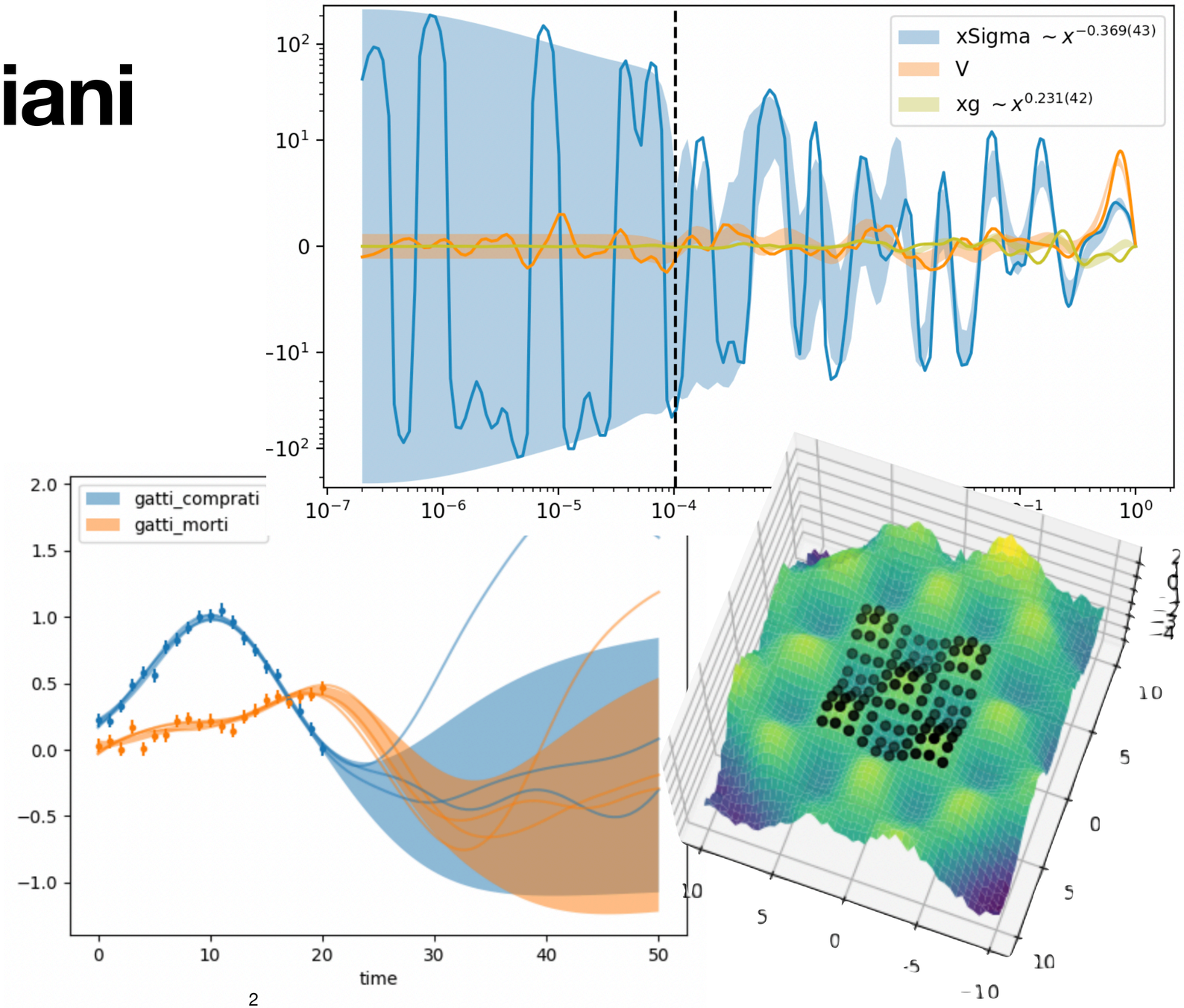
Seminario primo anno PhD statistica

Studio e calcolo di funzioni di covarianza per la regressione con i processi Gaussiani

Giacomo Petrillo, DiSIA UNIFI, 23 settembre 2022

Processi Gaussiani

- Modello di regressione nonparametrico
- Tradizionalmente usato in geostatistica (kriging)
- Però un po' ovunque in realtà (machine learning, ingegneria, fisica...)



Processi Gaussiani

Definizione

- Processo Gaussiano = Normale multivariata in ∞ dimensioni
- Ogni marginale è Normale multivariata finito-dimensionale
- La predizione si fa con la formula di condizionamento per le Normali m.v.:
  $E[y_{\text{inc}} \mid y_{\text{obs}}] = \Sigma_{\text{inc}-\text{obs}} \Sigma_{\text{obs}}^{-1} y_{\text{obs}}$

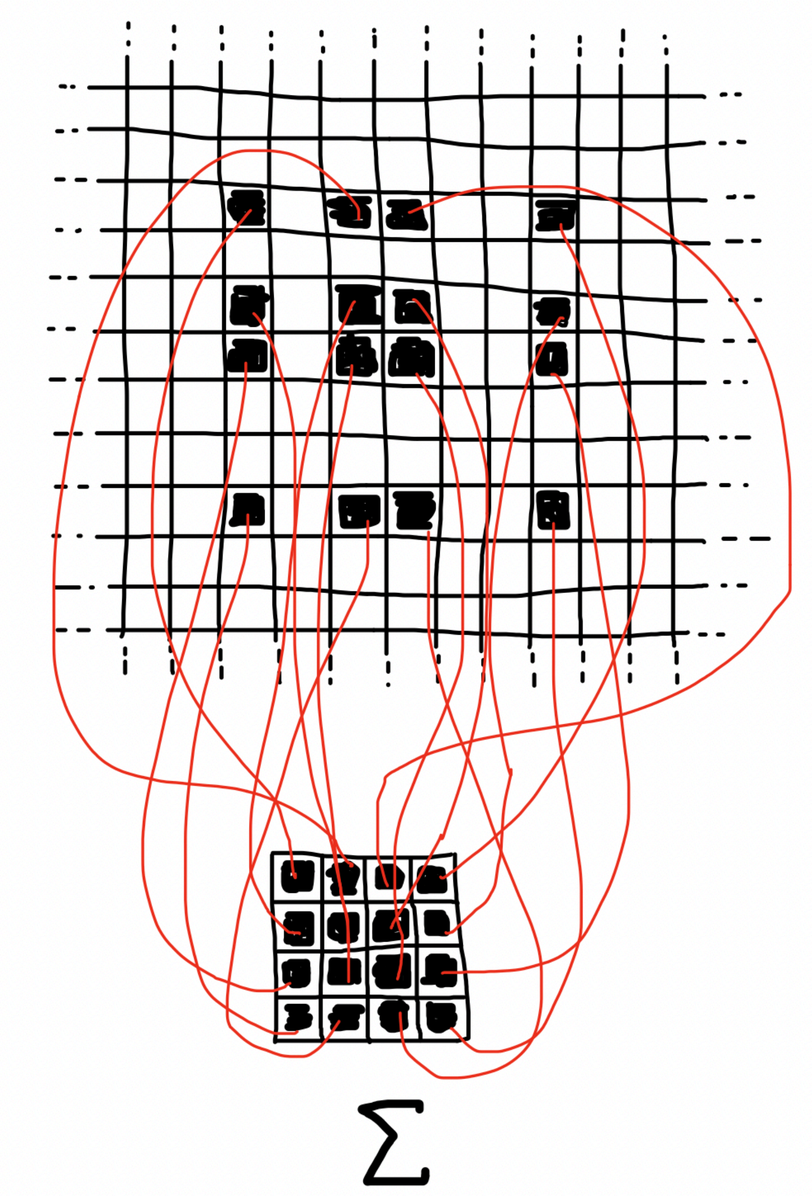
$$\begin{bmatrix} y_{inc} \\ y_{obs} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} ; \begin{bmatrix} \Sigma_{inc} & \Sigma_{inc\text{-}obs} \\ \Sigma_{obs\text{-}inc} & \Sigma_{obs} \end{bmatrix} \right)$$

