Modelli Grafici

Vers. 1.0.0

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Il Modello Statistico e Sua Rappresentazione Grafica

Modello Statistico

Il modello statistico:

- è un modello matematico che incorpora un insieme di assunzioni riguardanti la generazione di dati campionari
- è solitamente specificato come una relazione matematica tra una o più variabili aleatorie e altre variabili non aleatorie (parametri e/o iperparametri).
- rispetto a un modello matematico, il modello statistico è non deterministico
- è (o può essere) "una rappresentazione formale di una teoria".

Modello Statistico

L'esempio più semplice, e forse banale, è la regressione

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \epsilon_i$$

con $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ e le $x_{i,.}$ sono fisse.

Possibile estensioni

- le $x_{i,.}$ non sono fisse ma variabile aleatorie
- le ϵ_i non sono iid, ma $\epsilon_i \sim N(0, \sigma_i^2)$
- σ_i^2 proviene da una distributione dipendente da altre variabili

Modello Statistico

Da un punto di vista formale, possiamo usare la seguente definizione

Definizione - Modello Statistico

Un modello statistico è una coppia $(S, \mathcal{P})$, dove S è l'insieme delle possibili osservazioni, cioè lo spazio campionario, e $\mathcal{P}$ è un insieme di distribuzioni di probabilità su S. L'insieme $\mathcal{P}$ rappresenta tutti i modelli che sono considerati possibili.

L'insieme $\mathcal{P}$ è generalmente parametrizzato come $\mathcal{P} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, dove

- F_θ è una congiunta (discreta, continua, o mista)
- $\boldsymbol{\theta}$ sono i parametri del modello (la cui stima e interpretazione è generalmente l'oggetto di interesse)

Modello Statistico

Definizione - Identificabilità

Se una parametrizzazione è tale che valori distinti dei parametri danno origine a distribuzioni distinte, cioè $F_{\theta_1} = F_{\theta_2} \Rightarrow \theta_1 = \theta_2$ (in altre parole, la mappatura è iniettiva), si dice che il modello è identificabile.

I modelli devono essere identificabili (anche se in alcuni casi possiamo farne a meno)
L'identificabilità è il motivo per cui in una regressione ϵ_i ha media nulla

Una classe di modello molto utilizzata sono i **modelli gerarchici**

Modello Statistico

Modello Gerarchico

Un **modello gerarchico** in statistica è un modello che incorpora più livelli di parametri, riflettendo una struttura annidata o multilivello nei dati.
I parametri a un livello inferiore dipendono da quelli a un livello superiore, creando una gerarchia di variabili casuali.
Questi modelli sono utili per analizzare dati con strutture complesse, come dipendenze tra gruppi o cluster.

Una classica struttura di modello gerarchico ha 3 livelli

Livello 1: Dati Osservati

$$\mathbf{y}|\mathbf{w}, \boldsymbol{\theta}_y \sim F_y(\mathbf{w}, \boldsymbol{\theta}_y)$$

Livello 2: Variabili latenti/Processi/livello di Gruppo

$$\mathbf{w}|\boldsymbol{\theta}_w \sim F_w(\boldsymbol{\theta}_w)$$

Livello 3: Iperparametri/Parametri Globali

$$(\boldsymbol{\theta}_y, \boldsymbol{\theta}_w) \sim F_\theta(\boldsymbol{\theta}_)$$

Indipendenza Condizionata

Se due variabili A e B sono indipendenti ($A \perp B$) ho che

$$f(A, B) = f(A)f(B)$$

Se due variabili A e B sono condizionatamente indipendenti, dato una variabile Z ($(A \perp B)|Z$ oppure $A|Z \perp B|Z$) significa che

$$f(A, B|Z) = f(A|Z)f(B|Z)$$

questo però non implica che A e B siano indipendenti marginalmente.

Indipendenza Condizionata

Facciamo un piccolo esempio

Indipendenza condizionata

Ipotizziamo che $Z = A + W_1$ e $B = A + W_2$ con

$$A \sim Bern(0.5) \quad W_1 \sim Bern(0.5) \quad W_2 \sim Bern(0.5)$$

Abbiamo che la distribuzione di B dipende da Z, visto che se

$$P(B = 0|Z = 0) = 0.5 \quad P(B = 1|Z = 0) = 0.5 \quad P(B = 3|Z = 0) = 0$$

$$P(B = 0|Z = 1) = \sum_{i=0}^{1} P(B = 0|A = i, Z = 1)P(A = i|Z = 1) = 0.5 \times 0.5 + 0 = 0.25$$

$$P(B = 1|Z = 1) = \sum_{i=0}^{1} P(B = 1|A = i, Z = 1)P(A = i|Z = 1) = 0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$$

$$P(B = 2|Z = 1) = \sum_{i=0}^{1} P(B = 2|A = i, Z = 1)P(A = i|Z = 1) = 0 + 0.5 \times 0.5 = 0.25$$

Indipendenza Condizionata

Ma se condiziono al valore di A diventano indipendenti:

$$P(B = A|A, Z) = 0.5 \qquad P(B = A + 1|A, Z) = 0.5$$

DAG

DAG

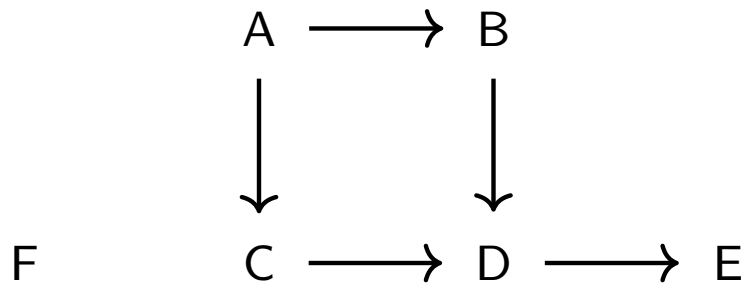


Figure: Esempio di un DAG

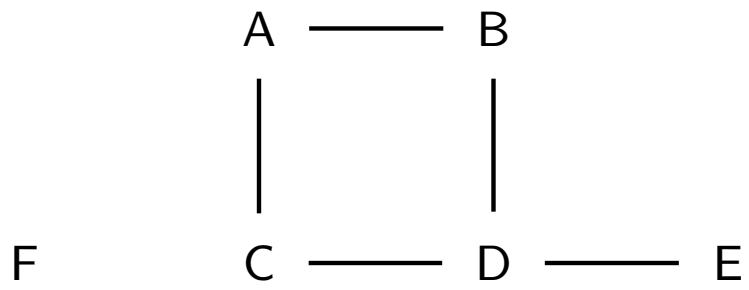
Un Grafo Aciclico Diretto (DAG) è un grafo che è:

- **Diretto:** Tutti gli archi hanno una direzione, puntando da un nodo a un altro.
- **Aciclico:** Non ci sono cicli, il che significa che è impossibile partire da un nodo e seguire un percorso che riporti allo stesso nodo.

I DAG

- sono strumenti potenti per rappresentare e analizzare le dipendenze in un modello.
- Aiutano nella formalizzazione chiara e sistematica dei modelli.

