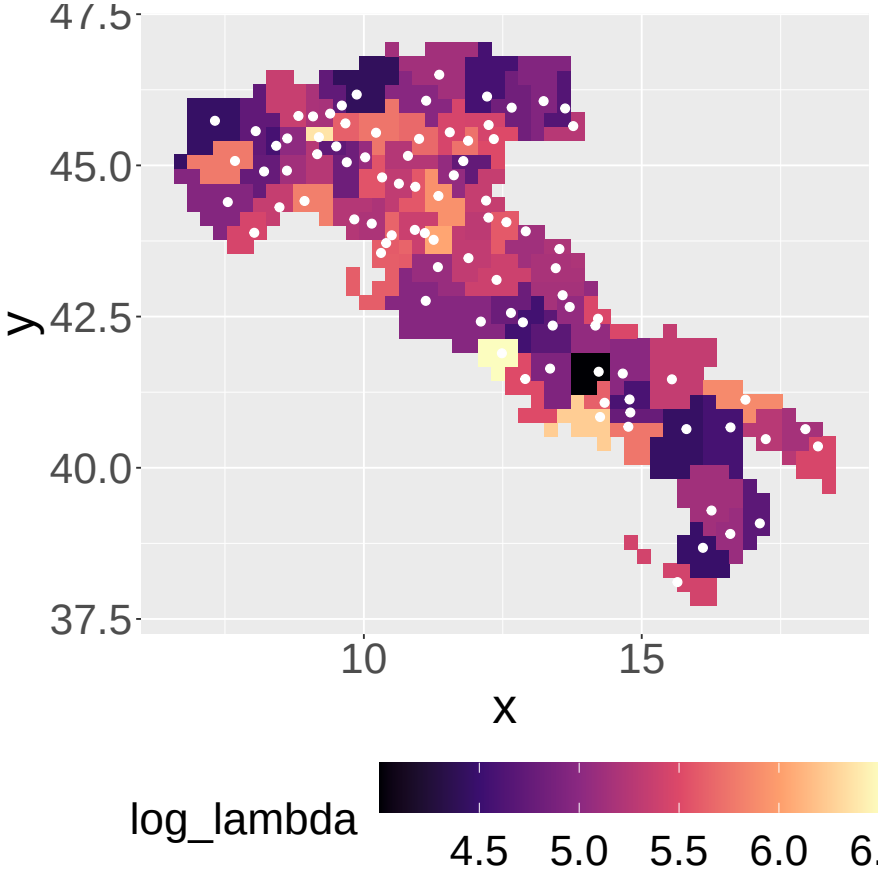
$$y_i|w_i \sim P(p_i exp(\mu + \beta s_2 + w_i))$$
$$\mathbf{w} \sim ICAR()$$

