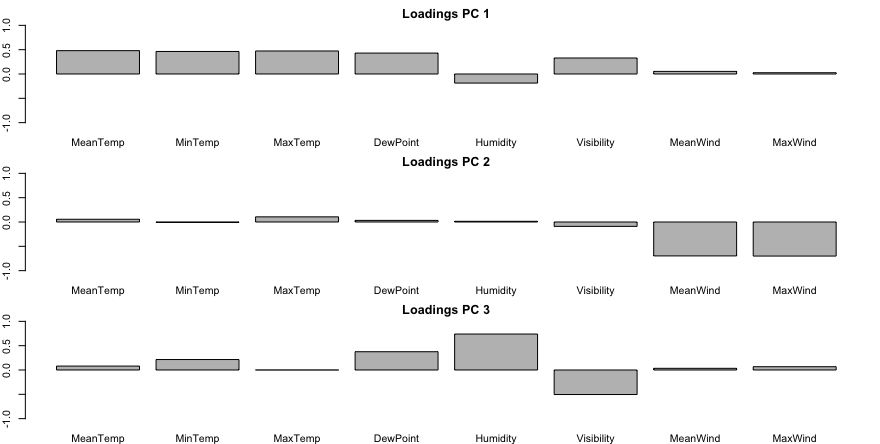
TDE 07/09/20

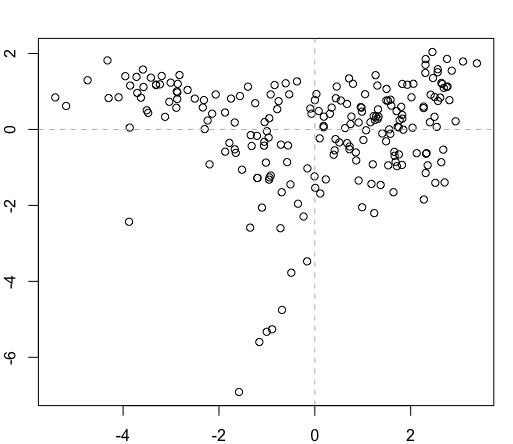
Problema 1

1. Vogliamo analizzare I dati attraverso una PCA quindi cerchiamo le direzioni come combinazioni della variabili originali lungo le quali i dati sono più varaibili ovvero a =argmin(a in Rp) Var(a’X). In particolare cerchiamo un Sistema di riferimento che catturi la variabilità del nostro dataset in maniera decrescente lungo le covariate e in modo che le nuove variabili siano non correlate. La soluzione a questo problema sono gli autovettori della matrice di covarianza campionaria e la parte di varianza spiegata da ognuno di essi è rappresentata dall’autovalore a cui è associato. In particolare per decidere se utilizzare la PCA sul dataset che ci arriva o sullo standardizzato guardiamo le variane delle singole variabili. Notiamo che la variabile umidità ha una varianza che è 3 volte quelle delle altre variabili quindi per studiare la PCA come valutazione della correlazione delle variabili più che valutazione della magnitudine delle varianze andiamo a utilizzare la PCA sulle unità standardizzate.

Possiamo visualizzare i loading delle prime 3 componenti principali che spiegano il 94% della variabilità totale del dataset