DAG



Partiamo da un grafo non-diretto, costruito assumendo che

- ogni nodo è una variabile aleatoria (anche multivariata)
- c'è dipendenza marginale tra due nodi se c'è un percorso che li connette
- due nodi N_1 e N_2 sono condizionatamente indipendenti dato un set di nodi $\mathbf{M}$ intermedi, se non è' possibile trovare un percorso tra N_1 e N_2 che non passi per un nodo in $\mathbf{M}$

in questo caso abbiamo che (un subset delle condizioni)

$$F \perp A \quad F \perp A, B, C, D, E \quad A \not\perp C \quad A \not\perp D$$

$$A|C \not\perp D|C \quad A|B, C \perp D|B, C \quad E|D \perp A, B, C|D$$

DAG

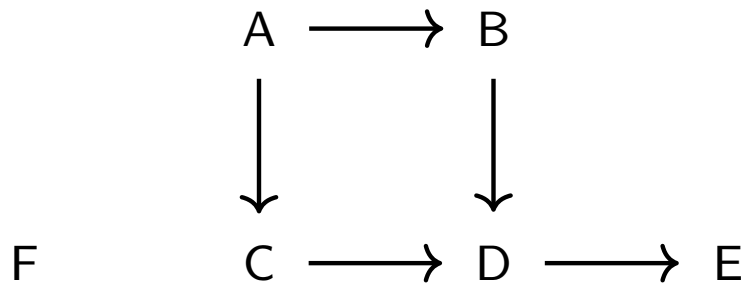


Figure: Esempio di un DAG

Il nostro scopo è definire una congiunta, che in questo caso è

$$f(A, B, C, D, E, F)$$

che abbia la struttura di dipendenza definita dal grafo non diretto che può potenzialmente dipendere da dai parametri.
Partendo dal grafo non-diretto

- troviamo un DAG, che si può sempre trovare, ma non è unico
- per ogni nodo definiamo la distribuzione condizionate del nodo dato solo i nodi che puntano a lui direttamente, e.g $f(D|B, C)$, ma non $f(D|B, C, A)$
- la congiunta è' definita come il prodotto delle distribuzioni di ogni nodo

DAG

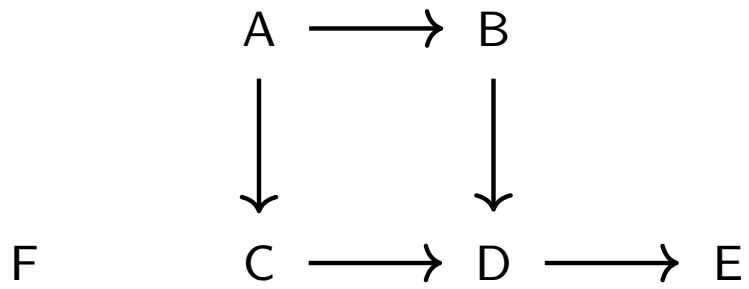


Figure: Esempio di un DAG

Quindi

$$f(A, B, C, D, E, F) = f(A)f(B|A)f(C|A)f(D|C, B)f(E|D)f(F)$$

Se le condizionate sono valide, abbiamo per costruzione che la congiunta è valida.
Questo è un caso particolare della classica decomposizione

$$f(A, B, C, D, E, F) =$$

$$f(A)f(B|A)f(C|B, A)f(D|C, B, A)f(E|D, C, B, A)f(F|E, D, C, B, A)$$

Un Esempio Climatico

Prendiamo un esempio reale in cui, su $N = 2468$ punti spaziali in Europa sono state misurate per T =365 giorni (da primo gennaio 2011 al 31 dicembre) le seguenti 3 variabili

- Temperatura minima $x_{i,t}$, $i = 1, \ldots 2468,$, $t = 1, \ldots, 365$, realizzazione del processo $X(\mathbf{s}, h)$
- Temperatura massima $z_{i,t}$, $i = 1, \ldots 2468,$, $t = 1, \ldots, 365$, realizzazione del processo $Z(\mathbf{s}, h)$.
- Precipitazioni $y_{i,t}$, $i = 1, \ldots 2468,$, $t = 1, \ldots, 365$, realizzazione del processo $Y(\mathbf{s}, h)$.

dove

$$\mathbf{s} = (s_1, s_2) \in \mathcal{D} \subset \mathbb{R}^2,$$

indica lo spazio e

$$h \in \mathcal{T} \subset \mathbb{R}^+$$

il tempo

Un Esempio Climatico

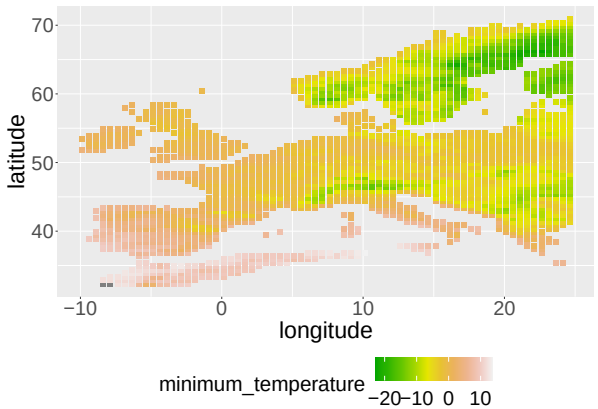
Diamo qualche definizione (le diamo per x ma definizioni equivalenti valgono per z e y)

- $\mathbf{x}_{.,t} = (x_{1,t}, \ldots, x_{N,t})'$ tutte le osservazioni al tempo t
- $\mathbf{x}_{i,.} = (x_{i,1}, \ldots, x_{i,T})'$ tutte le osservazioni del punto griglia i
- $\mathbf{x}^{d}_{i,.} = (x_{i,d+1}, x_{i,d+1}, \ldots, x_{i,T})'$ tutte le osservazioni del punto griglia i tranne le prime d
- $\mathbf{x}^{-d}_{i,.} = (x_{i,1}, x_{i,d+1}, \ldots, x_{i,T-d-1})'$ tutte le osservazioni del punto griglia i tranne le ultime d

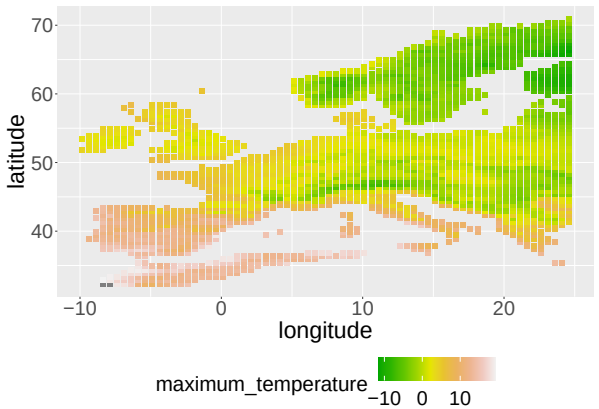
per formalizzare il modello (i.e., definire la congiunta), dobbiamo capire o avere qualche idea delle relazioni tra le variabili.

