Statistics for High Dimensional Data (and CompStat Lab) a.a. 2022/2023 (1st edition)

Prof. Francesco Finazzi francesco.finazzi@unibg.it

Prof. Alessandro Fassò alessandro.fasso@unibg.it

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

Lesson 7

NB: l'*alta risoluzione* è meno usata poiché è decisamente più costosa rispetto all'alta frequenza per via della necessità di installare un maggior numero di sensori nello spazio.

Towards spatio-temporal functional models

NB: per *alta frequenza* si intede dal punto vista temporale, mentre per *alta risoluzione* da quello spaziale.

- In many cases, data are observed at high frequency/resolution in at least one dimension (spatial or temporal)
- It usually happens with the temporal dimension
  - For instance, pollutant concentrations observed hourly or every 15 minutes
  - In general, high frequency observations (100, 1000, 10.000 per day)
- In space it is less common because a high resolution sampling is usually very expensive but…
  - In 3D space, one dimension may be sampled at higher resolution than the others

Esempio: l'inquinante non viene misurato dalla stazione una volta al giorno, ma *ogni 15 minuti*. Ciò porta ad avere dei dati in *alta frequenza*: 64 osservazioni/giorno vs 1 osservazione/giorno.

Esempio: i palloni sonda non possono essere lanciati ad alta risoluzione geografica (es. un lancio a Dalmine e un altro a Bergamo), tuttavia campionano ogni 10 metri, pertanto si ottiene un'alta frequenza *in altezza*.

NB: i modelli visti finora, ossia DCM e HDGM, vanno ancora bene per trattare questa particolare casistica? Limitiaoci per ora al caso solo *spaziale*: abbiamo trattato dataset nello spazio *bidimensionale* (no 3D), ossia con coordinate sul piano o sulla sfera. Volendo i modelli spaziali possono essere estesi alle tre dimensioni; sarà necessario determinare la correlazione tra coppie di punti la cui distanza è nello spazio tridimensionale, quindi otterremo un processo gaussiano 3D.

Towards spatio-temporal functional models

- Which are the problems with high frequency/resolution data?
  - Usually the original data set is very large and so the computational burden (of classic spatio-temporal models)
  - Data may be collected asynchronously over time (e.g., different monitoring stations may have different clocks)
  - Data may have large gaps over time (how does the Kalman Smoother perform in this case?)
  - Temporal correlation is usually very high (the Markovian model may explain the data but it is not very useful for prediction)

NB: noi per ora abbiamo utilizzato per i modelli precedenti un dataset con 194 stazioni e 365 tempi (1 osservazione/giorno). Se la frequenza aumenta, allora aumenta anche il numero di osservazioni giornaliere (da 1 a 2500, per esempio), pertanto è un problema di inferenza *ancora* trattabile con il DCM o HDCM? Se una singola iterazione dell'algoritmo EM prima richiedeva 3 secondi, adesso sono necessari 2500*3 secondi, ossia circa **2 ore**. Ciò è *insostenibile*.

NB: ogni sensore ha il suo clock. Questo è un problema per i modelli a *tempo discreto* visti precedentemente in quanto essi gestiscono un clock comune a tutte le stazioni di monitoraggio. Volendo è possibile fare delle approssimazioni per questi istanti che non si ossevano, ma ciò comporta l'introduzione di un errore che né il DCM né l'HDGM sono in grado di gestire.

Esempio: dal momento in cui la frequenza è molto alta, la temperatura al minuto t risulta essere *fortemente* correlata a quella misurata al minuto $t-1$. Pertanto, la matrice G della dinamica markoviana $\mathbf{z}(t) = G*\mathbf{z}(t-1) + \boldsymbol{\eta}(t)$ è prossima a 1 [G è simile alla matrice A di un modello state-space, ossia dice come devono essere combinati $z\_1(t-1)$, ..., $z\_n(t-1)$ per ottenere lo stato $z\_i(t)$] e la varianza di $\boldsymbol{\eta}(t)$ vicina a 0 poiché il cambiamento è minimo. Una matrice G prossima a 1 non solo rende il sistema instabile, ma non risulta nemmeno essere utile per la previsione (predizione fino a 5 minuti).

NB: un sensore che misura ogni minuto può improvvisamente guastarsi e rimanere non operativo per 3 giorni, pertanto si perdono ben 4320 osservazioni, una serie storica avente un "buco" non indifferente. Il filtro di Kalman (smoother) in questo caso va a stimare la media per i dati mancanti con un' incertezza molto alta, un comportamento che potrebbe non piacere.