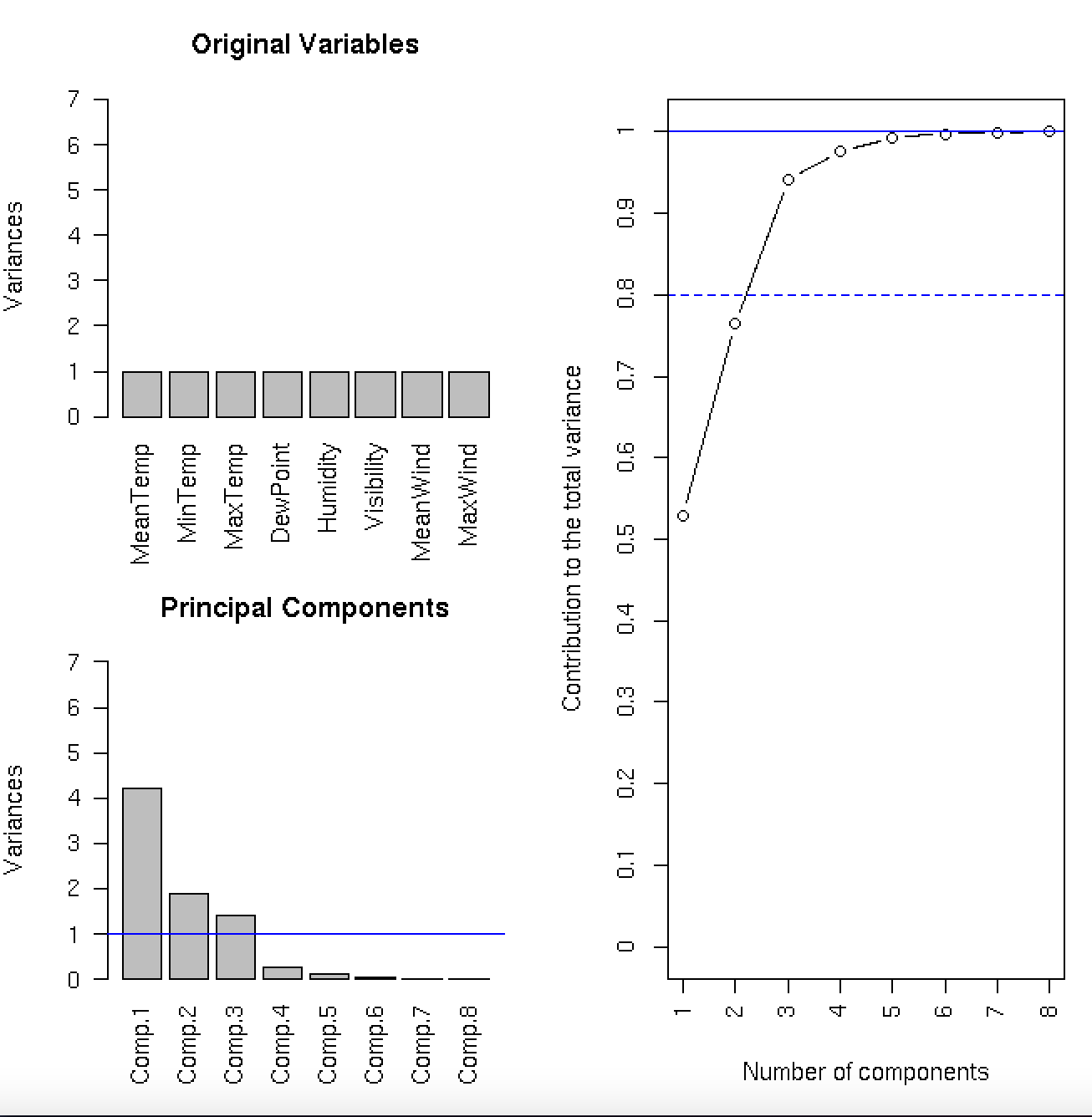


La prima componente principale possiamo notare che è una media tra i vari indici di temperatura con una piccola contribuzione data anche dalla visibilità e un contrasto rispetto all’umidità. La seconda componente principale rappresenta invece una media dei venti e la terza infine rappresenta un contrasto tra la visibilità e la temperatura percepita.



Ci aspettiamo che nel primo quadrante in alto si trovino i giorni in cui abbiamo forte vento ed alta temperature, nel quadrate in alto a sinistra i giorni in cui ci sono alti verni ma temperature tendenzialmente più basse della media nel quadrante in basso a sinistra venti e temperature più basse della media e in quelli in basso a destra alte temperature e basso vento.

Abbiamo in particolare che i primi giorni del mese tendono a trovarsi nel quadrante in alto a sinistra per poi sposarsi con il variare dei giorni in senso antiorario nei quadranti



In questo caso ridurrei la dimensione a tre componenti per riuscire a spiegare almeno 80% della variabilità totale. La varianza spiegata dalle singole componenti è data appunto dal dall’ autovalore associato all’autovettore corrispondente alla PC.