Un Esempio Climatico

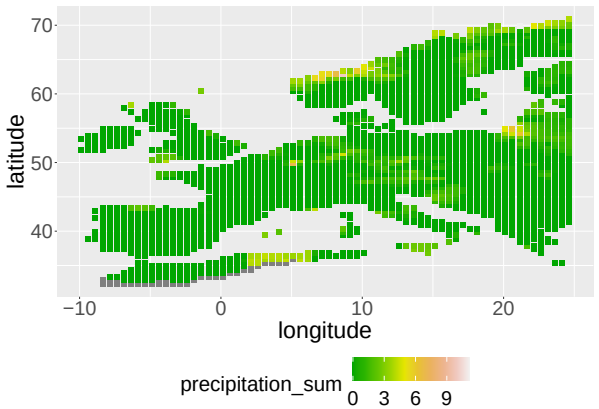
Distribuzione spaziale al tempo 1 e 8



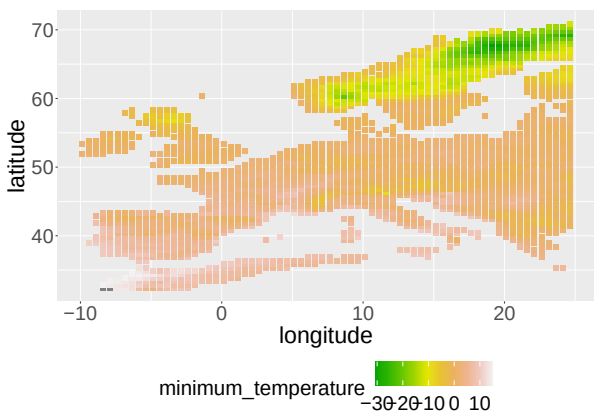
(a) $\mathbf{x}_{\cdot,1}$



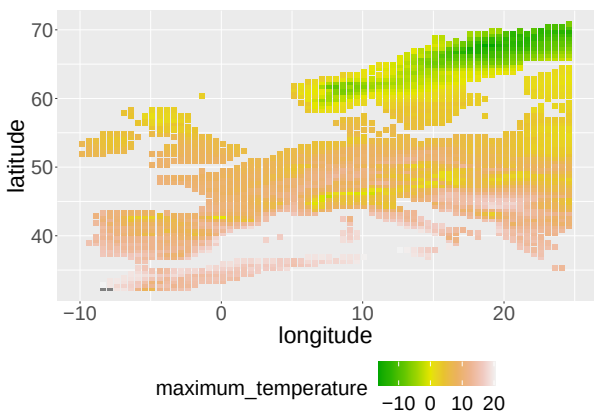
(b) $\mathbf{z}_{\cdot,1}$



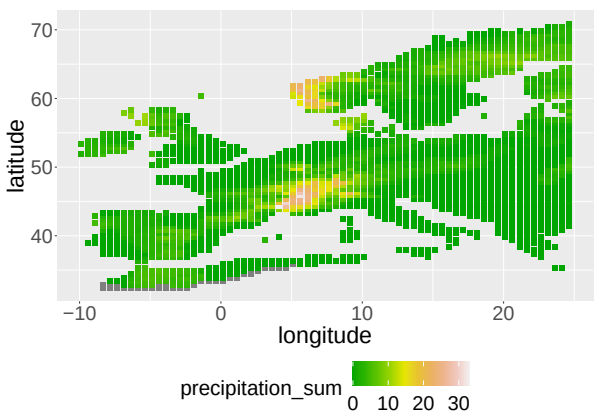
(c) $\mathbf{y}_{\cdot,1}$



(d) $\mathbf{x}_{\cdot,8}$



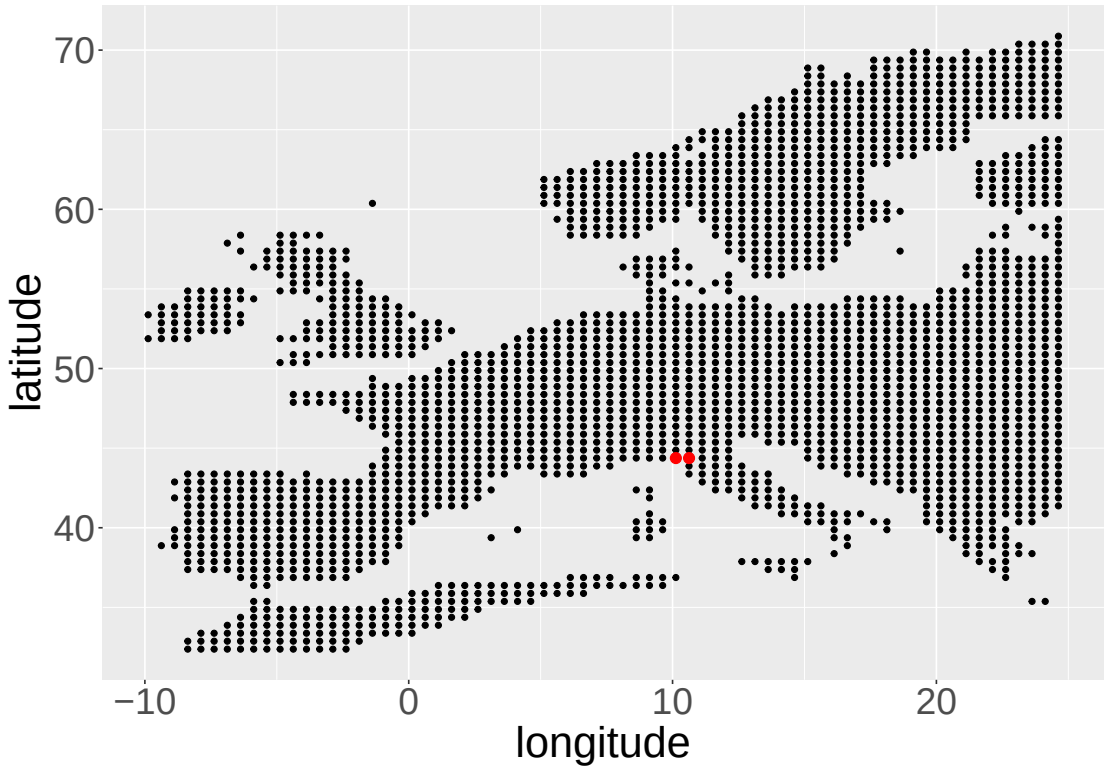
(e) $\mathbf{z}_{\cdot,8}$



(f) $\mathbf{y}_{\cdot,8}$

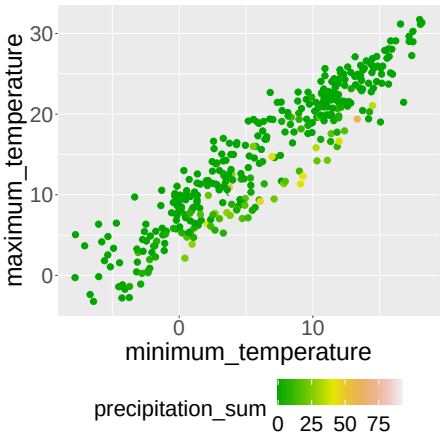
Un Esempio Climatico

Analizziamo due punti griglia (in rosso) nel dettaglio

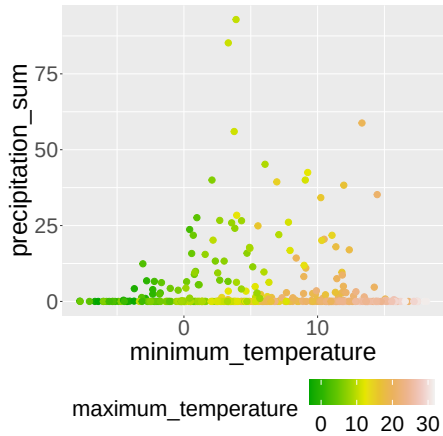


Un Esempio Climatico

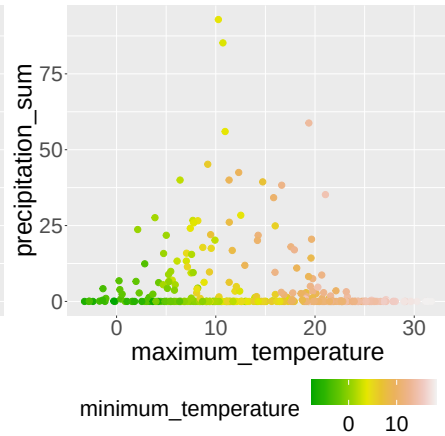
Distribuzione temporale per la stazione 671 e 672



