

Processi Gaussiani (GP)

Vers. 1.1.1

Gianluca Mastrantonio

email: gianluca.mastrantonio@polito.it

Introduzione

Un modello generale

Partiamo da una classe di modelli molto ampia. Ipotizziamo di avere una variabile y_i che segue qualche distribuzione, la cui media è modellizzata come

$$E(y_i) = \mu + \sum_{j=1}^p g_j(x_{i,j}) + \eta_i$$

dove μ rappresenta un valore media comune, g_j sono funzioni generali, che dipendono dal valore di una

- covariata
- tempo
- spazio
- etc..

η_i è un extra termine che contiene tutto ciò che non viene colto dalla restanti parti. Con questi tipi di modelli lo scopo è spesso quello di stimare le $g_j()$.

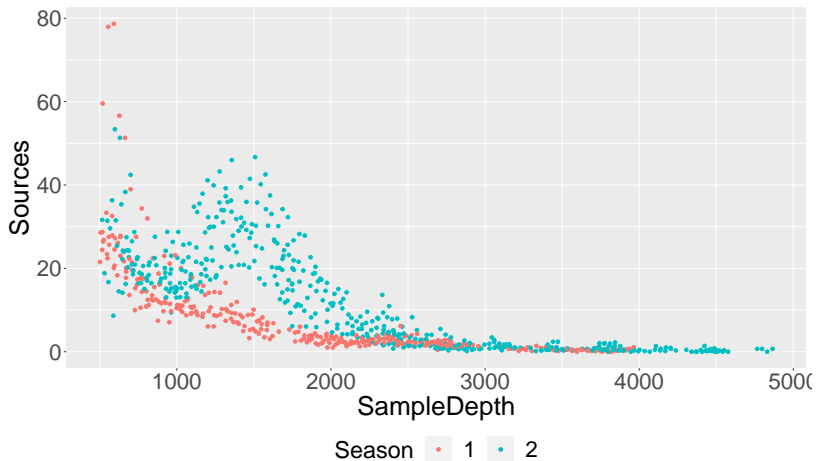
Per farlo si possono fare delle assunzioni, per esempio la linearità, ma in questo caso vogliamo

- mettere meno ipotesi possibili
- flessibili
- “facilmente” implementabile/stimabile

Esempio: nel data ISIT abbiamo il livello di bioluminescenza (Sources) in relazione alla stagione e profondità

(<https://r.qcbs.ca/workshop08/book-en/introduction-to-gams.html>)

Un modello generale



Almeno nella seconda stagione abbiamo una relazione complessa che vogliamo modellizzare. Formalizziamo il modello in maniera più precisa, ipotizzando si utilizzare solo i dati della stagione 2, e diciamo che:

- y_i sono i livelli di luminescenza
- la distribuzione è Normale (poco precisa \rightarrow provate a pensare come possiamo fare meglio)
- una sola covariata x_i che rappresenta la profondità (nel dataset ce ne sono molte)

per semplicità inoltre

- ordiniamo i dati in funzione della profondità (per semplificare l'implementazione)
- prendiamo un sottoinsieme dei dati (per velocizzare l'algoritmo)
- facciamo in modo che la profondità sia definita tra 0 e 1 (per motivi numerici)

Un modello generale

Il modello è

$$y_i = \mu + g(x_i) + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$ (!! ha davvero senso una varianza costante?). Se applichiamo una logica bayesiana, tutto ciò che non conosciamo è una variabile aleatoria, quindi, potremmo dire che $g()$ è una variabile aleatoria. Se vogliamo che questo abbia senso, dobbiamo dare a $g()$ qualche proprietà \Rightarrow per adesso richiediamo solo la continuità.

Al momento non siamo in grado di farlo, ma possiamo definire

$$w_i = g(x_i)$$

e mettere una distribuzione sul set delle w_i . La cosa più semplice è una distribuzione normale.

Il modello diventa un modello gerarchico, visto che

$$y_i | w_i \sim N(\mu + w_i, \sigma^2)$$

$$\mathbf{w} \sim N(\mathbf{0}, \Sigma)$$

Σ ha troppi parametri $((n-1)^2 + n)$, e non la possiamo stimare, anche perchè per stimare la relazione tra $[\Sigma]_{1,2}$ ho solo w_1, w_2 . Inoltre per far in modo il modello abbia senso, dobbiamo fare un ipotesi su Σ : più due punti sono vicini, più la covarianza deve essere forte (la correlazione deve tendere a 1). Il tutto ricordando che Σ deve essere definita positive.

