1 Riassunto: trend, patterns, modelli

Riassumiamo prima di tutto alcune idee concettuali emerse dall'indagine "a mano" svolta nella scorsa lezione. Abbiamo cercato di farci guidare dall'intuito, ovvero di realizzare con metodi matematici quello che l'intuito umano, il senso umano di analogia, analisi e sintesi, ci portava a immaginare circai valori successivi della serie storica. Ciascuno di noi avrebbe potuto proseguire con carta e penna la serie storica, secondo il proprio intuito. Abbiamo quindi cercato di analizzare questo imp'agabile processo mentale.

In esso abbiamo rilevato alcuni elementi. Uno è la ricerca di *pattern*, forme geometriche caratteristiche, *ricorrenti*. Le tre grandi oscillazioni dell'ultimo anno e mezzo sono piuttosto evidenti e simili tra loro. La loro somiglianza va però intesa nello stesso senso in cui somigliano tutte le lettere A maiuscolo scritte da esseri umani. Perciò non si tratta di una ripetizione esatta di una forma, e nemmeno di una ripetizione sporcata da rumore: da una oscillazione all'altra cambia leggermente il periodo di oscillazione, cambia l'ampiezza (distanza tra massimo e minimo), cambia l'altezza da cui parte l'oscillazione... nonostante questo, la mente umana le individua come forme simili, quindi vede un *pattern ricorrente*.

Un secondo elemento è la nostra spontanea ricerca di una *struttura*, una *logica*, addirittura un *modello*. Nell'accezione più semplice questo può significare che riconosciamo una crescita generica, oppure una crescita ma con saturazione, o in altri casi una crescita sempre più rapida. Quindi il cosidetto *trend*. In un'accezione più evoluta questo può significare varie cose, ad esempio che riconosciamo degli andamenti che rispecchiano avvenimenti o situazioni o stato economico ecc.; oppure che scriviamo un modello del tipo dati=retta+funzione_periodica+noise; oppure infine che arriviamo addirittura ad immaginare delle relazioni causa-effetto.

Tutti questi elementi ci aiutano a prevedere, ad immaginare come proseguirà la serie storica. Ad esempio, se abbiamo identificato un *trend*, sufficientemente semplice da essere racchiuso in una formula estrapolabile, siamo portati ad ipotizzare che nel futuro immediato si continui con lo stesso trend. Se, sovrapposto a questo trend, abbiamo individuato un *pattern* ripetitivo, siamo portati a pensare che si ripeterà. Se addirittura avessimo un modello con certi predittori (fattori, cause) e la serie storica stessa come variabile dipendente (effetto), useremmo i valori dei predittori, se in nostro possesso, per predire i valori futuri della serie.

Nell'esempio della serie *Bulgari* abbiamo applicato l'idea di trend e di

pattern ricorrente. Non avendo trovato una funzione di trend semplice su tutto l'itervallo temporale in nostro possesso (dati da Gennaio 2004 in poi), abbiamo ristretto i dati ad un periodo finale di circa un anno e mezzo e lì abbiamo cercato la retta di regressione, come funzione che descrivesse il trend. Immagineremo quindi che, nell'immediato futuro, ad esempio nei successivi 30 giorni, quel trend si conservi.

In parallelo o ancor prima abbiamo notato la presenza di un pattern, una forma di oscillazione, che si ripete tre volte. Però sulla serie originale essa è complicata dalla crescita media descritta dal trend, oltre che da altri fattori. Abbiamo allora sottratto il trend, con l'effetto che le oscillazioni ora sono "in piano". Nonostante questo, la loro ampiezza cresce nel tempo. Le abbiamo allora riscalate (la funzione $c(t)$) in modo che avessero la stessa ampiezza. Poi abbiamo individuato un periodo sufficientemente rappresentativo delle tre oscillazioni, alla ricerca di una forma abbastanza comune ad esse. Scelto il periodo, le abbiamo sovrapposte ed abbiamo cercato una funzione regolare che le interpolasse. Questa funzione è la forma idealizzata del pattern.

Ora dobbiamo tornare indietro, mettere tutto insieme ed effettuare la previsione.