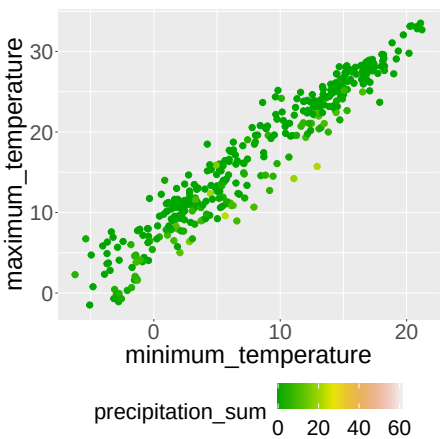
(g) $\mathbf{x}_{671,.}, \mathbf{z}_{671,.}$



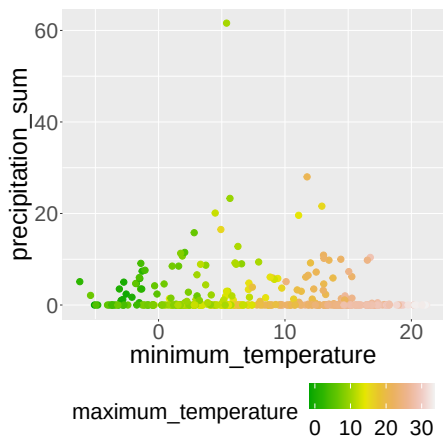
(h) $\mathbf{x}_{671,.}, \mathbf{y}_{671,.}$



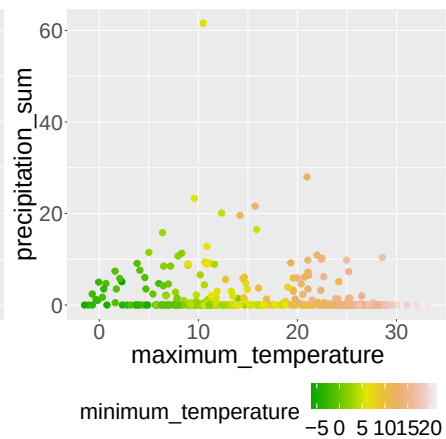
(i) $\mathbf{z}_{671,.}, \mathbf{y}_{671,.}$



(j) $\mathbf{x}_{672,.}, \mathbf{z}_{672,.}$

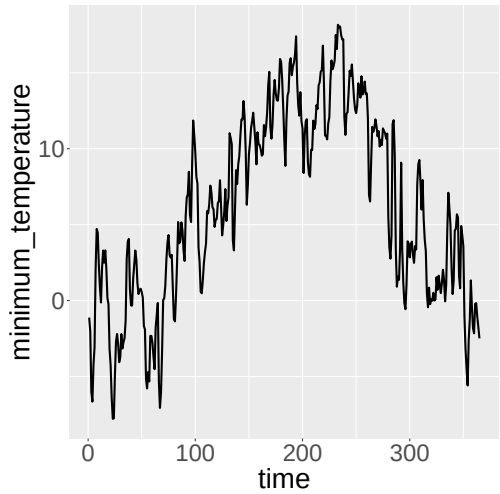


(k) $\mathbf{x}_{672,.}, \mathbf{y}_{672,.}$

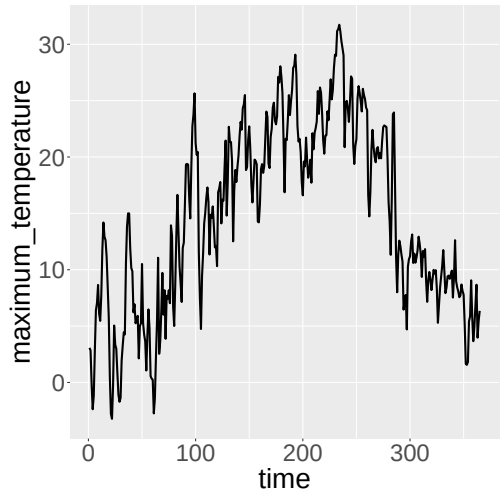


(l) $\mathbf{z}_{672,.}, \mathbf{y}_{672,.}$

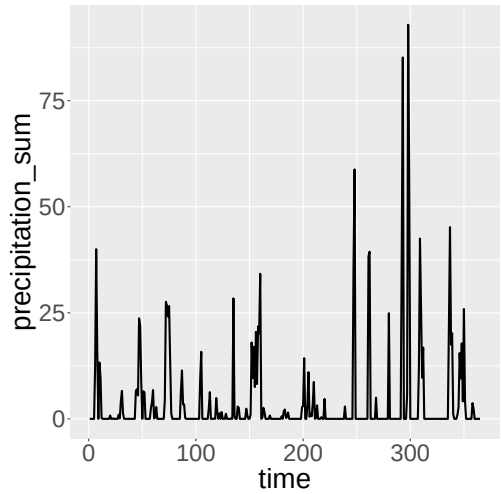
Serie storia e autocorrelazione per la stazione 671