Un modello generale

Il tutto diventa più facile se lavoro con le condizionate, quindi la distribuzione di w_1 , $w_2|w_1$, $w_3|w_1, w_2$. La mia idea è di definire il modello come

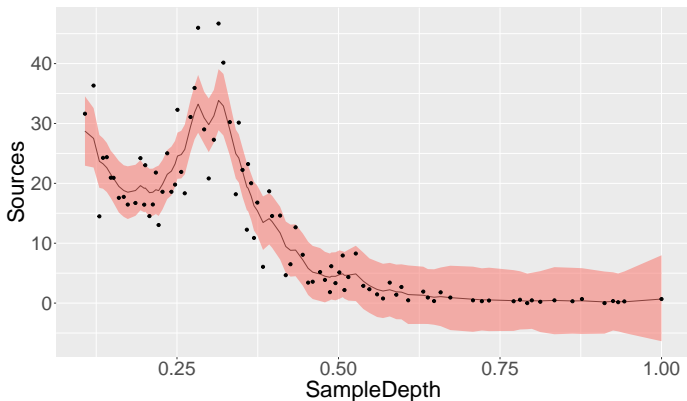
$$\begin{aligned}y_i|w_i &\sim N(\mu + w_i, \sigma^2) \\w_i|w_{i-1} &\sim N\left(\exp(-\phi|x_i - x_{i-1}|)w_{i-1}, \frac{(1 - \exp(-2\phi|x_i - x_{i-1}|))}{2\phi}\tau^2\right) \\w_1 &\sim N(0, 10000)\end{aligned}$$

Per completare la gerarchia del modello dobbiamo definire le prior, che sono

$$\sigma^2, \tau^2 \sim IG(1, 1) \quad \mu \sim N(0, 10000^2) \quad \phi \sim U(3/(\max \text{ dist}), 3/(\min \text{ dist}))$$

Proviamo a stimare il modello in STAN (il codice è nel materiale) e vediamo i risultati.

Un modello generale



In nero trovare la media a posteriori di $\mu + w_i$, e la zona colorata è il 95%CI. Il modello ha funzionato? problemi? soluzioni?

NB è possibile implementare l'MCMC di questo modello abbastanza facilmente.

Gaussian Processes

Riprendiamo i processi e parliamone più in generale. Un processo è una collezione di variabili aleatorie indicizzate, per esempio da $s \in \mathcal{S}$:

$$\{X(s) : s \in \mathcal{S}\}$$

che spesso viene anche indicato come

$$\{X_s : s \in \mathcal{S}\}$$

In questa sezione assumeremo che $\mathcal{S} \in \mathbb{R}^d$ ma non è richiesto in generale e altri spazi si possono usare. Il processo X_s può avere valori in \mathbb{R}^c , \mathbb{Z} etc

Il processo viene totalmente descritto dalle distribuzioni **finito-dimensionali**, che sono la distribuzioni di tutti i possibili set di $n < \infty$ osservazioni del processo.

Il processo Gaussiano è un processo le cui distribuzioni finito-dimensionali sono gaussiane. Per specificare il processo abbiamo bisogno della

- funzione media $m(\cdot)$;
- funzione covarianza $C(\cdot, \cdot)$

e si scrive

$$X(\mathbf{s}) \sim GP(m(\cdot), C(\cdot, \cdot))$$

Se assumiamo di aver osservato n punti del processo nelle coordinate $\mathbf{s}_1, \dots, \mathbf{s}_n$, e definiamo $x_j = x(\mathbf{s}_j)$, allora abbiamo che

$$\mathbf{x} = (x_1, \dots, x_n)' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

dove

$$[\boldsymbol{\mu}]_j = m(\mathbf{s}_j) \quad [\boldsymbol{\Sigma}]_{jk} = C(\mathbf{s}_j, \mathbf{s}_k)$$

Prima di continuare vediamo qualche proprietà dei processi stocastici.

Definizione - Stazionarietà forte

Un processo si dice stazionario in senso forte se per ogni n naturale, e ogni n -upla $(s_1, s_2, \dots, s_n) \in \mathcal{S}^n$, e per ogni scelta di \mathbf{h} tale che $s_i + \mathbf{h} \in \mathcal{S}$, per ogni $i = 1, 2, \dots, n$, abbiamo

$$\begin{aligned} f(X(s_1) = x_1, X(s_2) = x_2, \dots, X(s_n) = x_n) = \\ f(X(s_1 + \mathbf{h}) = x_1, X(s_2 + \mathbf{h}) = x_2, \dots, X(s_n + \mathbf{h}) = x_n) \end{aligned}$$

Abbiamo quindi che la distribuzione non cambia se la trasliamo di \mathbf{h} .

Definizione - Stazionarietà debole

Un processo si dice stazionario in senso debole di ordine $c \in \mathbb{N}$ se per ogni $n \leq c$ naturale, e ogni n -upla $(s_1, s_2, \dots, s_n) \in \mathcal{S}^n$, e per ogni scelta di \mathbf{h} tale che $s_i + \mathbf{h} \in \mathcal{S}$, abbiamo che per ogni $i = 1, 2, \dots, n$,

$$E(X(s_1)X(s_2) \dots X(s_n)) = E(X(s_1 + \mathbf{h})X(s_2 + \mathbf{h}) \dots X(s_n + \mathbf{h})) < \infty$$

Nel caso in cui il processo sia Gaussiano, se è stazionario debole di ordine 2 è stazionario forte visto che

$$E(X(\mathbf{s}_j)) = E(X(\mathbf{s}_j + \mathbf{h}))$$

$$E(X(\mathbf{s}_j)X(\mathbf{s}_j)) = E(X(\mathbf{s}_j + \mathbf{h})X(\mathbf{s}_j + \mathbf{h}))$$

segue che

$$Cov(X(\mathbf{s}_j), X(\mathbf{s}_j)) = Cov(X(\mathbf{s}_j + \mathbf{h}), X(\mathbf{s}_j + \mathbf{h}))$$

In generale $m(s) = 0$ (poi capiremo perchè), ma la funzione di covarianza è più difficile da trattare. Questo perchè Σ deve essere definita positiva $\Rightarrow C()$ deve essere una funzione definita positiva. Spesso, invece delle funzioni di covarianza, viene specificata la funzione di correlazione.