NB: ci sono due possibili approcci:

- utilizzare un funzione polinomiale più o meno complessa per interpolare tutti i dati, rumore compreso;
- utilizzare una funzione più *smooth* (es. parabola) in modo tale da evitare sia di inseguire il rumore sia l'overfitting.

Differenza:

- analisi statistica classica: l'oggetto dell'inferenza è un **vettore** di parametri **θ**. Dei punti rappresentano un campione, quindi c'è la possibilità di calcolare la media;
- FDA: l'oggetto dell'inferenza è una **funzione** (oggetto matematico *continuo*). Dei punti rappresentano un campione di **una** funzione, quindi non può essere calcolata la media.

Functional data analysis (FDA)

- In FDA, the object of the statistical inference is a continuous function rather than scalar/vector values
- For instance, the temperature measured by a sensor over the 24h of the day can be described by a (smooth?) function
  - Independently of how many observations we take
  - Independently of where in time these observations are taken
- Which function or class of functions should we use?
  - The function should describe the «global» data pattern
  - In a way, the function filters out the data noisy
  - The researcher should be able to control the function smoothness

NB: la temperatura solitamente non varia rapidamente.

Vantaggio: possibilità di ricostruire la funzione senza introdurre nel modello errori di interpolazione dovuta all' irregolarità dei dati che sarebbero inevitabili se si uilizzasse l'approccio classico.

Esempio: in una stanza ci sono due sensori di temperatura che non campionano alla stessa frequenza, ma sono completamente *asincroni.* I dati vengono condivisi per creare la serie storica della temperatura nella stanza. Se immaginiamo di utilizzare un modello temporale (es. ARMA) per ricostuire la funzione, allora saremmo in difficoltà perché essi si basano su tempi discreti e frequenza di campionamento fissa. Volendo si potrebbero fare delle approssimazioni (suddivisione del dominio temporale (es. 24 ore) in sotto-intervalli e interpolazione per ogni sotto-intervallo), ma ciò comporta un errore non gestibile dai modelli visti finora.

Esempio: nella statistica classica l'insieme di N valori di temperatura misurato da un sensore in un giorno identifica il campione di dati temporalmente correlato. Per la FDA, invece, essi rappresentano il campione di un elemento, ossia la curva T(t) che si desidera stimare e non N valori provenienti da N variabili casuali temporalmente correlate (es. processo stocastico).

Splines

- Spline is a class of functions which allows to easily control the function smoothness
  - By selecting the proper basis functions
  - By selecting the proper knots
- B-spline basis are useful to describe non-periodic functions
  - Knots can be placed ad-hoc along the function domain (more knots where the function should change more rapidly)
- Fourier basis can describe periodic functions
  - Smoothness is controlled by the number of basis

NB: due sono i parametri che devono essere opportunamente scelti:

- numero di **basi**: funzioni la cui combinazione lineare descrive l'andamento della funzione da stimare in uno specifico intervallo del dominio. Un intervallo --> un set di coefficienti.
- numero di **nodi** e dove collocarli: all'aumentare del numero di nodi, aumenta la capacità di following del modello.

Esempio: possiamo mettere più nodi laddove so che la funzione varia più rapidamente. In questo modo avremo tanti intervalli e pertanto un set di coefficienti per ogni intervallo in modo tale da aumentare la qualità della stima nella regione di rapida variazione della funzione in cui si desidera una maggiore risoluzione. Se, per esempio, sappiamo che in università il riscaldamento viene acceso alle 7, allora andremo a collocare 4 nodi in corrispondenza di quell'orario (2 prima e 2 dopo) in quanto la funzione cambierà rapidamente dalle 7 in poi.

NB: con le basi di Fourier (seni e coseni con un frequenza via via più alta) non ci sono i nodi, pertanto la smoothness dipende solo dal numero di basi che vengono incluse per descrivere la funzione.

NB: nella stanza ci sono 4 sensori di temperatura collacati ai 4 angoli, ognuno dei quali campiona alla propria frequenza. Siamo interessati alla stima della temperatura al centro della stanza in un istante *t*, pertanto ci serve un **modello spazio-temporale**.

How to describe functional data in a space-time model?