2 Passi conclusivi sull'esempio

Nell'esempio Bulgari, ristretto ad una finestra finale di circa un anno e mezzo, abbiamo usato il modello

$$d(t) = (at + b) + F(t)c\,(t) + \varepsilon\,(t)$$

e con R abbiamo trovato la retta $(at + b)$, la retta $c\,(t)$ ed una possibile funzione $F(t)$ (su cui torneremo tra un momento). Allora, se t_f indica il giorno finale dei dati in nostro possesso, e vogliamo prevedere i valori della serie nei giorni $t_f + 1$, $t_f + 2$, ... , $t_f + 30$, calcoliamo sempicemente

$$\widehat{d}(t_f + k) = (a\,(t_f + k) + b) + F(t_f + k)c\,(t_f + k)\,, \quad k = 1, ..., 30.$$

Ecco come appare la sequenza di comandi nel suo complesso.

- C'è una prima fase di caricamento dati, ricerca della retta di regressione

$$r(t) = (at + b)$$

della retta di appiattimento $c(t)$ e della funzione

$$\widehat{F}(t) := \frac{d(t) - (at + b)}{c(t)}.$$

Questi vengono chiamati nel seguito *r.t, c.t, F.hat.*

```
B <- read.table(file="C:/Franco/dottoratoING/time_series/ordine/bulgari.txt")
lenB<-length(B[,1])
B<-B[1:lenB,1]
$$ questa è la serie dal 1/1/04 al 12/10/2006
lag <- 30
B0 <- B[1:(lenB-lag)]
lenB0<-length(B0)
ts.plot(B0)
$$ questa è la serie dal 1/1/04 al 30/8/2006
x0<-1:lenB0
x<-1:lenB
reg <- lm(B0[200:lenB0] ~x[200:lenB0])
abline(reg$coefficient)
r.t<-(reg$coefficient[1]+reg$coefficient[2]*x)
y <- B0-r.t[1:lenB0]
ts.plot(y)
est <- read.table(file="C:/Franco/dottoratoING/time_series/ordine/estremi.txt")
xx<-est$V1
yy<-est$V2
reg.est <- lm(yy ~xx)
abline(reg.est$coefficient)
c.t<-(reg.est$coefficient[1]+reg.est$coefficient[2]*x)
F.hat<- y[200:lenB0]/c.t[200:lenB0]
ts.plot(F.hat)
```

- Viene trovata una funzione periodica $F(t)$. che su tre periodi somigli a $\widehat{F}(t)$. Descriveremo sotto i comandi.

- Si calcolano le previsioni

$$\widehat{d}(t_f + k) = (a\,(t_f + k) + b) + F(t_f + k)c\,(t_f + k)\,, \quad k = 1, ..., 30.$$

Nel seguito le previsioni verranno indicate con *B.hat*. In realtà, per visualizzare varie informazioni simultaneamente, chiamiamo *B.hat* anche i valori del modello sui tre periodi prima della previsione. Con F indichiamo i valori della funzione periodica trovata col metodo che spiegheremo sotto.

```
B.hat<-1:lenB
B.hat[1:(lenB0-3*per)]<-B0[1:(lenB0-3*per)]
F.4P<-c(F,F,F,F)
B.hat[(lenB0-3*per+1):lenB]<-r.t[(lenB0-3*per+1):lenB]+F.4P[1:(3*per+lag)]*c.t[(lenB
3*per+1):lenB]
ts.plot(B0)
points( B.hat ~x, col = "red")
```

2.1 Analisi dei residui

Un'altra cosa che si può fare subito è l'analisi della distribuzione statistica dei residui

$$\varepsilon\,(t) = d(t) - [(at + b) + F(t)c\,(t)]$$

$\varepsilon\,(t)$ allo scopo di fornire un'indicazione di *intervallo di confidenza* per la previsione. Supponiamo che da un istogramma ed un Q-Q plot dei residui risulti accettabile l'ipotesi di gaussianità (altrimenti bisogna usare quantili spermentali o di distribuzioni più complicate). Detta $\widehat{\sigma}$ la deviazione standard dei residui, fissata una confidenza, es. 95%, i valori di $\varepsilon\,(t)$ stanno nell'intervallo $[-\widehat{\sigma}q_{0.975}, \widehat{\sigma}q_{0.975}]$ con probabilità 0.95. Quindi potremo dichiarare per i valori futuri della serie storica:

$$\widehat{d}(t_f + k) = (a\,(t_f + k) + b) + F(t_f + k)c\,(t_f + k) \pm \widehat{\sigma}q_{0.975}$$
$$k = 1, ..., 30, \text{ con livello di confidenza } 95\%.$$

Possiamo anche tracciare graficamente due bande di confidenza, come abbiamo fatto per la retta di regressione. In realtà è naturale aspettarsi che l'intervallo di confidenza debba allargarsi al crescere del tempo futuro. Arrivare a tale realismo richiede un'analisi più complessa del modello stesso e comunque non è univoca e semplice, quindi non la discutiamo e ci accontentiamo dell'intervallo dato, come prima indicazione.