Assumiamo che

- siamo in uno spazio, in cui è possibile calcolare una distanza scalare (e.g., se s rappresenta il tempo, lo spazio **Euclideo**, o qualcosa di affine, per esempio una covariate)
- assumiamo stazionarietà

e definiamo

$$h_{jk} = \mathbf{s}_j - \mathbf{s}_k \in \mathbb{R}^+$$

allora possiamo usare le classiche scelte che sono disponibili in letteratura:

- Spherical $\begin{cases} \sigma^2(1 - \frac{3h}{2\phi} + \frac{h^3}{2\phi^3}) & \text{if } 0 < h \leq \phi \\ 0 & \text{if } h > \phi \end{cases}$
- Powered Exponential $\sigma^2 \exp(-|\frac{h}{\phi}|^\kappa)$ $0 < \kappa \leq 2$
- Exponential $\sigma^2 \exp(-\frac{h}{\phi})$
- Rational Quadratic $\sigma^2 \left(1 - \frac{h^2}{\phi^2 + h^2}\right)$
- Matèrn $\frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)} (h/\phi)^\kappa K_\kappa(h/\phi)$ $\kappa > 0$
- Cauchy $\sigma^2 (1 + (h/\phi)^2)^{-k}$, $k \in [0.5, 2]$

In tutte queste funzioni σ^2 è la varianza e se la mettete = 1 queste diventano funzioni di correlazione. Quando la funzione di covarianza dipende solo dalla distanza, si può scrivere

$$C(\mathbf{s}_i, \mathbf{s}_k) \equiv C(h_{jk})$$

P.S. Nell'esponenziale, ma anche in altre, si preferisce lavorare con $\lambda = 1/\phi$ che viene chiamato "decay", mentre ϕ è il "range"

Notate bene: Queste funzioni di covarianza sono valide su spazi euclidei, sulla sfera, per esempio, non lo sono più (ci sono delle funzioni di covarianza definite sulla sfera).

Se invece $s = (\mathbf{a}, t)$, dove \mathbf{a} è lo spazio Euclideo e t il tempo, potremmo ancora usarle, ma assumiamo che una distanza di 1 nello spazio sia uguale a una distanza di 1 nel tempo, che è un'assunzione forte. Una soluzione è definire due funzioni di Correlazione per lo spazio ($C_s()$) e il tempo ($C_t()$) e definire la funzione di covarianza come

$$C(\mathbf{s}_j, \mathbf{s}_k) = aC_s(\|\mathbf{a}_j - \mathbf{a}_k\|) + bC_t(|t_j - t_k|)$$

visto che somma di funzioni definite positive è ancora definita positiva. Notate che se definite

$$W(t) \sim GP(0, C_t(,)) \quad Z(\mathbf{a}) \sim GP(0, C_s(,))$$

allora

$$X(\mathbf{s}) = \sqrt{a}W(t) + \sqrt{b}Z(\mathbf{a})$$

il processo ha esattamente la funzione di covarianza $aC_t(|t_j - t_k|) + bC_s(\|\mathbf{a}_j - \mathbf{a}_k\|)$. Fate attenzione che tutto questo ha senso se il processo non ha ripetizioni nelle coordinate perchè in questo caso, per esempio, ci potrebbero essere due punti tra gli osservati di $W(t)$ che hanno correlazione 1. Una possibile soluzione è usare

$$C(\mathbf{s}_j, \mathbf{s}_k) = a(C_s(\|\mathbf{a}_j - \mathbf{a}_k\|) + bC_t(|t_j - t_k|)) + dC_s(\|\mathbf{a}_j - \mathbf{a}_k\|)C_t(|t_j - t_k|)$$

che è sempre definita positiva fintanto che $\mathbf{s}_j \neq \mathbf{s}_k$ se $j \neq k$.

In tutti e due i casi stiamo assumendo **separabilità** tra lo spazio e il tempo, che non sempre è ragionevole.

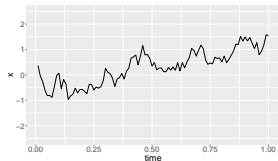
Esistono (poche) alternative, per esempio la funzione di covarianza di Gneiting

$$C(\mathbf{s}_j, \mathbf{s}_k) = \frac{\sigma^2}{(\phi_t |t_j - t_k|^{2\alpha} + 1)^\tau} \exp \left(-\frac{\phi_s \|\mathbf{a}_j - \mathbf{a}_k\|^{2\gamma}}{(\phi_t |t_j - t_k|^{2\alpha} + 1)^{\eta\gamma}} \right)$$

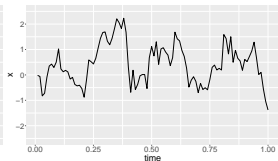
con $\phi_t, \phi_s > 0$, $\alpha, \gamma \in (0, 1]$, $\eta \in [0, 1]$ e $\tau > 1$.