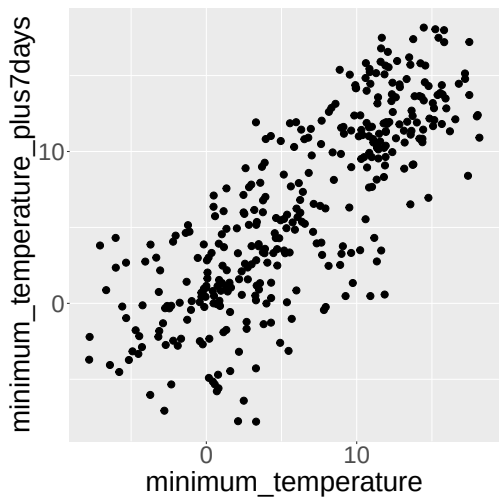
(m) $\mathbf{x}_{671,.}$



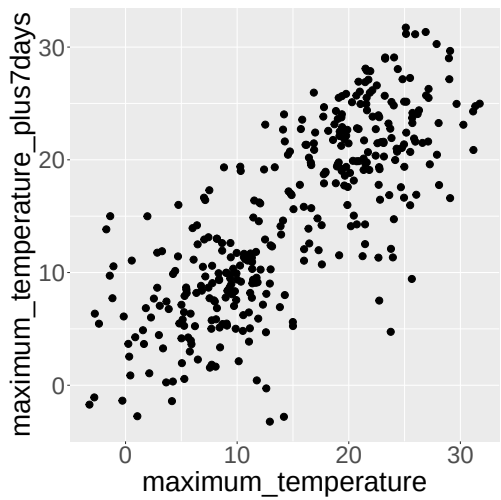
(n) $\mathbf{z}_{671,.}$



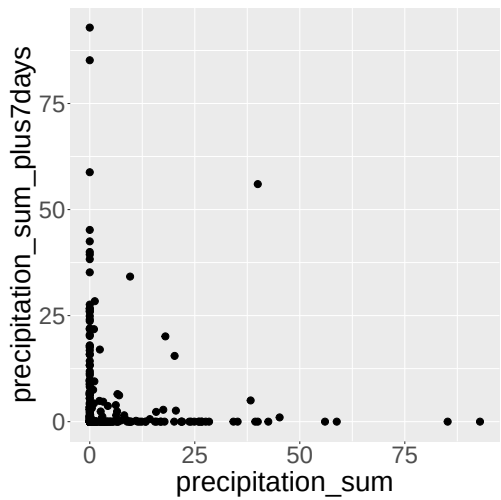
(o) $\mathbf{y}_{671,.}$



(p) $\mathbf{x}^{-7}_{671,.}, \mathbf{x}^{7}_{671,.}$

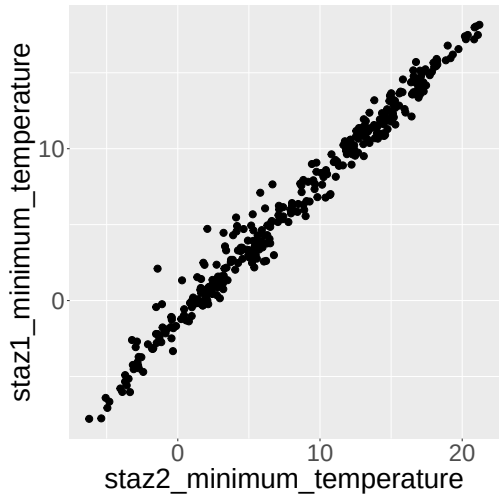


(q) $\mathbf{z}^{-7}_{671,.}, \mathbf{z}^{7}_{671,.}$

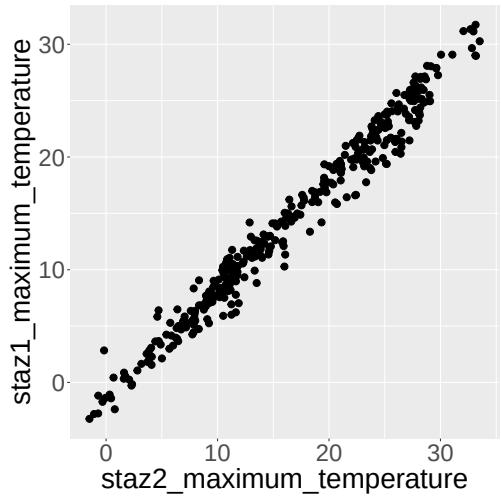


(r) $\mathbf{y}^{-7}_{671,.}, \mathbf{y}^{7}_{671,.}$

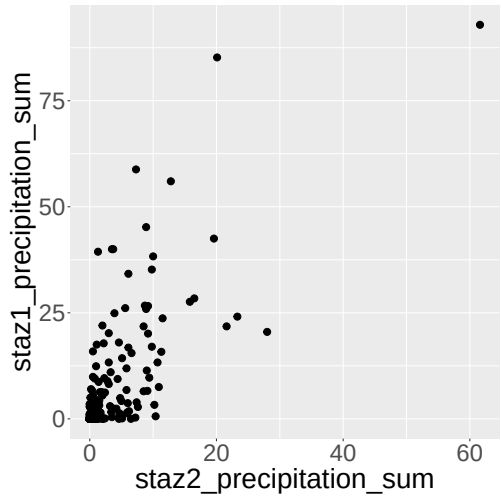
Dipendenza tra stazioni e cross-correlation



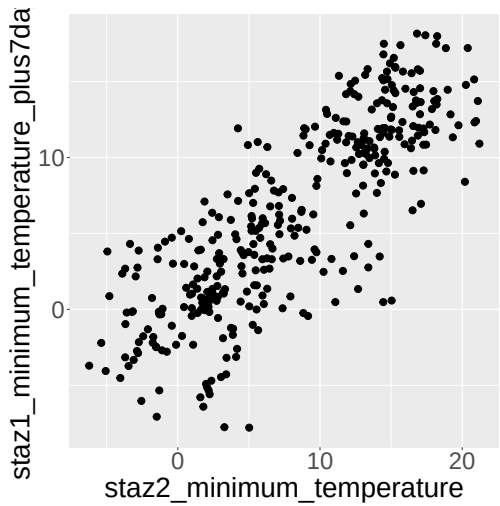
(s) $\mathbf{x}_{672,.}, \mathbf{x}_{671,.}$



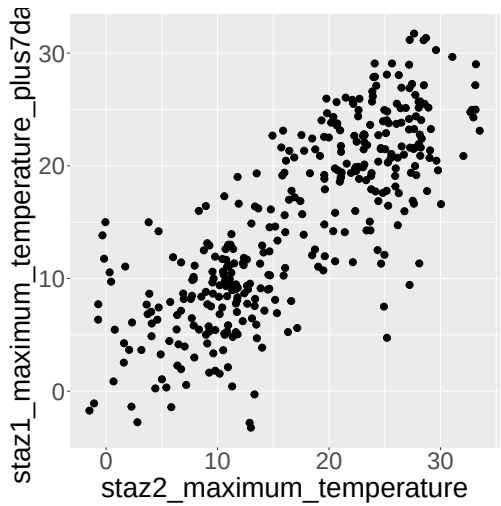
(t) $\mathbf{z}_{672,.}, \mathbf{z}_{671,.}$



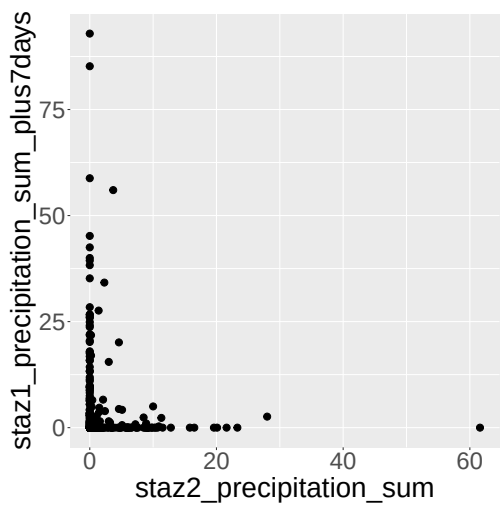
(u) $\mathbf{y}_{672,.}, \mathbf{y}_{671,.}$



(v) $\mathbf{x}^{-7}_{672,.}, \mathbf{x}^{7}_{671,.}$



(w) $\mathbf{z}^{-7}_{672,.}, \mathbf{z}^{7}_{671,.}$



(x) $\mathbf{y}^{-7}_{672,.}, \mathbf{y}^{7}_{671,.}$

All models are wrong, but some are useful

Potremmo essere tentati di costruire un modello in cui tutto è dipendente da tutto, ma poi potrebbe essere difficile da stimare. Inoltre, i due principi guida da usare per definire un modello sono

- Parsimonia (nei parametri)
- Interpretabilità (dei parametri)

Spesso si prendono decisione che si sanno non essere corrette

- in termini di dipendenza
- nelle distribuzioni

se l'effetto che hanno è poco influente.

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about safety from mice when there are tigers abroad. (**George E. P. Box**)

Modello Grafico

$$\cdots \rightarrow \mathbf{x}_{.,t-1} \rightarrow \mathbf{x}_{.,t} \rightarrow \mathbf{x}_{.,t+1} \rightarrow \cdots$$

$$\cdots \rightarrow \mathbf{z}_{.,t-1} \rightarrow \mathbf{z}_{.,t} \rightarrow \mathbf{z}_{.,t+1} \rightarrow \cdots$$

$$\cdots \rightarrow \mathbf{y}_{.,t-1} \rightarrow \mathbf{y}_{.,t} \rightarrow \mathbf{y}_{.,t+1} \rightarrow \cdots$$