Argomento:

Studio e calcolo di funzioni di covarianza

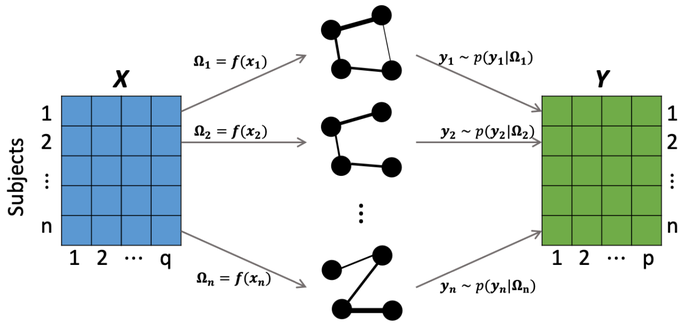
La funzione di covarianza (o kernel) di un processo f è

$$k(x, y) = \mathrm{Cov}[f(x), f(y)]$$

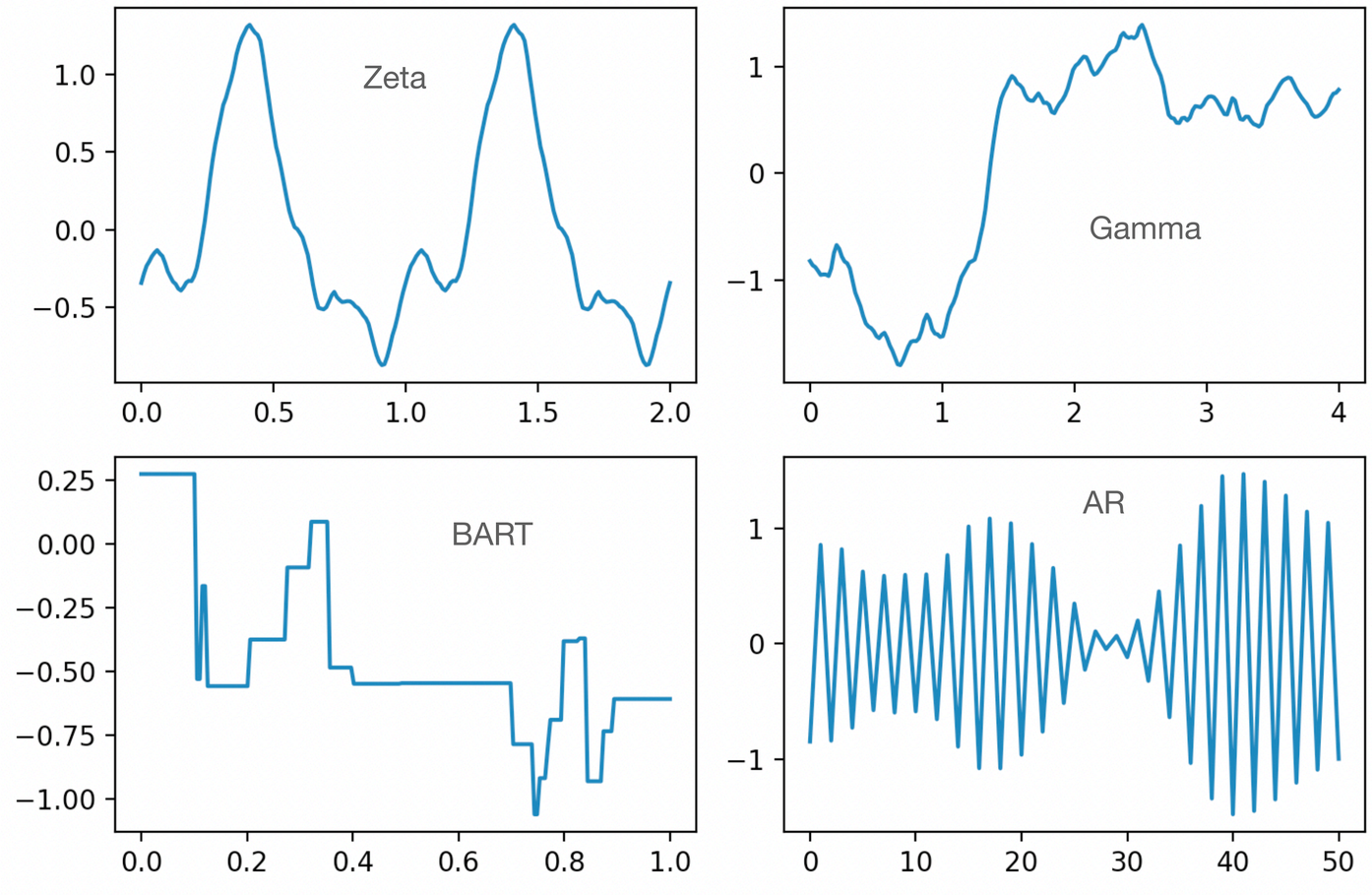


Kernels che introduco:

- Zeta di Lerch
- Gamma incompleta
- BART
- AR(p)
- Qualcosa sui grafi, da definirsi