Gaussian Processes

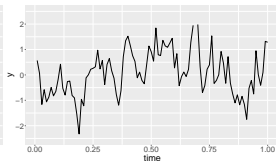
Vediamo qualche simulazione da un processo definito sul tempo, con una funzione di covarianza esponenziale (prima riga) e Cauchy (seconda)



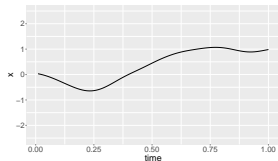
(a) $\lambda = 3$



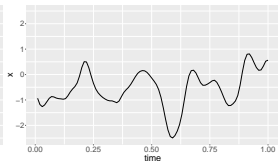
(b) $\lambda = 15$



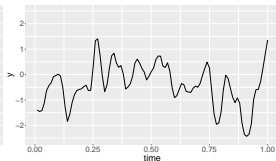
(c) $\lambda = 30$



(d) $\lambda = 3 \quad k = 1$

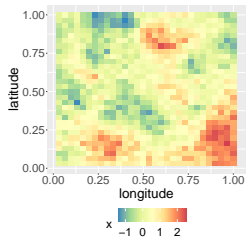


(e) $\lambda = 15 \quad k = 1$

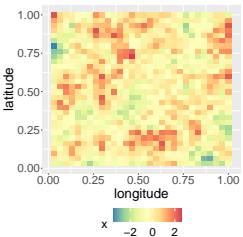


(f) $\lambda = 30 \quad k = 1$

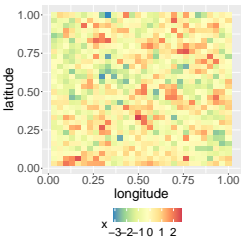
Possiamo vedere anche dei processi spaziali, con covarianza esponenziale



(g) $\lambda = 3$

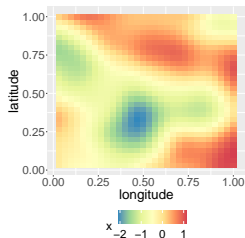


(h) $\lambda = 15$

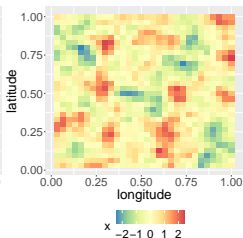


(i) $\lambda = 30$

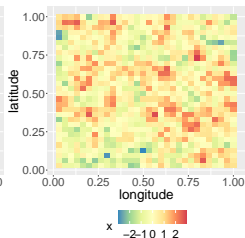
Possiamo vedere anche dei processi spaziali, con covarianza esponenziale Cauchy



(j) $\lambda = 3$



(k) $\lambda = 15$



(l) $\lambda = 30$

Il più grande problema dei processi spaziali è che la densità di un set di osservazioni è

$$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{\mathbf{x}'\Sigma^{-1}\mathbf{x}}{2}\right)$$

e il calcolo dell'inversa è molto complesso dal punto di vista computazionale.

Applicazione

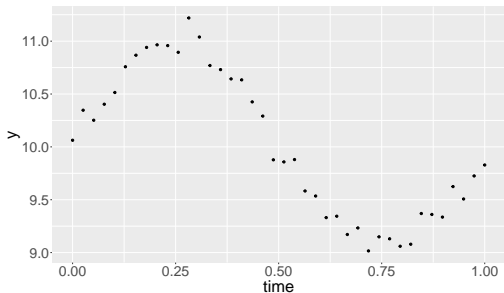
Simulazioni

Iniziamo con dei dati simulati. Le simulazioni seguono il modello

$$y(t) \stackrel{ind}{\sim} GP(\sin(t * 2 * \pi), 0.1^2)$$

e simuliamo $n = 40$ dati equispaziati tra $t \in [0, 1]$.

Mostriamo i dati

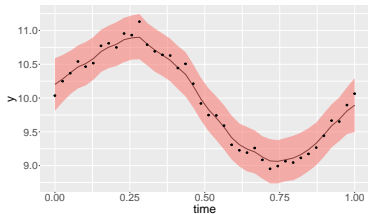


Immaginiamo di non sapere la funzione media che ha generato i dati, e assumiamo un modello

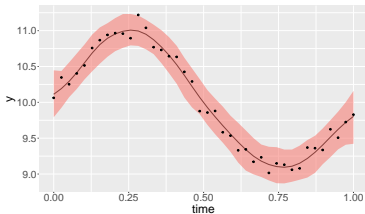
$$y_i | \mathbf{w} \sim N(\mu + w_i, \sigma^2)$$

$$w(t) \sim GP(0, C(,))$$

Come covarianze usiamo un'esponenziale e una Cauchy. Stimiamo il modello e i risultati sono i seguenti



(m) Esponenziale



(n) Cauchy

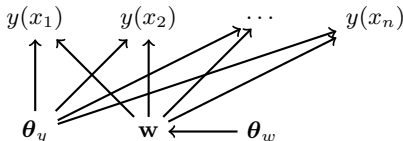
Adesso possiamo modellizzare i dati visti all'inizio con il seguente modello

$$y(x)|w(x) \stackrel{ind}{\sim} GP(\mu + w(x), \sigma^2)$$
$$w(x) \sim GP(0, C(,))$$