2.2 Sulla funzione $F(t)$

Riguardo però all'esempio specifico dell'ultima lezione, va rilevato una sorta di errore che abbiamo commesso. E' stato osservato da un partecipante che la funzione $F(t)$ trovata con *loess* non è periodica. E' questione di intendersi. Se rinunciamo alla continuità, ogni funzione definita su un periodo può essere usata come germe per una funzione periodica: basta ripeterla - eliminando l'ultimo valore, che andrebbe in conflitto col primo. Se invece vogliamo la continuità, bisogna che la funzione F sul periodo dato abbia lo stesso valore agli estremi. Bene, la funzione trovata con *loess* non è continua *e* periodica.

Siccome è molto difficile trovarne una periodica, si può decidere di rinunciare alla continuità. Si può anche cercare di allieviare il peso psicologico di questa scelta, con piccoli trucchi: ad esempio di cerca con *loess* una funzione interpolante in un intervallo un po' più piccolo del periodo, e poi si unisce il grafico con un segmento di retta.

2.3 Ricerca ottimale della funzione periodica

Descriviamo un procedimento di ricerca di una $F(t)$ periodica che, pur non essendo l'unico ed avendo comunque un certo livello di arbitrarietà, contiene al suo interno un tentativo di ottimizzazione rispetto a certi parametri e ad un certo indice di performance.

La funzione *periodicSpline* del package *splines* cerca una regressione con polinomi locali che sia periodica. Purtroppo il grado di regolarità non è facile da impostare, per cui la useremo in un modo un po' arzigogolato. Ecco la lista di comandi:

```
require(splines)
deg<-20
per<-145
spn<-0.07
ff<-length(F.hat)
F1<-F.hat[(ff-3*per+1):(ff-2*per)]
F2<-F.hat[(ff-2*per+1):(ff-per)]
F3<-F.hat[(ff-per+1):ff]
Fmean <- (0.5*F1+F2+1.5*F3)/3
x.per<-1:per
F.loess<-loess(Fmean~x.per, span = spn, degree = 2)
```

```
F.nonper<-predict(F.loess, data.frame(time =seq(1, per, len = per)), se
= TRUE)$fit
kn<-seq(1, per, per/deg)
y.knots<-predict(F.loess, data.frame(time = kn), se = TRUE)$fit
F.spline <- periodicSpline( kn, y.knots, period = per)
F<-predict(F.spline, seq(1, per, len = per))$y
F.hat.3P<-F.hat[(ff-3*per+1):ff]
F.nonper.3P<-c(F.nonper,F.nonper,F.nonper)
F.3P<-c(F,F,F)
x.3P<-1:(3*per)
ts.plot(F.hat.3P)
points( F.nonper.3P ~x.3P, col = "black")
points( F.3P ~x.3P, col = "red")
err<-sum(abs(F.hat.3P-F.3P))/(3*per)
err
```

Il significato è questo. Caricato il package *splines*, fissiamo tre parametri chiave che si capiranno tra un momento, ovvero *deg, per* e *spn.* E' rispetto ad essi che svolgeremo l'ottimizzazione. *per* è il periodo. Ad occhio avevamo scelto periodo 145. Ora cercheremo il periodo che produce l'errore minore. Prendiamo poi la traccia *F.hat* sugli ultimi tre periodi, divisa nei tre pezzi. (*F1,F2,F3*). Di questi calcoliamo il tracciato medio pesato, dando più importanza all'ultima oscillazione:

$$Fmean < -(0.5 * F1 + F2 + 1.5 * F3)/3.$$

Si vedrà molto più avanti che quest'idea ricorre nei metodi iterativi di tipo smoothing esponenziale.