- We now want to model the generic observation $y(\boldsymbol{s}, t, h)$
  - $\boldsymbol{s}$ and t are the usual spatial and temporal indexes
  - $h \in \mathbb{R}$ is the «functional» dimension (spatial or temporal)
- Examples
  - h could describe the continuous time within the day while t is the index of days
  - h could describe altitude in a 3D space while $\boldsymbol{s}$ describes the generic location across the globe

Idea: per ogni giorno (dominio *t*) abbiamo una funzione (dominio *h*), pertanto l'alta frequenza all'interno del giorno viene descritta mediante una funzione **continua**, mentre ciò che accade da un giorno al successivo dall'indice *t*. La scrittura *y*(**s**, *t*, *h*) permette di descrivere il processo complessivo tramite un *insieme* di serie storiche a bassa frequenza (piuttosto che mediante una sola ad alta frequenza), ognuna delle quali viene descritta tramite una funzione nel dominio *h*. Quello che dovremo fare sarà stimare la spline, nello specifico i coefficienti delle basi (sempre le stesse) per ogni istante *t* e per ogni punto **s**.

Esempio: supponiamo di misurare la temperatura in una stanza con un sensore che effettua 10000 misurazioni al giorno per un anno, quindi abbiamo ben 3.650.000 osservazioni. Fare previsione utilizzando un filtro di Kalman avvalendosi di questa mole di dati risulterebbe problematico.

The functional HDG model in D-STEM

- D-STEM implements the (univariate) functional version of the HDG model:

$$y(\boldsymbol{s},t,h) = \boldsymbol{x}(\boldsymbol{s},t,h)'\boldsymbol{\beta}(h) + \boldsymbol{\phi}(h)'\boldsymbol{z}(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t,h)$$
$$\boldsymbol{z}(\boldsymbol{s},t) = \boldsymbol{G}\boldsymbol{z}(\boldsymbol{s},t-1) + \boldsymbol{\eta}(\boldsymbol{s},t)$$

- $\boldsymbol{\phi}(h)$ are the basis functions, $\boldsymbol{z}(\boldsymbol{s},t)$ are the spline coefficients
- All details are in Wang et al. (2021) Journal statisti software

NB: η(**s**, *t*) è un processo gaussiano, pertanto la correlazione spaziale tra punti **s** è descritta dalla sua matrice di varianze e covarianze ρ(|**s** - **s**'|, **θ**).

PN: è possibile avere delle **covariate**. Per esempio, la temperatura in questa stanza il 29/12 alle ore 20 dipende dal fatto che il riscaldamento sia acceso o meno e dal numero di individui presenti. Non è obbligatorio che le covariate siano funzione di tutti e tre gli indici.

NB: β(*h*) è **funzionale**, ossia l'effetto della covariata non è globale, ma cambia nel dominio funzionale *h*. Per esempio, il numero di studenti in aula influenza diversamente la temperatura a seconda dell'ora del giorno *h*.

NB: la variabile latente **z**(**s**, *t*) serve per spiegare ciò che **x**(**s**, *t*, *h*)∗β(*h*) non esplicita, ossia la correlazione spazio-temporale tra punti

NB: Φ(*h*) e **z**(**s**, *t*) dipendono *solo* dal numero di basi e non dal numero di osservazioni. Se avessimo utilizzato il modello HDG, avremmo dovuto stimare la variabile latente **z** per ogni punto **s** per ogni istante *h* di ogni istante *t*, mentre con il modello funzionale *n_b* (numero di basi) una **z** per ogni punto **s** per ogni istante *t* poiché il dominio funzionale *h* viene descritto dalle basi (comuni) Φ(*h*) (noto) --> **semplificazione**.

Idea: la variabile latente **z**(**s**, *t*)interagisce con le basi per descrivere ciò che **x**(**s**, *t*, *h*)∗β(*h*) non esplicita. Tolte le covariate, i 4 sensori agli angoli della stanza hanno dei residui *y*(**s**, *t*, *h*) - **x**(**s**, *t*, *h*)∗β(*h*) che risulteranno essere spazialmente correlati, pertanto essi vengono descritti utilizzando una variabile latente **z**(**s**, *t*) avente una dinamica markoviana rispetto all'indice *t* (da un giorno all'altro tutto può cambiare, cosa che invece non accade nel dominio funzionale *h* --> G != 1) correlata spazialmente.

NB: una spline si ottiene combinando linearmente le basi tramite i coefficienti (da stimare), ossia **z**(**s**, *t*). Essi sono costanti in *h*, ma cambiano in *t* (giorni diversi --> coefficienti diversi) e in **s** (punto spaziale diverso --> coefficienti diversi). Tali coefficienti hanno una dinamica markoviana (il presente dipende dal passato) e un'innovazione spazialmente correlata.