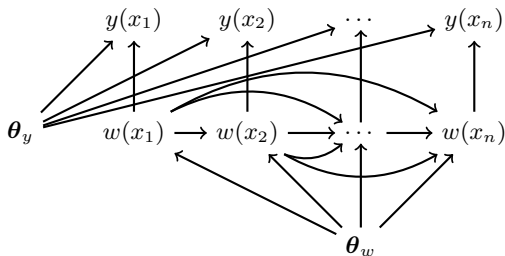
Il primo livello della gerarchia si può scrivere anche come

$$y(x) = \mu + w(x) + \epsilon(x)$$

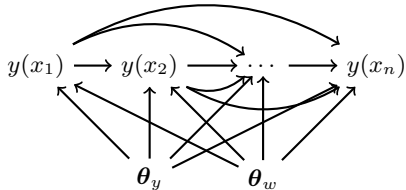
Anche in questo caso, proviamo la covarianza esponenziale. Questo modello, come tutti quelli che seguono, hanno una struttura del tipo



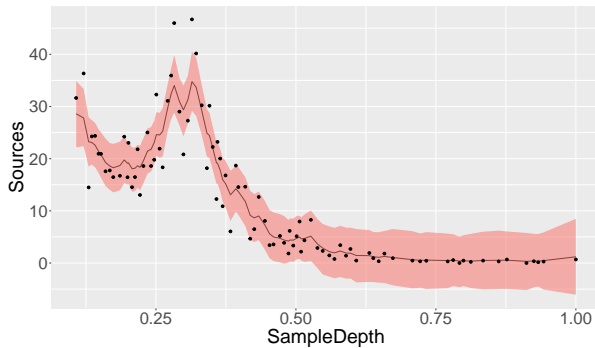
se indico con θ_w il vettore dei parametri del processo e θ_y quello delle osservazioni. Che può essere scritta anche come



Se marginalizziamo rispetto al processo, otteniamo



Vediamo qualche risultato



(o) Esponenziale

Da ora in poi, visto che l'inversione della matrice di covarianza è computazionalmente pesante, stimeremo modelli con i parametri della covarianza fissi, se volete ne potete testare diversi e poi decidere quale usare usando il WAIC o altri indici.

Proviamo 3 modelli

- Modello 1:

$$y(x)|w(x) \stackrel{ind}{\sim} GP(\mu + w(x), \sigma^2)$$
$$w(x) \sim GP(0, C(,))$$

con covarianza esponenziale $\tau^2 \exp(-dist)$

- Modello 2: varianza non costante

$$y(x)|w(x) \stackrel{ind}{\sim} GP(\mu + w(x), \sigma^2 \exp(-\eta x))$$
$$w(x) \sim GP(0, C(,))$$

con covarianza esponenziale

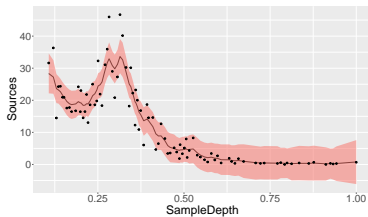
- Modello 3: con distribuzione troncata

$$y(x)|w(x) \stackrel{ind}{\sim} GP(\mu + w(x), \sigma^2 \exp(-\eta x)) I_{>0}$$
$$w(x) \sim GP(0, C(,))$$

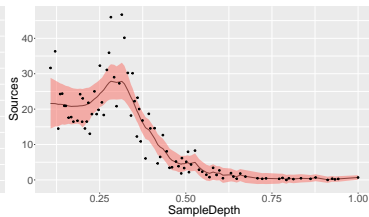
con covarianza esponenziale. La prima riga serve a dire che il GP è troncato (i dati seguono una normale troncata a 0)

Dopo aver lanciato il modello, possiamo calcolare il WAIC dei 3 modelli che sono 558.948, 490.3879, 437.4644 e quindi il terzo è il migliore. Le stime di $\mu + w(x)$ sono riportate sotto