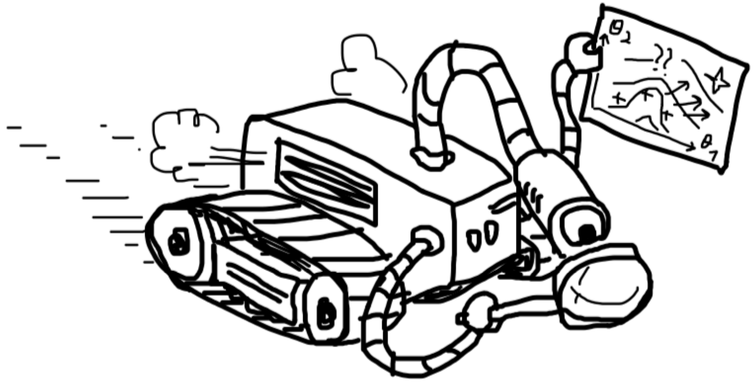


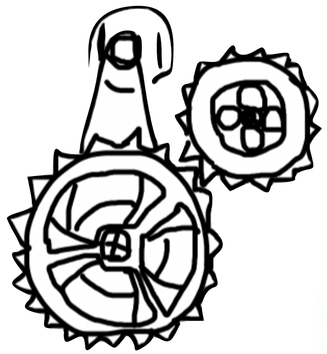
Ni, Stingo, Baladandayuthapani 2022



Quali sono gli obiettivi nel definire una funzione di covarianza?

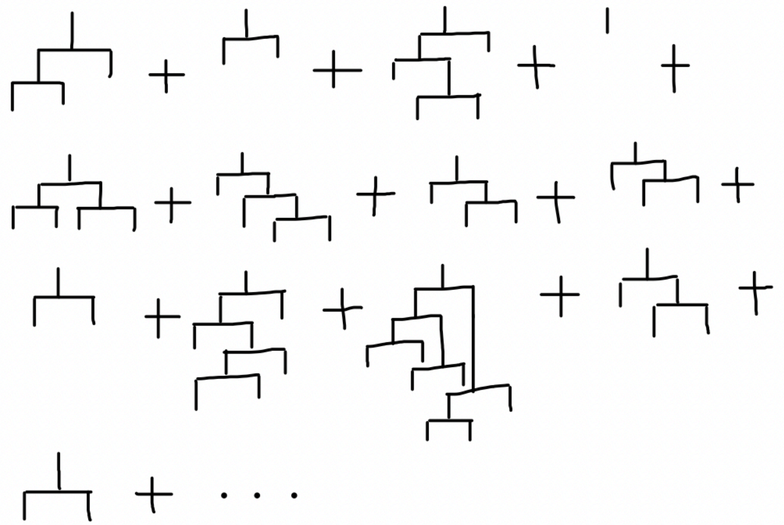
- Parametrizzazione flessibile, per farci inferenza
- Calcolabile efficientemente
- Calcolabili anche le derivate prime e seconde rispetto ai parametri
- Deve applicarsi a problemi concreti





BART

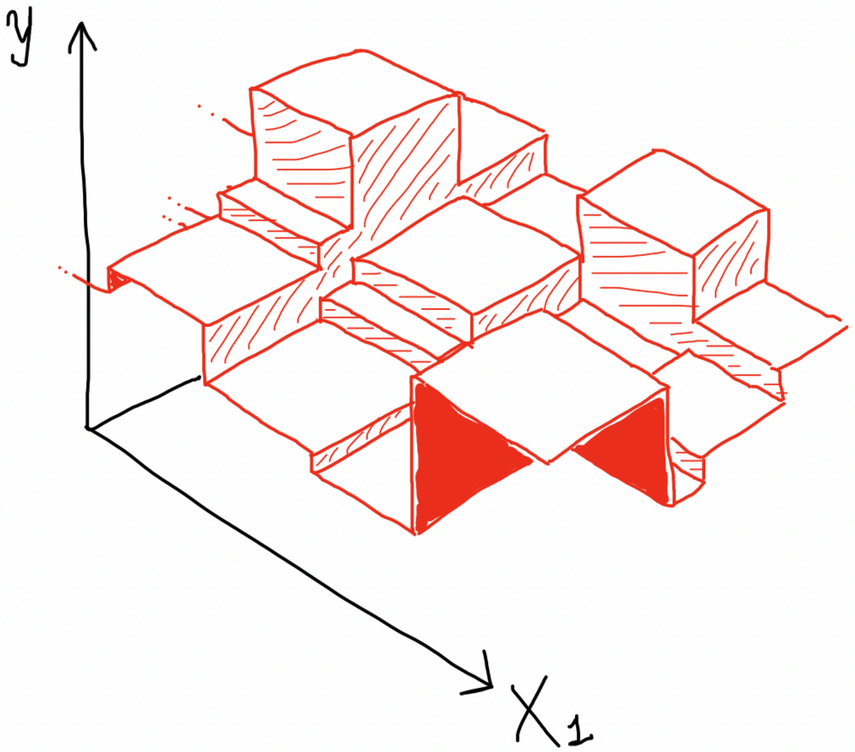
Siccome ho poco tempo, spiego solo il BART.



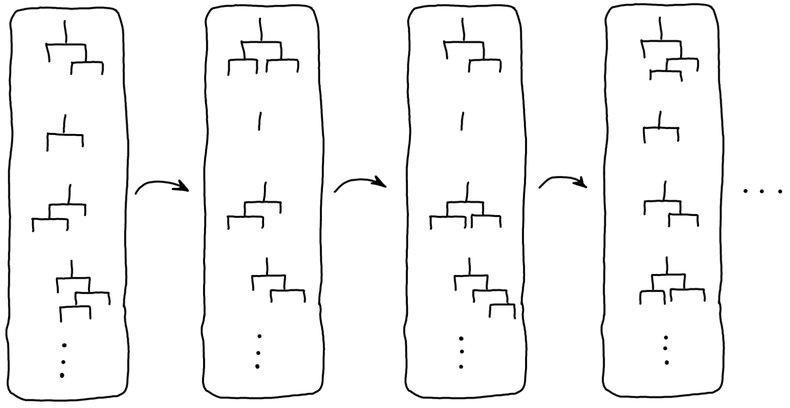
BART = Bayesian Additive Regression Trees