In questo modello le 3 variabili sono indipendenti, mentre c'è una dipendenza temporale del primo ordine. Potremmo assumere

$$\mathbf{x}_{1,.} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \qquad \mathbf{x}_{t,.}|\mathbf{x}_{t-1,.} \sim N(\boldsymbol{\mu}_x + \mathbf{x}'_{t-1,.}\boldsymbol{\beta}_x, \boldsymbol{\Sigma}_x)$$

$$\mathbf{z}_{1,.} \sim N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \qquad \mathbf{z}_{t,.}|\mathbf{z}_{t-1,.} \sim N(\boldsymbol{\mu}_z + \mathbf{z}'_{t-1,.}\boldsymbol{\beta}_z, \boldsymbol{\Sigma}_z)$$

$$\mathbf{y}_{1,.} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \qquad \mathbf{y}_{t,.}|\mathbf{y}_{t-1,.} \sim N(\boldsymbol{\mu}_y + \mathbf{y}'_{t-1,.}\boldsymbol{\beta}_y, \boldsymbol{\Sigma}_y)$$

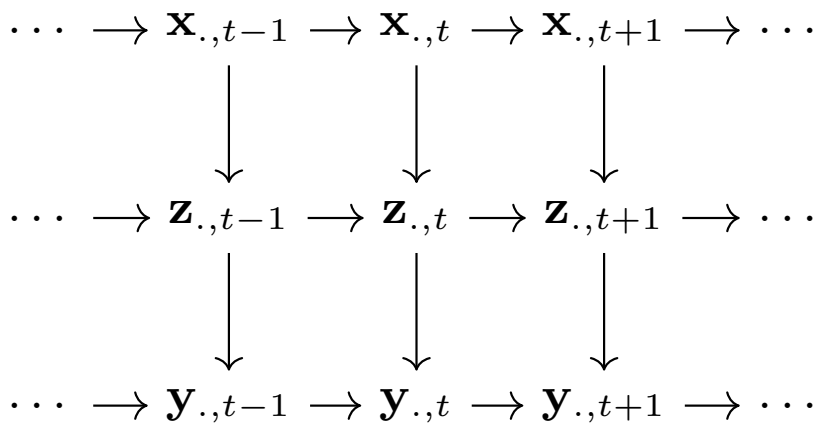
Modello Grafico

Cosa su cui ragionare

- indipendenza tra le variabili
- ha senso usare la stessa varianza/covarianza per le marginali del primo tempo e le condizionate?
- ha senso una dipendenza temporale solo rispetto al tempo precedente
- potremmo definire una covarianza tra le osservazioni dello stesso tempo che dipende dalle distanze (la matrice deve essere definita positiva) $\rightarrow$ **Processo Gaussiano**
- La pioggia è una variabile mista, con una massa a zero e poi valori solo positivi. Ha senso una normale multivariata? ma cosa usare altrimenti?

Potremmo provare ad aggiungere dipendenza tra le variabili

Modello Grafico



In questo modello le 3 variabili sono indipendenti, mentre c'è una dipendenza temporale del primo ordina. Potremmo assumere

$$\mathbf{x}_{1,.} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \qquad \mathbf{x}_{t,.}|\mathbf{x}_{t-1,.} \sim N(\boldsymbol{\mu}_x + \mathbf{x}'_{t-1,.}\boldsymbol{\beta}_x, \boldsymbol{\Sigma}_x)$$

$$\mathbf{z}_{1,.}|\mathbf{x}_{1,.} \sim N(\boldsymbol{\mu}_z + \mathbf{x}'_{1,.}\boldsymbol{\gamma}_z, \boldsymbol{\Sigma}_z) \qquad \mathbf{z}_{t,.}|\mathbf{z}_{t-1,.}, \mathbf{x}_{t,.} \sim N(\boldsymbol{\mu}_z + \mathbf{z}'_{t-1,.}\boldsymbol{\beta}_z + \mathbf{x}'_{t,.}\boldsymbol{\gamma}_z, \boldsymbol{\Sigma}_z)$$

$$\mathbf{y}_{1,.}|\mathbf{z}_{1,.} \sim N(\boldsymbol{\mu}_y + \mathbf{z}'_{1,.}\boldsymbol{\gamma}_y, \boldsymbol{\Sigma}_y) \qquad \mathbf{y}_{t,.}|\mathbf{y}_{t-1,.}, \mathbf{z}_{t,.} \sim N(\boldsymbol{\mu}_y + \mathbf{y}'_{t-1,.}\boldsymbol{\beta}_y + \mathbf{z}'_{t,.}\boldsymbol{\gamma}_y, \boldsymbol{\Sigma}_y)$$

Modello Grafico

Cosa su cui ragionare

- ~~indipendenza tra le variabili~~
- ha senso usare la stessa varianza/covarianza per le marginali del primo tempo e le condizionate?
- ha senso una dipendenza temporale solo rispetto al tempo precedente
- potremmo definire una covarianza tra le osservazioni dello stesso tempo che dipende dalle distanze (la matrice deve essere definita positiva) $\rightarrow$ **Processo Gaussiano**
- La pioggia è una variabile mista, con una massa a zero e poi valori solo positivi. Ha senso una normale multivariata? ma cosa usare altrimenti?
- (**New**) l'ordine dato tra le 3 variabili, ha qualche tipo di effetto? (otterrei lo stesso modello se definissi il DAG come $\mathbf{y}_{.,t} \rightarrow \mathbf{z}_{.,t} \rightarrow \mathbf{x}_{.,t}$?)

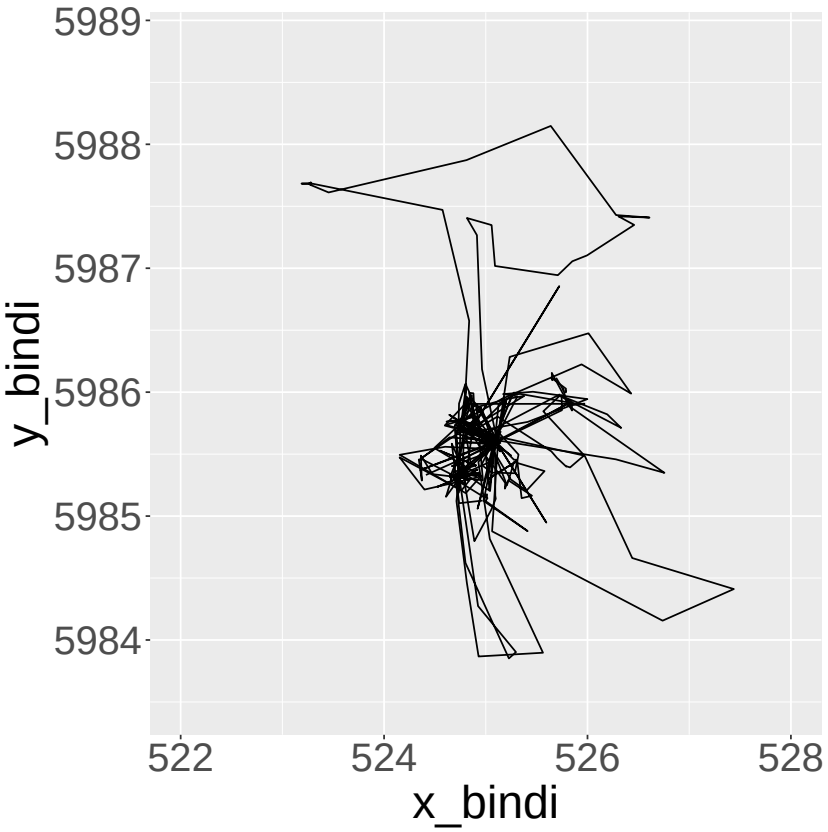
Un esempio Biologico

Abbiamo le coordinate, registrate ogni 30 secondi, di 2 cani pastore in una proprietà in Australia. Gli animali sono osservati per diversi periodi e spesso non insieme. Definiamo