  dove s_2 è la latitudine

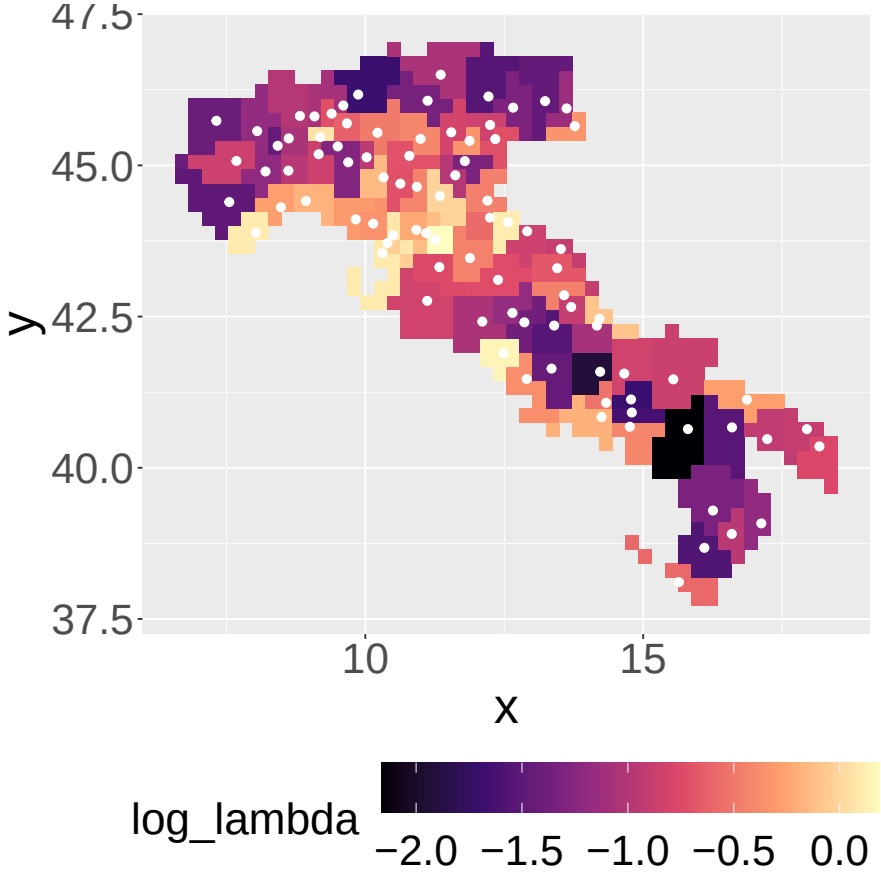
Crime Data

I modelli hanno, rispettivamente, WAIC pari a 785.0702, 787.9209 e 790.0982.

Mostriamo la stima di $E(\mathbf{y})$

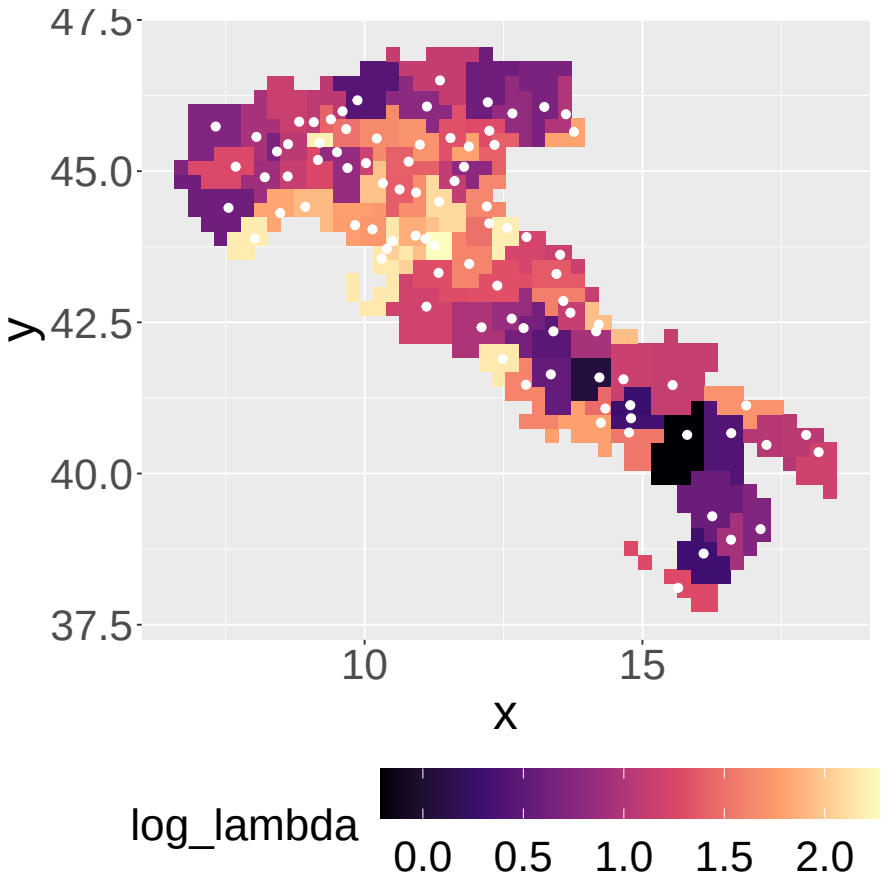


(a)



(b)

Crime Data



(c)