= somma di 100-1000 alberi di decisione usati in modo Bayesiano

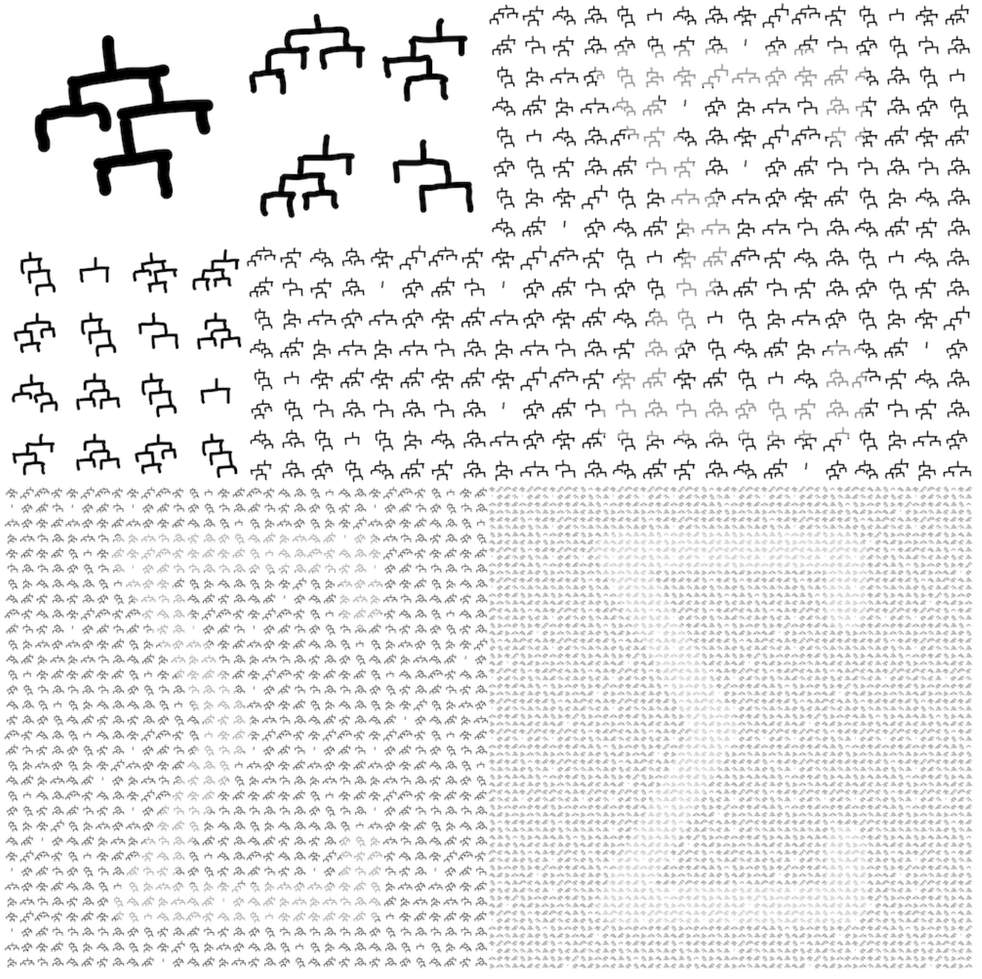
$$y = f(x) + \varepsilon, \qquad f(x) = \sum_{i=1}^{n} T_i(x)$$



BART



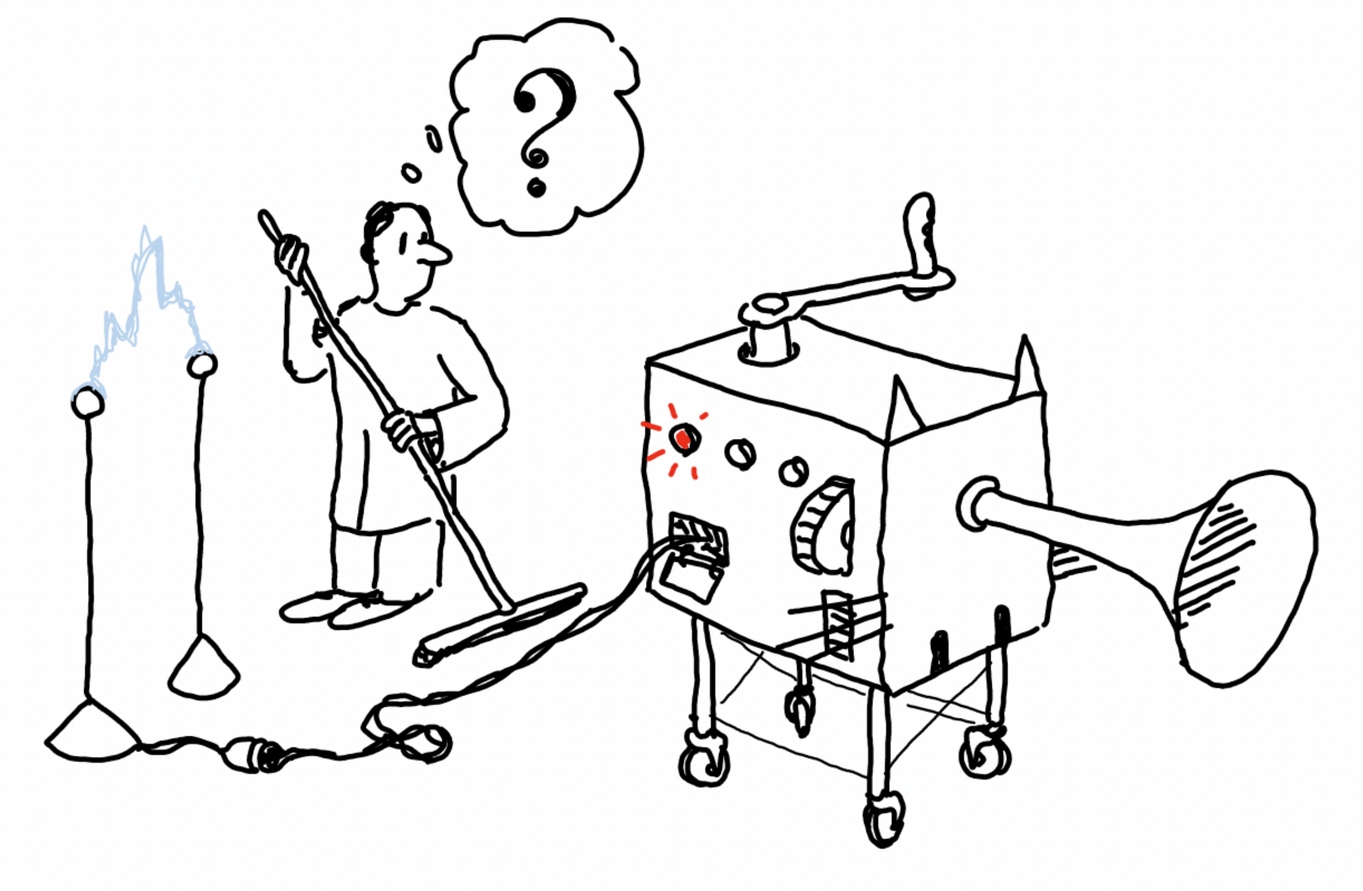
- L'inferenza si fa con MCMC ad hoc
- Ma nel limite di infiniti alberi diventa un processo Gaussiano (Linero 2017)
- ⇒ Si può fare inferenza come processo Gaussiano, senza MCMC
- ⇒ Obiettivo: calcolare la matrice di covarianza del processo Gaussiano che corrisponde al BART



BART

BART come GP:

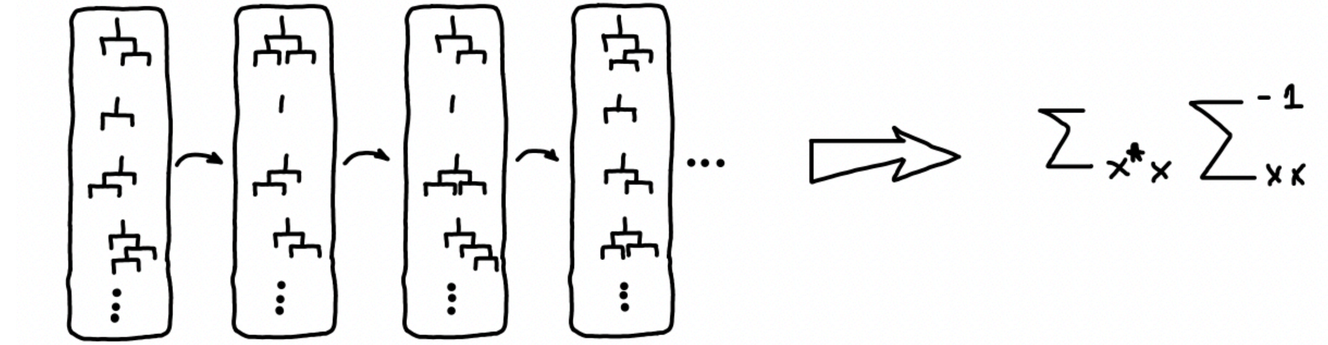
- A cosa serve?
- Come si fa?

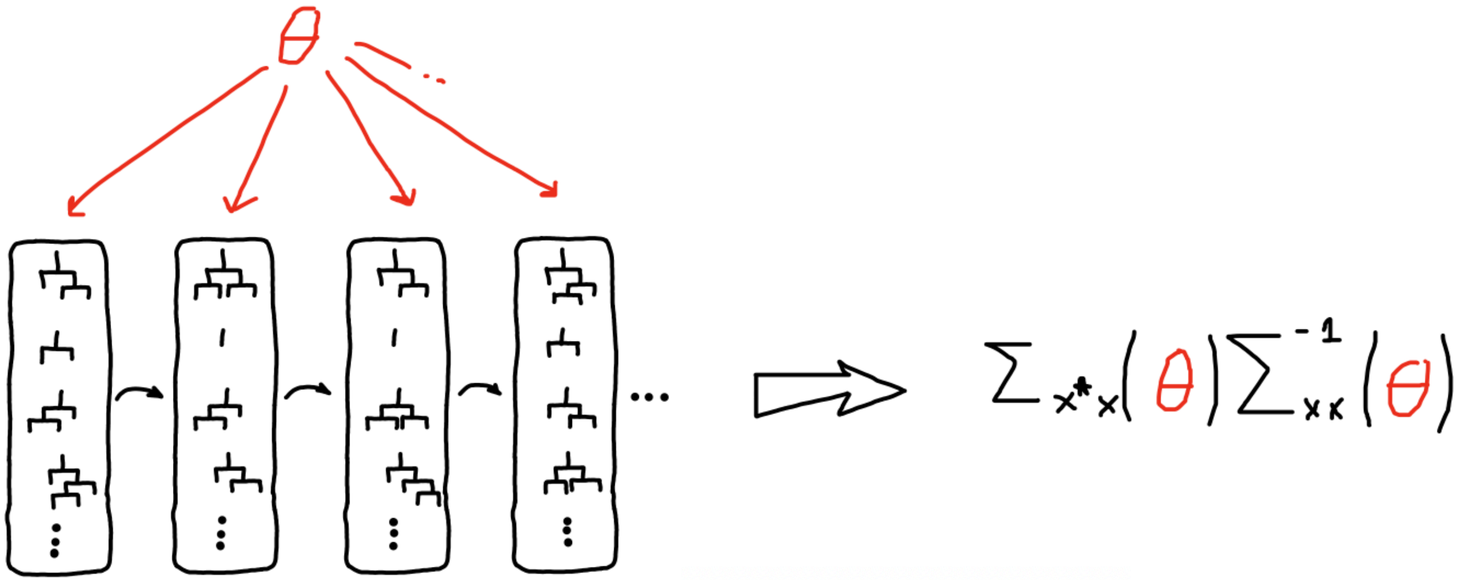


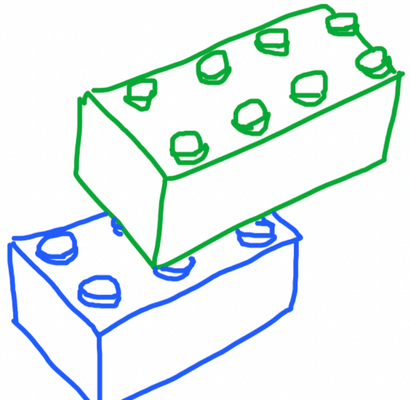
BART

A cosa serve BART come GP?

- Metodo alternativo di stima del BART
- Inferenza Bayesiana completa (MCMC non stima tutti i parametri)
- Facilità di combinare il BART come mattoncino insieme ad altri processi Gaussiani