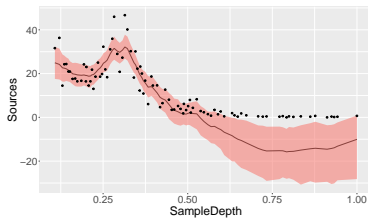
Applicazione



(p) Modello 1

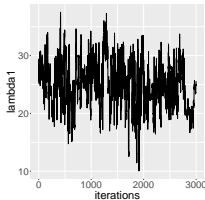


(q) Modello 2

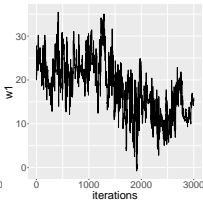


(r) Modello 3

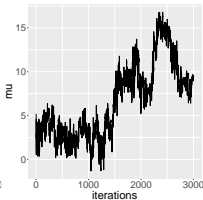
Vediamo i plot del modello migliore



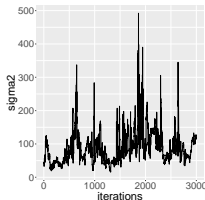
(s) $\mu + w(1)$



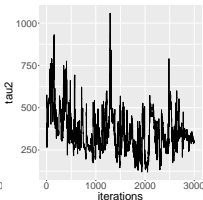
(t) $w(1)$



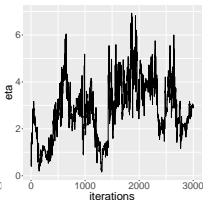
(u) μ



(v) σ^2



(w) τ^2



(x) η

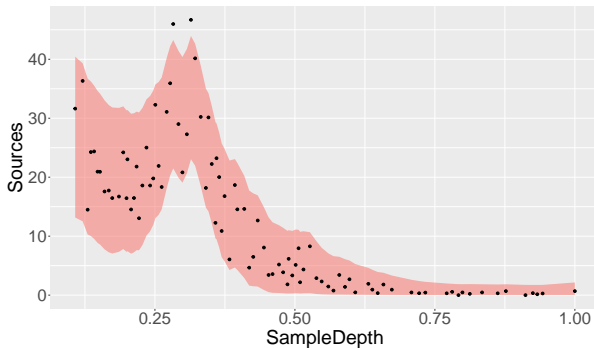
Dalle figure vediamo che forse servirebbe qualche iterazione in più, anche se i plot della seconda riga non sono male.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
μ	9.35	0.81	2.86	2.90	7.77	9.05	10.80	16.73	12.45	1.01
η	2.38	0.39	1.10	0.60	1.62	2.23	2.97	4.78	7.78	1.06
τ^2	547.61	42.63	193.94	226.72	411.15	517.47	673.49	1006.02	20.70	1.00
σ^2	75.88	9.34	37.12	28.73	50.08	67.88	91.45	172.71	15.80	1.04

Per capire meglio come il modello si sta adattando ai dati, possiamo calcolare l'intervallo che contiene il 95% della probabilità della normale troncata. Indichiamo con $\boldsymbol{\theta}$ tutti i parametri del modello e con $q_{nt}(p, \boldsymbol{\theta}, \mathbf{w})$ la funzione quantilica della normale troncata di probabilità p , allora cerchiamo il seguente valore atteso

$$\int q_{nt}(p, \boldsymbol{\theta}, \mathbf{w}) f(\boldsymbol{\theta}, \mathbf{w} | \mathbf{y}) d\boldsymbol{\theta} d\mathbf{w} \approx \frac{\sum_{b=1}^B q_{nt}(p, \boldsymbol{\theta}^b, \mathbf{w}^b)}{B}$$

Utilizzando quest'approccio abbiamo la seguente figura che rappresenta l'intervallo al 95% centrale



Attenzione: questa figura non è simile a quella di prima, visto che questo rappresenta la distribuzione dei dati mentre quella di prima la distribuzione di λ

Ipotizziamo di aver scelto il modello 2. In questo caso è molto facile conoscere il valore, o meglio la sua distribuzione predittiva, del processo temporale in una posizione x^* non osservato, visto che, se indichiamo con θ tutti i parametri del modello, abbiamo che

$$f(w(x^*)|\mathbf{y}) = \int f(w(t^*)|\theta, \mathbf{w}, \mathbf{y})f(\theta, \mathbf{w}|\mathbf{y})dw(t^*) = \int f(w(t^*)|\theta, \mathbf{w})f(\theta, \mathbf{w}|\mathbf{y})dw(t^*)$$

visto che (fate il DAG) $w(t^*)$ è indipendente da \mathbf{y} dato \mathbf{w} (condizionatamente indipendenti)

$$\begin{pmatrix} \mathbf{w} \\ w(x^*) \end{pmatrix} | \theta \sim N \left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_w & \Sigma'_{w,w^*} \\ \Sigma_{w,w^*} & \tau^2 \end{pmatrix} \right)$$

dove le varie matrici sono calcolate utilizzando la funzione di covarianza e la distanza temporale

$$[\Sigma_w]_{j,k} = \tau^2 \exp(-\phi|x_j - x_k|) \quad [\Sigma_{w,w^*}]_j = \tau^2 \exp(-\phi|x_j - x^*|)$$

Usando le regole per la distribuzione condizionata di una normale, possiamo facilmente trovare la distribuzione $f(w(t^*)|\boldsymbol{\theta}, \mathbf{y})$

$$w(t^*)|\boldsymbol{\theta}, \mathbf{w} \sim N(\boldsymbol{\Sigma}_{w,w^*} \boldsymbol{\Sigma}_w^{-1} \mathbf{w}, \tau^2 - \boldsymbol{\Sigma}_{w,w^*} \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\Sigma}'_{w,w^*})$$

Quindi, il campione b -esimo da $f(w(x^*)|\mathbf{y})$ si può ottenere prendendo il campione $\boldsymbol{\theta}^b$ e simulando da $w(t^*)|\boldsymbol{\theta}^b, \mathbf{y}$.