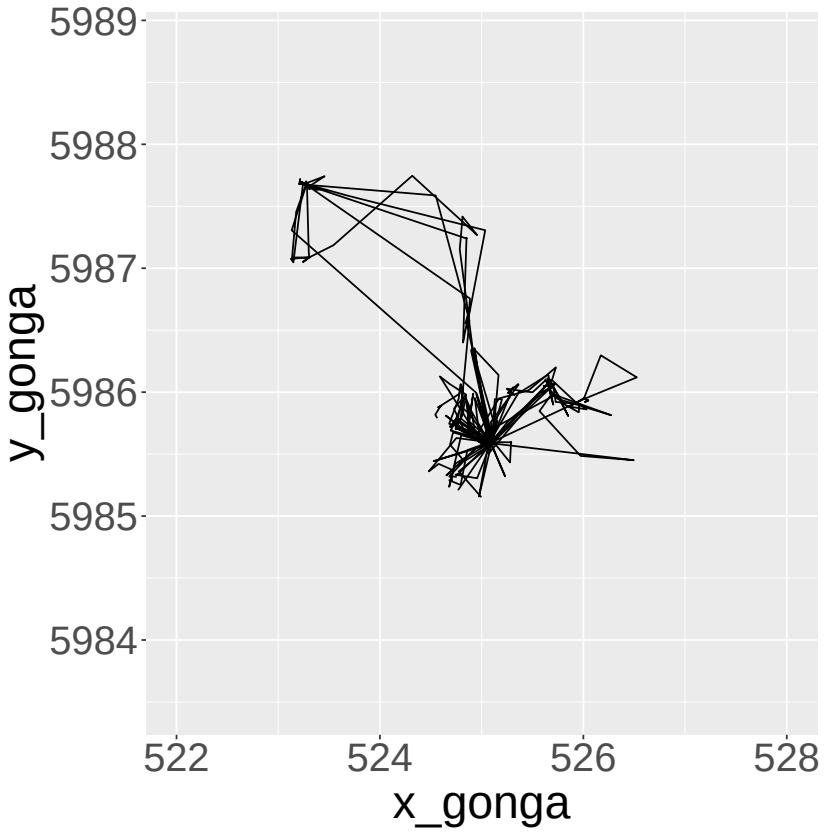
- $\mathbf{s}_{i,t} = (s_{i,t,1}, s_{i,t,2})' \in \mathbb{R}^2$ come le coordinate dell'animale $i-$esimo al $t-$esimo istante temporale
- $\lambda_{i,t} = ||\mathbf{s}_{i,t+1} - \mathbf{s}_{i,t}||_2 \in \mathbb{R}$ distanza percorsa (step-length)
- $\phi_{i,t} = \text{atan}^*(s_{i,t+1,2} - s_{i,t,2}, s_{i,t+1,1} - s_{i,t,1})$ angolo di spostamento (bearing-angle) dove atan* è l'inversa della tangente definita in $[02\pi)$
- $\eta_{i,t} = \phi_{i,t} - \phi_{i,t-1}$ angolo di rotazione rispetto all'ultimo movimento (tuning-angle)

Un esempio Biologico

Plot dei path dei singoli animali



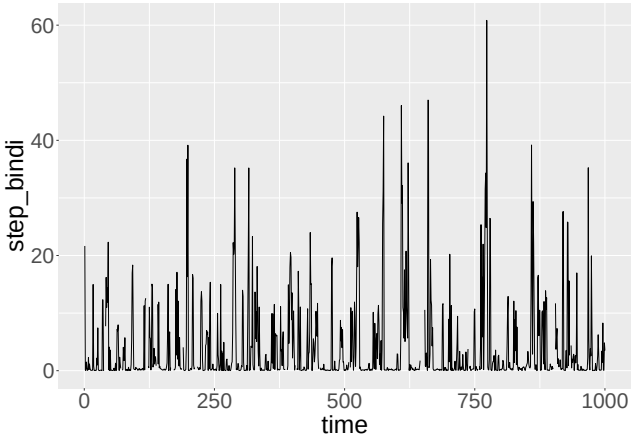
(a) $\mathbf{s}_{1,t}$



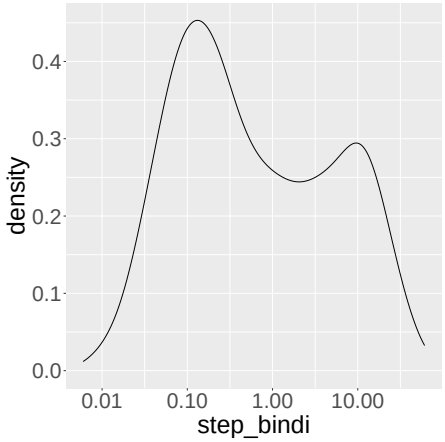
(b) $\mathbf{s}_{2,t}$

Un esempio Biologico

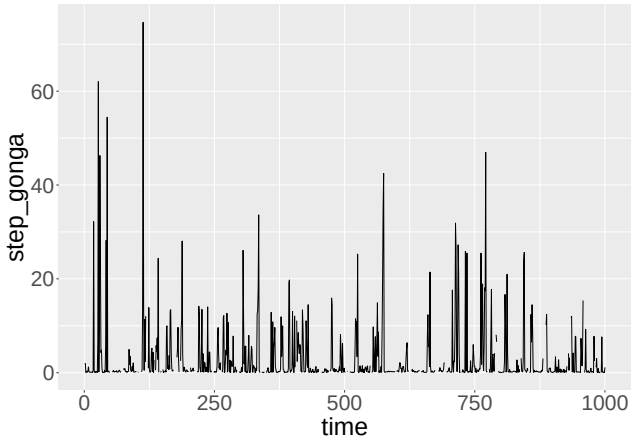
Vediamo gli step-length (serie storia e density (scala log))