BART

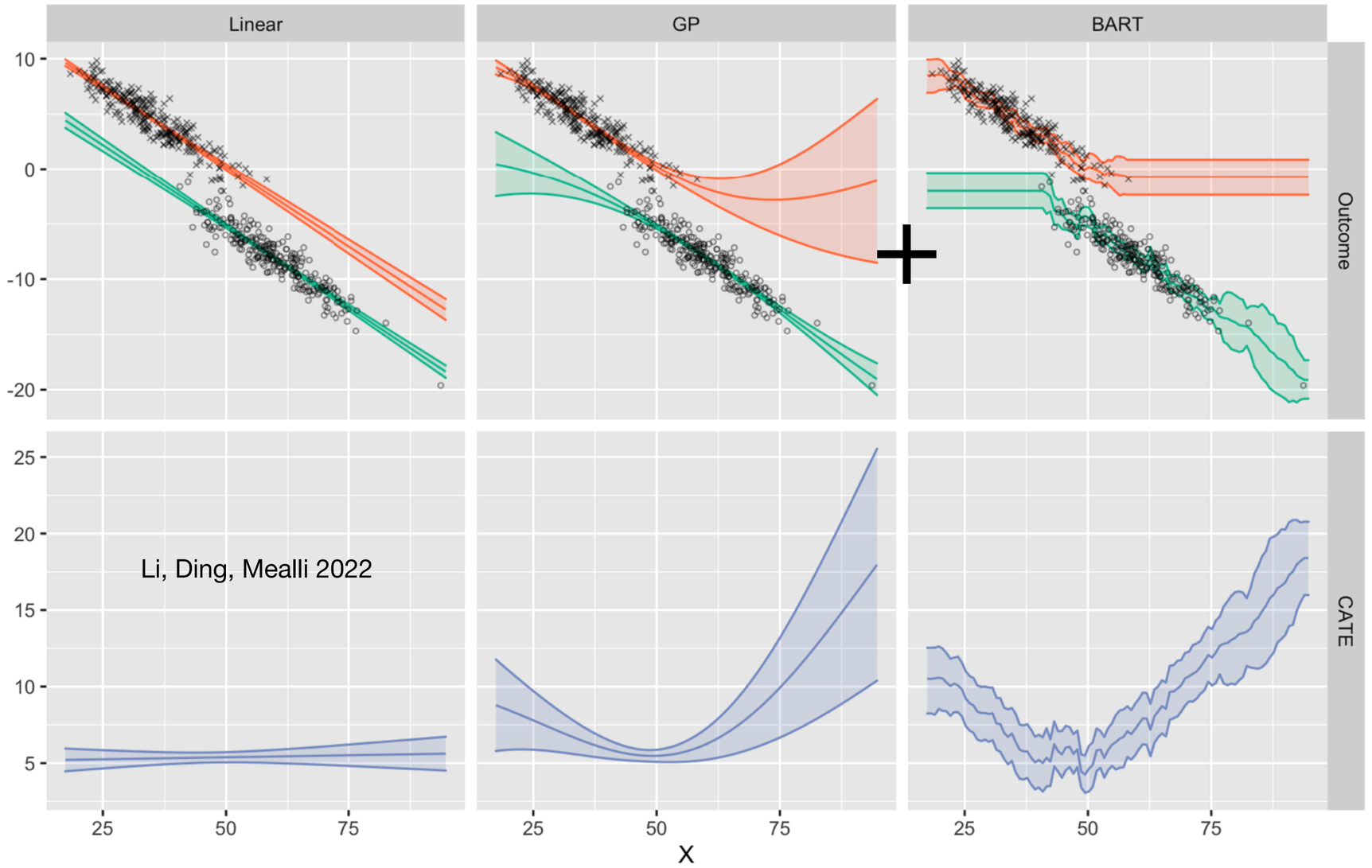
A cosa serve BART come GP?

$$\Sigma[f+g] = \Sigma[f] + \Sigma[g]$$

$$\Sigma[fg] = \Sigma[f]\Sigma[g]$$

…

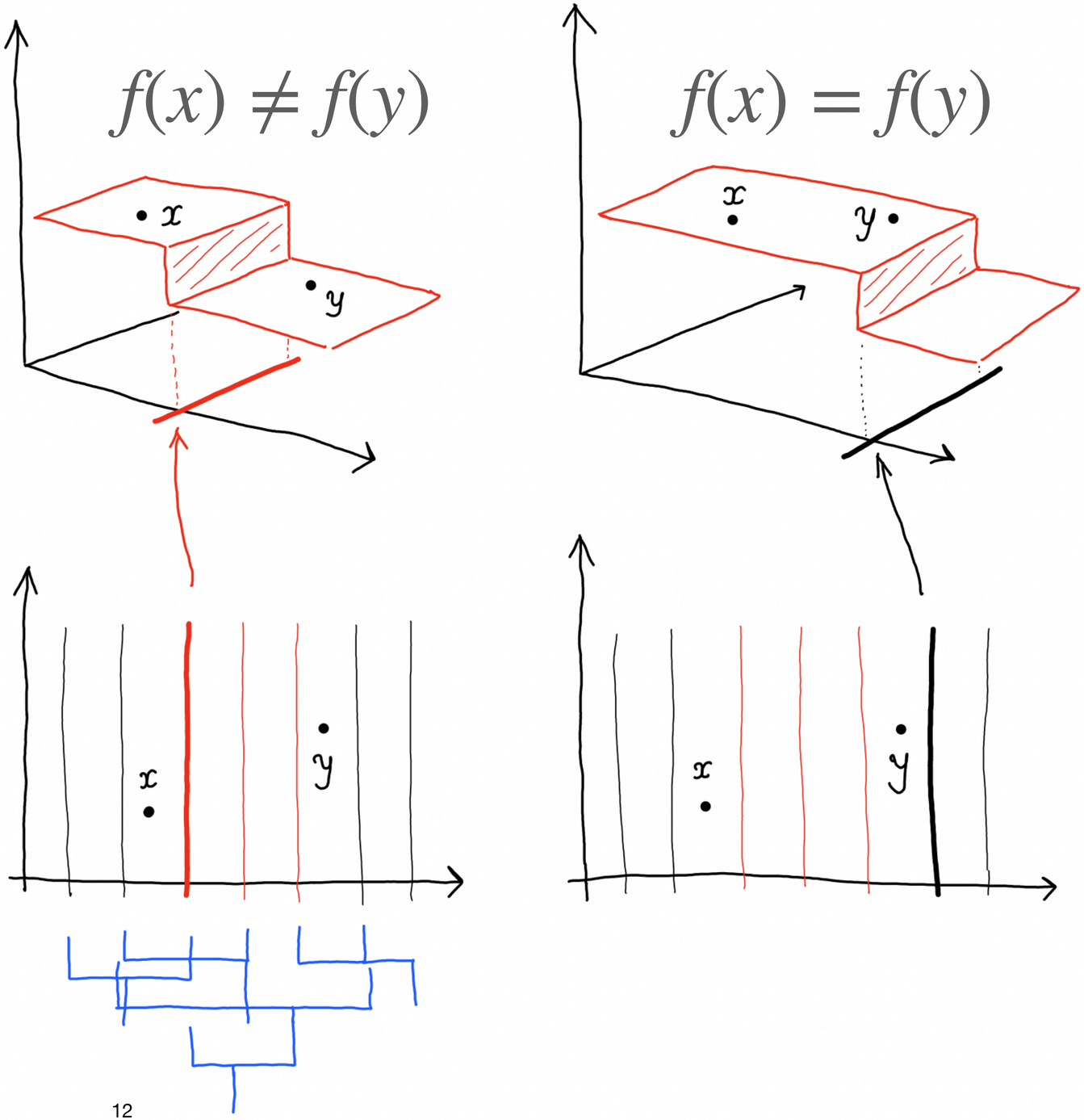
- Facilità di combinare il BART come mattoncino insieme ad altri processi Gaussiani



BART

Come si fa a usare BART come GP?

- Bisogna calcolare la funzione di covarianza
- Correlazione tra $f(x)$ e $f(y)$ = probabilità che l'albero *non* piazzi una divisione tra x e y
- Va calcolata ricorsivamente seguendo tutti i possibili alberi (es. X è 100x10 => ≈2^1000)
- ⇒ troppo lento



BART

Come faccio a calcolare la correlazione?