Possiamo anche simulare l'osservazione sullo stesso punto

$$f(y(x^*)|\mathbf{y}) = \int f(y(t^*)|w(t^*), \boldsymbol{\theta}, \mathbf{y}), f(w(t^*)|\boldsymbol{\theta}, \mathbf{y}) f(\boldsymbol{\theta}|\mathbf{y}) d\mathbf{w}(t^*)$$

simulare da questa è facile, una volta che abbiamo campioni di $w(t^*)$ visto che

$$f(y(t^*)|w(t^*), \boldsymbol{\theta}, \mathbf{y}) \equiv f(y(t^*)|w(t^*), \boldsymbol{\theta})$$

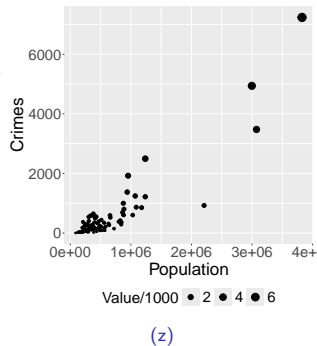
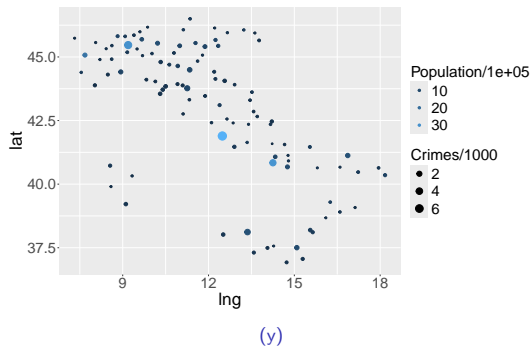
e da modello

$$y(t^*)|w(t^*), \boldsymbol{\theta} \sim N(\mu + w(t^*), \sigma^2 \exp(-\eta t^*))$$

Nel caso in cui la verosimiglianza è normale troncata, i calcoli non sono facili, ma se volte potete ottenere queste quantità direttamente durante il model fitting.

Esempio - Poisson spaziale

Abbiamo il numero di “furti di ciclomotori” commessi in Italia, nel 2006 (ci sono anche altri anni nel dataset e altri crimini), accorpati per provincia.



Se vogliamo dare una struttura spaziale, possiamo definire il modello come

$$y(\mathbf{s}) \sim P(\lambda(\mathbf{s}))$$

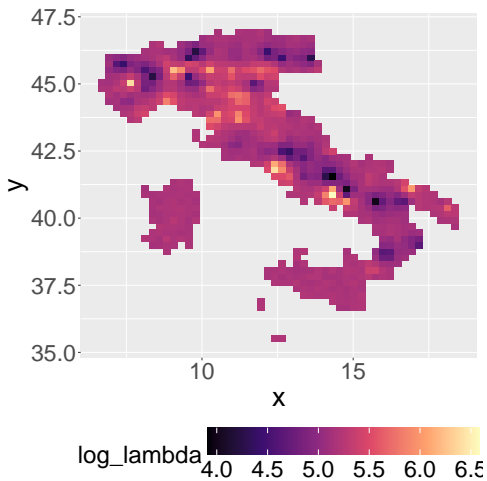
$$\log(\lambda(\mathbf{s})) = \mu + w(\mathbf{s})$$

$$W(\mathbf{s}) \sim GP(0, \sigma^2 \exp(-h\phi))$$

Usiamo le prior $\mu \sim N(0, 1000^2)$ e $\phi \sim U(0.3, 30)$ e $\sigma^2 \sim IG(1, 1)$.

Per la stima del modello elimino le isole, e poi, per vedere i risultati faccio la stima di $\log(\lambda)$ su una griglia regolare sovrainposta all'Italia. La previsione del processo si fa facilmente (come esercizio verificate come si ottengono campioni di λ . è simile al calcolo di $w(x)$ fatto prima)

Esempio - Poisson spaziale



In questo modello non teniamo conto della popolazione, che può essere una spiegazione del fenomeno. Quando in un modello Poisson si tiene conto di quante unità è basato un conteggio, si parla di **offset**. Se indichiamo con $p(\mathbf{s})$ la popolazione, quello che richiediamo è che

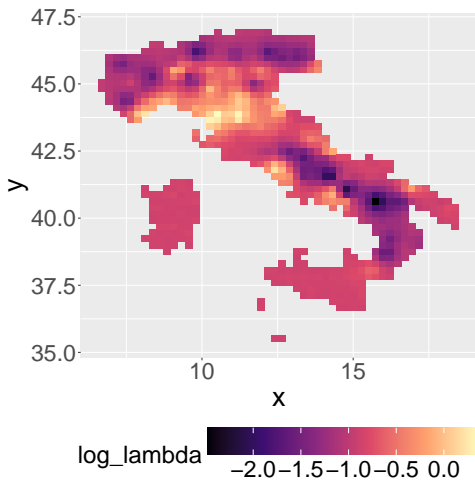
$$E\left(\frac{y(\mathbf{s})}{p(\mathbf{s})}\right) = \exp(\mu + w(\mathbf{s}))$$

e quindi il modello diventa

$$\begin{aligned}y(\mathbf{s}) &\sim P(p(\mathbf{s}) \exp(\mu + w(\mathbf{s}))) \\ \log(\lambda(\mathbf{s})) &= \log p(\mathbf{s}) + \mu + w(\mathbf{s}) \\ W(\mathbf{s}) &\sim GP(0, \sigma^2 \exp(-h\phi))\end{aligned}$$

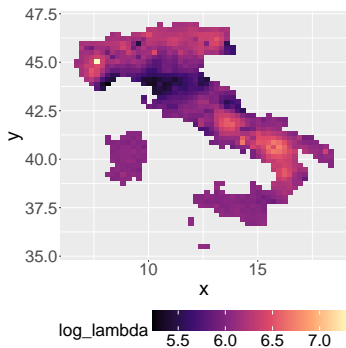
Usando le stesse prior di prima otteniamo

Esempio - Poisson spaziale



Esempio - Poisson spaziale

Da un punto di vista prettamente numerico, i due modelli sono veramente simili, dato che i WAIC sono rispettivamente 785.8078 e 786.1524, ma se facciamo la differenza tra i $\log(\lambda)$ stimate dai due modello ($\lambda(s)_{mod1} - \lambda(s)_{mod2}$)



vediamo che il primo modello ha valori "elevati al sud", mentre se teniamo conto della popolazione il livello medio decresce. Inoltre il primo modello ha rischio basso in Romagna.