Eseguiamo poi *loess* di *Fmean.* Questo produce un grafico regolare che interpola (nel senso della regressione) polinomialmente e localmente *Fmean.* La regolarità di questo grafico si può modificare tramite il valore del parametro *spn.* Il valore di default è 0.75, che produce funzioni molto regolari (cioè con poche oscillazioni su piccola scala). Più *span* è piccolo, più la funzione prodotta da *loess* oscilla. E' difficile stabilire inequivocabilmente se sia meglio una funzione più o meno oscillante. Se si permette al risultato di *loess* di oscillare molto, questo risulta molto più aderente a *Fmean* e quindi produce alla fine un errore minore. Però in un certo senso così si rincorrono le sotto-oscillazioni particolari di *Fmean*, che non è detto debbano ripetersi. In altre parole, mentre dalla ripetitività delle tre grandi oscillazioni siamo portati a

credere che tali oscillazioni abbiamo qualche ragione strutturale, e quindi ci aspettiamo che si ripetano, le sotto-oscillazioni su scala minore potrebbero invece essere fenomeni contingenti e più irregolari, salvo che le si veda in tutte e tre le grandi oscillazioni. Avendo scelto di pesare di più l'ultima oscillazione, le sue sotto-oscillazioni pesano di più e quindi un parametro *spn* molto piccolo riduce l'errore.

Calcolata *loess*, questa non è periodica. Allora la si interpola nuovamente tramite *periodicSpline*, ottenendo così una funzione periodica che le somiglia il più possibile. Il grado di regolarità del risultato di *periodicSpline* viene deciso dal parametro *deg*. Più esso è alto, più permettiamo al risultato si *periodicSpline* di oscillare. Dato il risultato di *loess*, se *deg* è alto, il risultato di *periodicSpline* somiglierà molto al precedente.

Abbiamo infine plottato i risultati e calcolato l'errore tra l'interpolazione periodica ed i dati originari, sui tre periodi. C'è arbitrarietà di scelta circa la funzione errore: noi abbiamo preso una media uniforme degli errori assoluti giornalieri, ma si poteva ad esempio persare di più l'errore sull'ultimo periodo.

Il gruppo di istruzioni va fatto girare un po' di volte, al variare di alcune scelte di *deg, per* e *spn*, alla ricerca di valori che rendano più basso possibile l'errore, oppure che soddisfino altri criteri, ad esempio che il grafico previsto corrisponda maggiormente a ciò che avremmo fatto ad occhio.

Dalla raffigurazione dei dati e delle loro interpolazioni, si nota purtroppo un difetto. Questi metodi di interpolazione automatica (*loess* e *periodicSpline*) appiattiscono moltissimo i picchi alti e stretti. Il massimo dell'interpolata è molto minore del massimo dei dati. Questo è più evidente se si cercano curve molto regolari quasi sinusoidali). Invece è meno evidente se si accetta un elevato grado di oscillazione nelle interpolazioni. Se si permette a *loess* di seguire il più possibile le sotto-oscillazioni dei dati (soprattutto del terzo periodo), allora l'interpolata riesce anche ad alzarsi un po' di più lungo il picco. Questa è una ragione a favore di un parametro molto più piccoli per *span*, ad esempio 0.07.

Si suggerisce di provare i seguenti due set alternativi di parametri, per vedere differenze marcate (e magari provarne tanti altri). Ci sono buone ragioni per ritenere interessante il risultato in entrambi i casi.

$$deg = 20, \quad per = 145, \quad spn = 0.07 \rightarrow err = 0.204$$
$$deg = 8, \quad per = 145, \quad spn = 0.5 \rightarrow err = 0.224$$

3 Giudizio a priori ed a posteriori sul metodo

Circa un giudizio a priori, cioè senza i dati reali del periodo previsto, c'è poco da dire. Avendo almeno un'altro metodo, si innescherebbe qualche ragionamento comparativo. Senza altri metodi, la comparazione avviene psicologicamente rispetto alla previsione ad occhio che avremmo fatto senza matematica. Di fatto questa comparazione l'abbiamo già effettuata strada facendo, giudicando la bontà dei vari passi. Si percepisca però il disagio dell'operatore che debba effettivamente impiegare risorse finanziarie in base alle proprie previsioni: non ha in mano niente se non l'intuito a conforto delle sue scelte.