Uso un algoritmo approssimato.
Requisiti:

- Buona approssimazione del BART
- Computazionalmente fattibile
- Generare una matrice valida (definita positiva)
- Al "limite" = esattamente BART

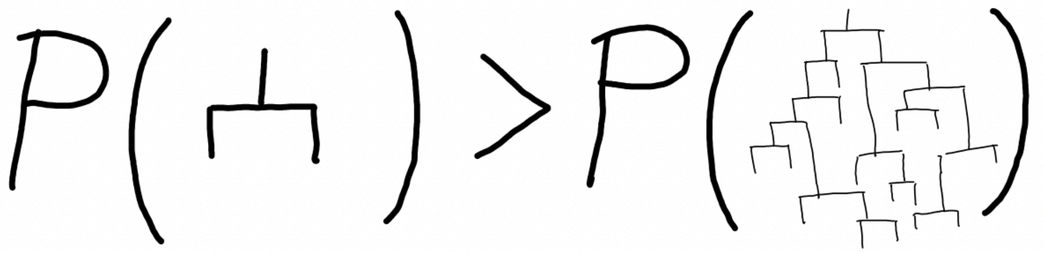
$$\tilde{\Sigma} \approx \Sigma_{\text{BART}}$$

$$\tilde{\Sigma} \geq 0$$

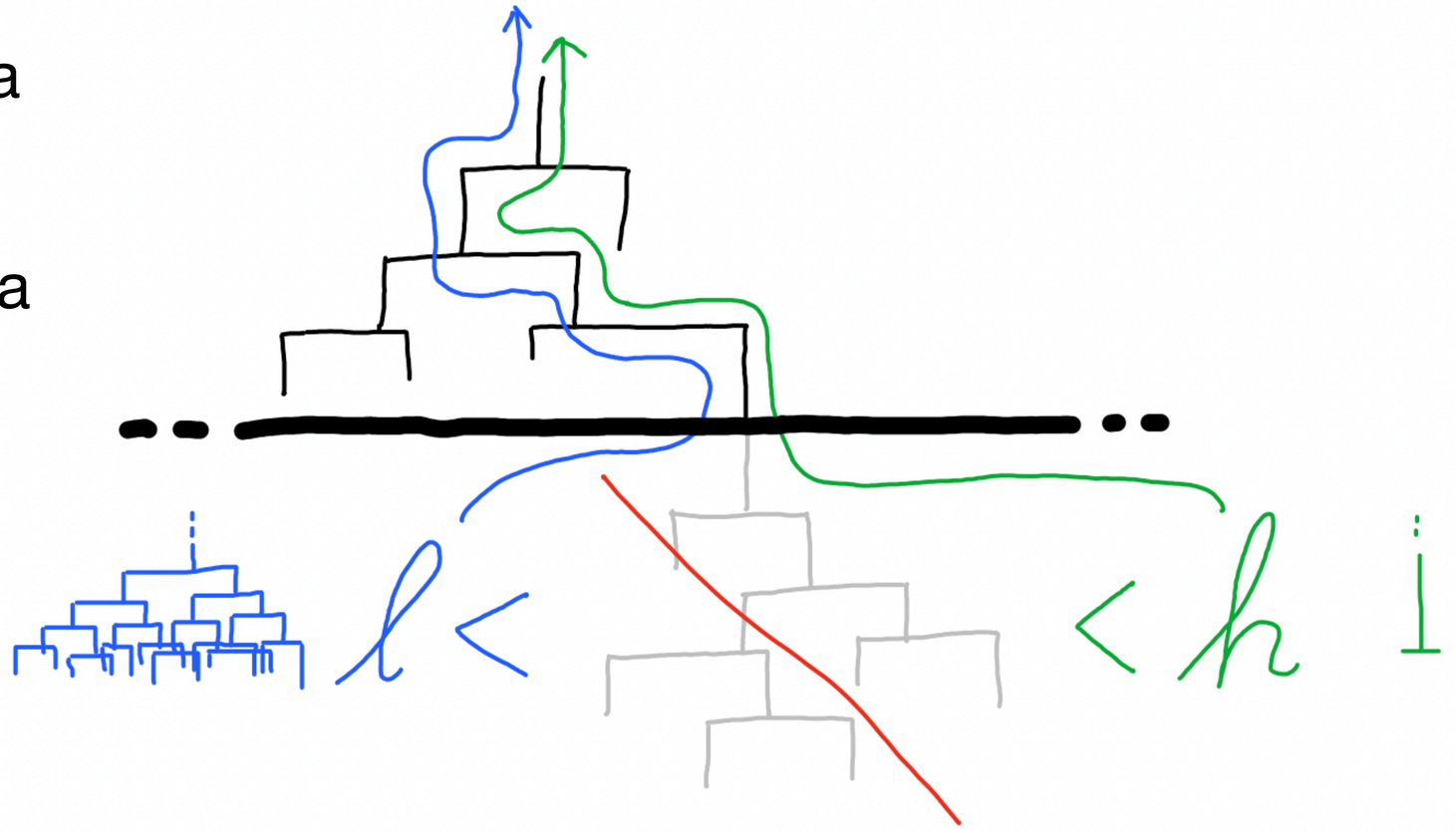
$$\tilde{\Sigma} \to \Sigma_{\text{BART}}$$

BART

Come faccio a non dover considerare tutti gli alberi possibili?



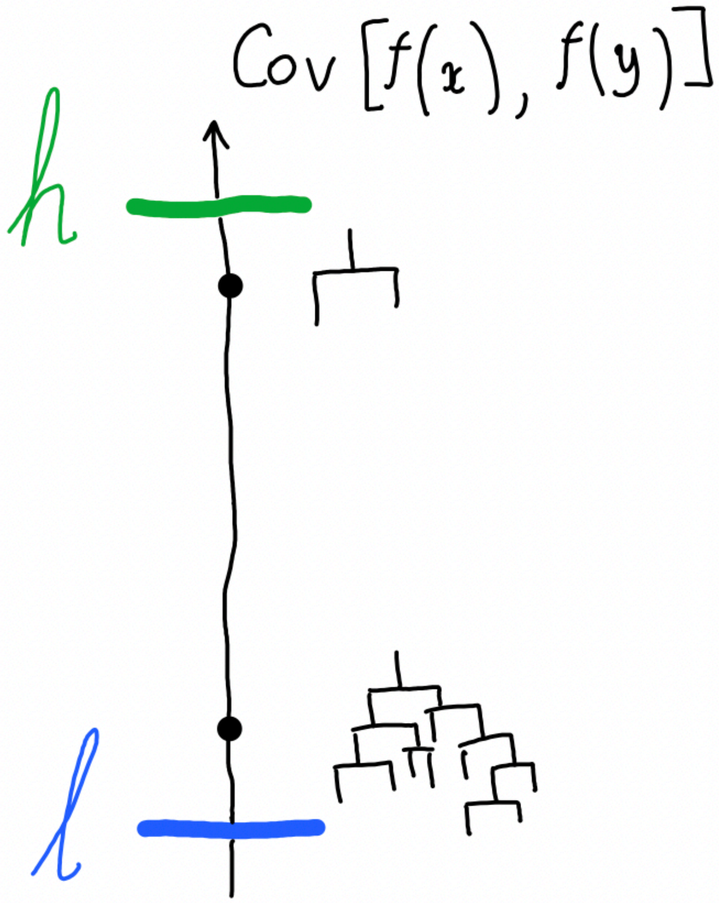
- Gli alberi profondi hanno poca probabilità
- ⇒ fisso la profondità massima
- Nelle foglie, metto un limite inferiore/superiore di come verrebbe il conto completo proseguendo l'albero
- Così ottengo un intervallo inf-sup più stretto alla radice