Nel modello precedente stiamo assumendo che

$$\log(E(y(\mathbf{s}))) = \log(p(\mathbf{s})) + \mu + w(\mathbf{s})$$

e quindi i conteggi aumentano proporzionalmente alla popolazione. Questa è un'ipotesi, che potremmo considerare molto forte. Possiamo testare la proporzionalità modificando il modello con

$$\log(E(y(\mathbf{s}))) = \beta \log(p(\mathbf{s})) + \mu + w(\mathbf{s})$$

Se la a posteriori (il suo CI) di β contiene l'1, abbiamo evidenza che ci sia proporzionalità, altrimenti possiamo vedere se l'effetto della popolazione fa incrementare o diminuire i conteggi rispetto alla proporzionalità. Il modello diventa

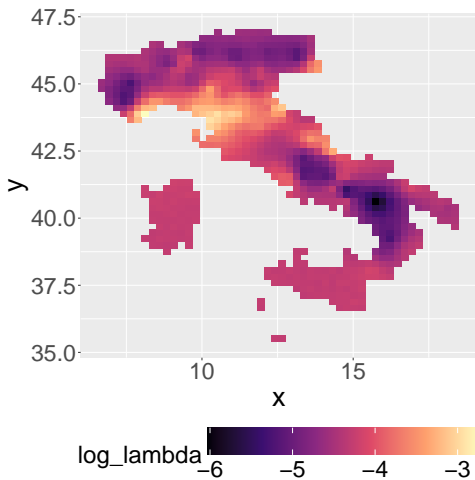
$$y(\mathbf{s}) \sim P(p(\mathbf{s})^\beta \exp(\mu + w(\mathbf{s})))$$

$$\log(\lambda(\mathbf{s})) = \beta \log(p(\mathbf{s})) + \mu + w(\mathbf{s})$$

$$W(\mathbf{s}) \sim GP(0, \sigma^2 \exp(-h\phi))$$

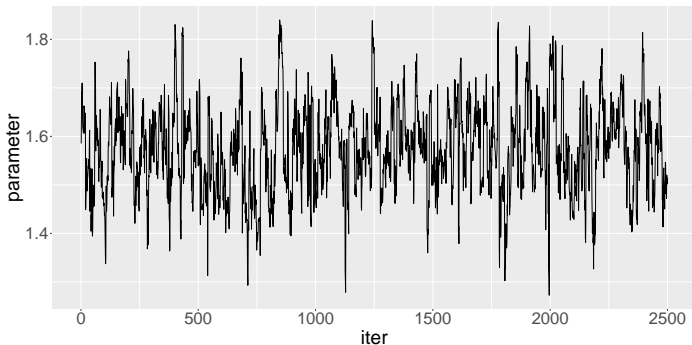
dove mettiamo una prior su $\beta \sim N(0, 10000)$

Esempio - Poisson spaziale



Esempio - Poisson spaziale

LA stima di λ è più liscia e se guardiamo il WAIC abbiamo che è 785.6504 che è più basso dei precedenti. In termine di distribuzione di β abbiamo



Esempio - Poisson spaziale

La distribuzione è concentrata su valori maggiori di 1, che ci dicono che la popolazione ha un effetto “esponenziale” sul numero di crimini

Chiudiamo questa sezione chiedendoci se veramente questo è il modello giusto, visto che piuttosto che un processo che cambia in maniera continua nello spazio (s), ha forse più senso un modello sullo spazio discreto (visto che stiamo modellizzando le provincie), e la dipendenza deve essere basata su una nozione di vicini piuttosto che la distanza.

Nearest Neighbour Gaussian Proces NNGP

Ci sono vari modi per cercar di risolvere i problemi computazionali dei processo gaussiani. Quello che funziona meglio (e si può implementare in STAN) è il NNGP. L'idea è molto facile, se avete un processo gaussiano osservato \mathbf{w} e ne volete calcolare la densità, avete che

$$f(\mathbf{w}) = f(w_1) \prod_{i=2}^n f(w_i | w_{i-1}, w_{i-2}, \dots, w_1)$$

se le componenti di \mathbf{w} sono ordinate in funzione della vicinanza (punti spazialmente vicini sono vicini anche negli indici), e visto che se due punti sono lontani tendono a essere indipendenti, allora

$$f(w_1) \prod_{i=2}^n f(w_i | w_{i-1}, w_{i-2}, \dots, w_1) \approx f(w_1) \prod_{i=2}^n f(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-c})$$

con c piccolo. Di fatto stiamo rimuovendo archi dal DAG. Questo permette di calcolare le full conditional in maniera veloce, e è accurato anche per $c \approx 10$. NNGP si può implementare in STAN

(<https://mc-stan.org/users/documentation/case-studies/nngp.html>)