Qualcosa in più si può fare a posteriori, cioè potendo confrontare previsioni e dati reali. Possiamo visualizzarli insieme ed apprezzare ad occhio se la previsione sembra ragionevole, ottima, fallimentare... ma, riprendendo una delle domande finali della volta scorsa:

- una volta visualizzata la previsione ed i dati veri, come giudicheremo se la previsione era buona? Infatti la previsione non coinciderà mai coi dati, quindi bisogna valutare se lo scarto è piccolo o grande.

Lo scarto si può calcolare. Ma come giudichiamo se è piccolo o grande? L'unica cosa che possiamo fare è vedere se il modello che abbiamo costruito per le ultime tre oscillazioni resta ugualmente buono nei successivi 30 giorni. In altre parole, in assenza di altri metodi, non possiamo dare giudizi comparativi, né tanto meno giudizi assoluti, ma solo dire se i dati seguono ancora il modello scelto, nei limiti di approssimazione che esso aveva sui dati passati. Ci sono vari modi di effettuare questa verifica, non equivalenti, matematicamente ben fondati a seconda che valga una qualche ipotesi. uno per noi semplice per non introdurre idee nuove è quello di usare le bande di confidenza descritte sopra. Ci poniamo il problema di vedere se il dati veri stanno nell'intervallo di confidenza previsto al 90%. Ipotizziamo che ad ogni istante il rumore sia indipendente. Su 100 dati previsti, mediamente dovremmo aspettarcene 10 fuori dai limiti. Su 30, ce ne aspettiamo mediamente 3. Vediamo allora quanti effettivamente stanno fuori. Non è un vero e proprio test con tanto di valore p, ma è comunque un'indicazione che proviene dai numeri e non dall'occhio.

3.1 Aggiustamento iterativo

Se davvero stiamo usando il metodo in pratica e viviamo la tensione dell'incertezza sul buon uso di risorse ad es. economiche, viene spontaneo di aggiustare il tiro strada facendo. Fin tanto che i dati stanno, salvo eccezioni dell'ordine di una su dieci, entro i limiti di confidenza al 90%, non aggiustiamo il tiro. Quando questo non è più vero, significa che il modello non è più appropriato. Si può allora cercare una nuova retta di regressione, un nuovo raddrizzamento ed una nuova funzione periodica adatte all'insieme di tutti i dati attualmente in nostro possesso.

Chiaramente, se però i dati manifestano una indiscutibile rottura con la struttura di trend e pattern usata fino ad ora, vuol dire che la natura economica è cambiata, nuovi pattern sono probabilmente in arrivo o in formazione, e non abbiamo elementi per predirli.

4 Mistura di modelli

Questo è un argomento spinoso a cui accenniamo solamente. Nell'identificazione del modello previsionale abiamo operato alcune scelte, evidenziando il grado di approssimazione e soggettività applicato. Si potrebbero *probabilizzare* alcuni parametri che abbiamo scelto arbitrariamente. Ad esempio, invece che scegliere $span = 0.07$ oppure 0.5, si potrebbe stabilire che esso è una v.a. con una certa distribuzione, es. gaussiana di media 0.3 e deviazione 1.5. I valori previsti diventano così variabili aleatorie, di cui possiamo fornire la media oppure gli intervalli di confidenza. Queste metodologie vengono usate nel problema delle previsioni atmosferiche. Una loro utilità quindi esiste, ma certamente sono più complessi.

5 Analisi multiscala

Infine, potremmo esaminare la *serie storica dei residui* (sul periodo noto), cercando in essa dei pattern, una struttura. Plausibilmente non troveremo pattern apprezzabili alla grande scala temporale già esaminata, altrimenti essi sarebbero rientrati nel modello già identificato. Dobbiamo quindi guardare i dati su una scala più breve. Se riusciamo a formulare un modello dei residui su una scala più breve, possiamo utilizzarlo nello stesso modo del modello su grande scala, per prevedere appunto i residui del futuro.

Un'altro aspetto multiscala consiste nell'esaminare i dati *raggruppati settimanalmente* o mensilmente (o su altra finestra temporale). Questo però è un modo alternativo di fare ciò che abbiamo già fatto, ovvero trovare un modello su grande scala. Forse dopo il raggruppamento certi pattern sono persino più evidenti e la ricerca della funzione periodica può essere facilitata.