BART

Come faccio a non dover considerare tutti gli alberi possibili?

- Altro trucco: "comprimo" la formula ricorsiva guadagnando un livello gratis
- Gli intervalli così vengono piccoli ma non a precisione macchina
- Ultimo tocco: lim inf e sup sono valide funzioni di covarianza. Le interpolo con una regola derivata empiricamente con prove numeriche (DA FARE)

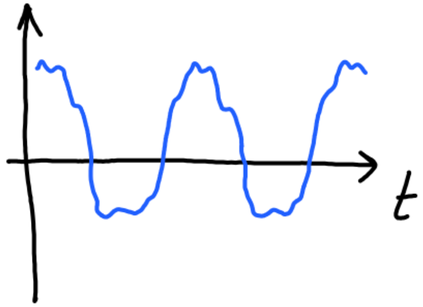


Le altre

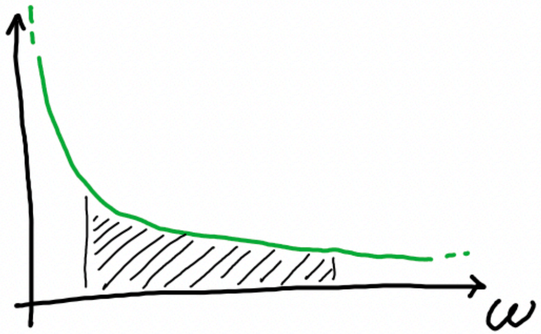
In breve

Zeta:

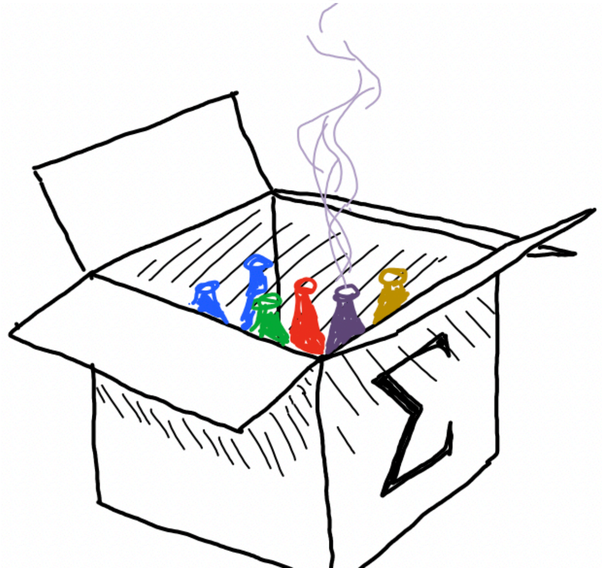
- Processi periodici o su sfere, spettro legge di potenza
- Approssima la Matérn circolare (incalcolabile)
- Ha correlazioni negative (a differenza della F-family che è > 0)

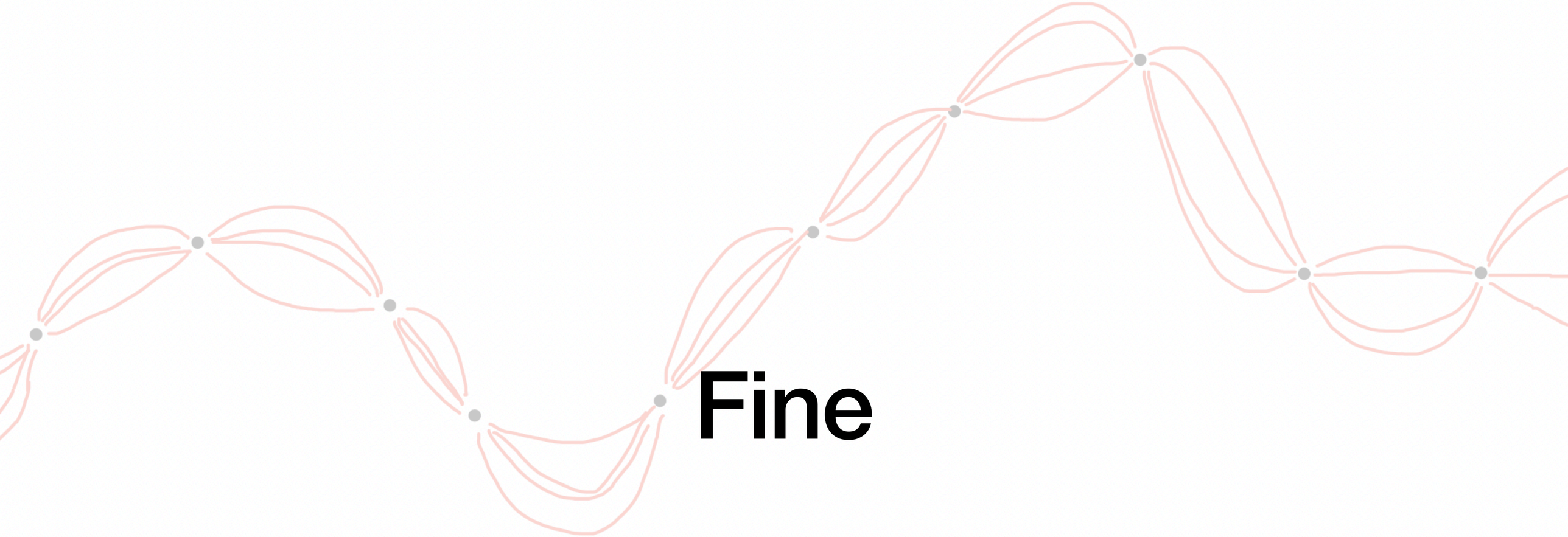


- Gamma: spettro legge di potenza (continuo)



- AR e altro: come BART, modelli già noti ma li infilo nei GP





Fine