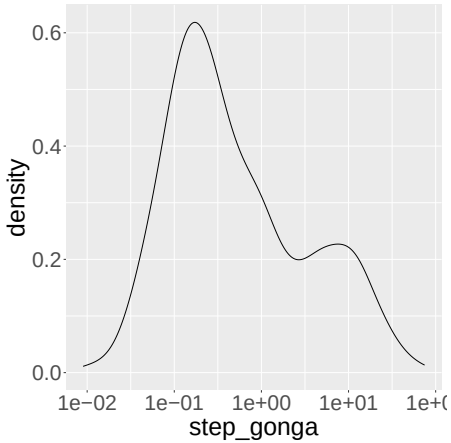
(a) $\lambda_{1,t}$



(b) $\lambda_{1,t}$



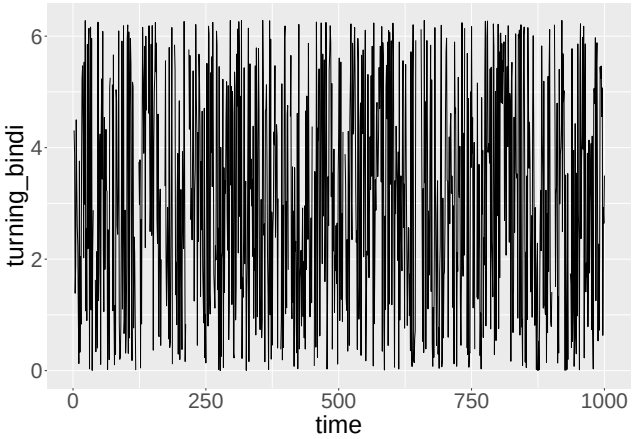
(c) $\lambda_{2,t}$



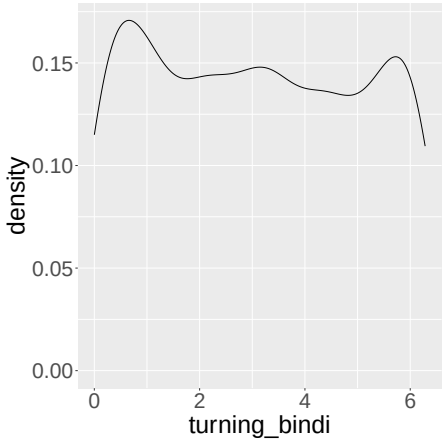
(d) $\lambda_{2,t}$

Un esempio Biologico

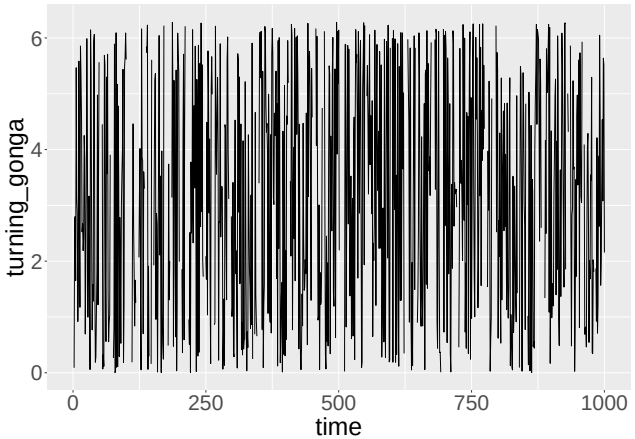
Vediamo i turning-angle (serie storia e density)



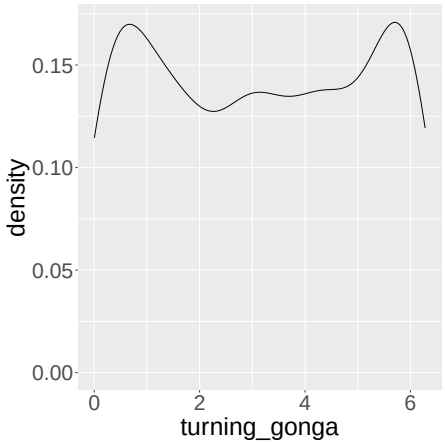
(a) $\eta_{1,t}$



(b) $\eta_{1,t}$



(c) $\eta_{2,t}$



(d) $\eta_{2,t}$

Variabili-latenti:

- variabili aleatorie non osservate
- vengono utilizzate per
  - facilitare l'implementazione
  - definire densità più flessibili
- possono rappresentare concetti "reali" o no.

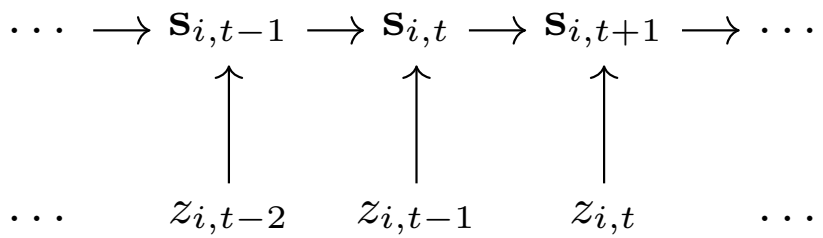
Possiamo spiegare parte dei dati se introduciamo il concetto di **comportamento** animale (e.g. dormire/cacciare/camminare) e il movimento di un animale dipende dal comportamento che sta avendo a un determinato tempo.

Introduciamo quindi

$$z_{i,t} \in \{1, 2, \ldots, K\}$$

che rappresenta il comportamente dell'animale i al tempo t. Possiamo provare a definire un modello, assumendo gli animali indipendenti.

Modello Grafico - Variabili Latenti - Modelli Gerarchici



assumendo $\mathbf{s}_{i,0}$ noto (potremmo definirlo come $\mathbf{s}_{i,0} = \boldsymbol{\mu}_i +$), possiamo definire

$$\mathbf{s}_{i,t+1} | \mathbf{s}_{i,t}, z_{i,t} \sim N(\boldsymbol{\mu}_i + \alpha_{i,z_{i,t}} \mathbf{s}_{i,t}, \sigma_{i,z_{i,t}} \mathbf{I})$$

$$z_{i,t} \sim Disc(\boldsymbol{\pi})$$

Questo è un modello è **gerarchico**

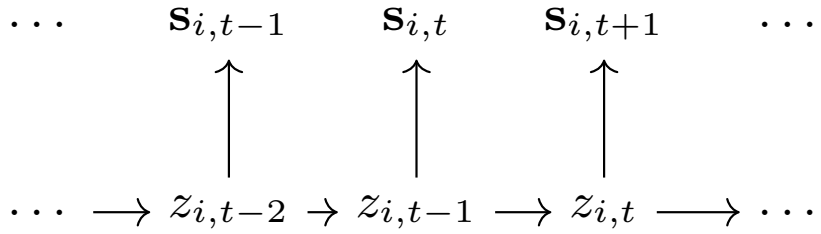
Cosa su cui ragionare

- i comportamenti sono indipendenti, ha senso?
- il modo in cui le posizioni sono dipendenti è veramente quello che vogliamo?
- cosa è μ_i?

Questo è un modello mistura

Modello Grafico - Variabili Latenti - Modelli Gerarchici

Un modello alternativo è



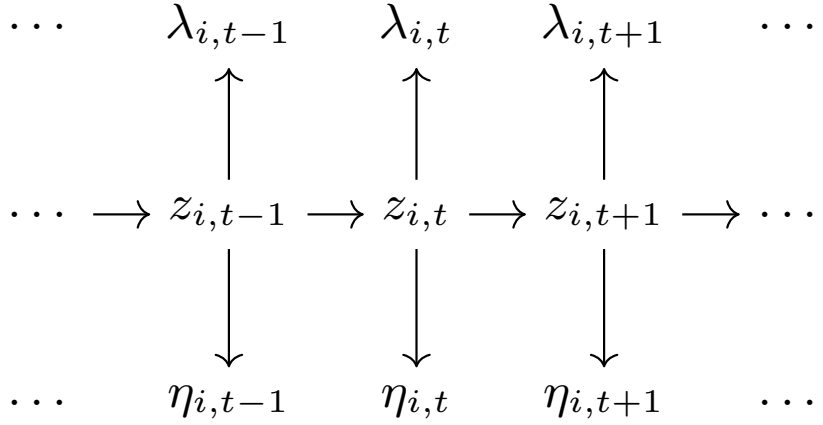
definendo

$$\mathbf{s}_{i,t+1}|z_{i,t} \sim N(\boldsymbol{\mu}_i, \sigma_{i,z_{i,t}}\mathbf{I})$$

- Come definiamo la distribuzione di $z_{i,t}|z_{i,t-1}$? $\rightarrow$ **Markov-chain e Hidden Markov model**

Modello Grafico - Variabili Latenti - Modelli Gerarchici

Un'altra alternativa



In questo modello, visto che $\eta_{i,t}$ è calcolato usando 3 punti spaziali e $\lambda_{i,t}$ 2 punti spaziali, stiamo implicitamente introducendo dipendenza tra le coordinate. Assumiamo

$$\eta_{i,t}|z_{i,t} \sim G(a_{z_{i,t}}, b_{z_{i,t}})$$

- Come definiamo la distribuzione di $z_{i,t}|z_{i,t-1}$? $\rightarrow$ **Markov-chain e Hidden Markov model**
- Come definiamo la distribuzione di $\eta_{i,t}|z_{i,t}$? $\rightarrow$ **Distribuzione per variabili circolare**