E in particolare sono [0.529159, 0.2350544, 0.1765319]

1. Quello che adesso vogliamo fare è data una nuova proiezioni proiettarla sul piano creato dalle prime 3 componenti principali

[ 46.30318, -8.708581]

Problema 2.

1. Abbiamo due popolazioni bivariante e indipendenti dato che i campioni non hanno un’ altissima numerosità supponiamo che siamo normali e abbiano la stessa struttura di covarianza. Queste assunzioni sono verificate attraverso un test approssimato di controllo in ogni direzione della normalità e dal calcolo delle due covarianze che sono molto simili. Valutiamo come gruppo 1 i dati di Sunshine Corp. E griuppo 2 i dati di Candle. Inc. Possiamo a questo punto utilizzare un test basato su queste assunzioni e sulla quantità pivotale T2 = (1/n1+1/n2)\* [Xmean1 -Xmean2- (mu1 - mu2)]’ \* Spooled ^(-1)\* [Xmean1 -Xmean2- (mu1 - mu2)] che è distribuita come (n1+n1-2)/n1+n2-p-1 Fisher con gradi di libertà p e n1+n2-1-p. Dalla quantità pivotale possiamo di conseguenza andare a fare il test e a calcolarne il p-value. In particolare il test ha come ipotesi nulla l’uguaglianza delle due media e ipotesi alternativa la loro differenza. Dal test rifiutiamo l’ ipotesi nulla e in particolare abbiamo evidenza statistica per dire che le due medie sono differenti
2. Il p-value del test descritto sopra è 3.09412e-08 questo significa che per ogni ragionevole livello del test la nostra ipotesi nulla verrà rifiutata.
3. In particolare possiamo andare a spiegare il nostro test in base a intervalli di confidenza di tipo Bonferroni sulle due componenti della media. I nostri intervalli saranno del tipo

[(Xmean1 Xmean2)i +- sqrt(t(1-alpaha/4)(n1+n2-1-p)\* (Spooled)I,i(1/n1+1/n2)]

IC.X1 [-2.4906343, -0.9277657]

IC.X2 [00.8841826, 2.5026174]

Possiamo vedere che entrambe le componenti differiscono nei due gruppi non contenendo entrambe lo zero.

1. Dobbiamo testare adesso una combinazione lineare delle medie in particolare dato a= (1, -1)’

[a’\*(Xmean1 Xmean2) + sqrt(t(1-alpha)(n1+n2-1-p)\*a’\* (Spooled)\*a(1/n1+1/n2)]

[-2.4906343 -0.9277657]

Si notiamo infatti che nell’intervallo di confidenza al 95% per questa combinazione lineare lo zero non è contenuto e i valori sono negativi

Problema 3

1. Dobbiamo utilizzare un modello ANCOVA che ci permetta di avere due modelli diversi in base all’utilizzo di un tipo di lievito o un altro. Il modello può essere rappresentato come:

rising = b[0] + b[1]\*time +b[2]\* time^2 +epsilon

dove epsilon si suppone siano normali a media nulla indipendenti e identicamente distribuite.

b[1] e b[2] hanno valori diversi in base a che lievito viene usato. In particolare per differenziarli utilizziamo una variabile che sia uguale a 1 nel caso di lievito sv e zero nel caso di lievito by e valutiamo le beta attraverso un least square method

* Caso lievito by:

Beta0 = 0.9839522

Beta1= 0.0141076

Beta 2= 0.0117339

* Caso lievito sv

Beta0 = 0.9839522

Beta1= 0.06704227

Beta 2= -0.0005003801

Le assunzioni vengono verificate attraverso i grafici dei residui e uno shapiro test

1. Dobbiamo preformare dei test su combinazioni lineari delle beta:

* Il primo test riguarda le differenza tra le Beta1 e Beta2 dei due modelli quindi ci concentriamo sulle beta dove è presente la variabile che indica i diversi tipi di lievito

Data beta : (Intercept) time I(time^2) time:dummy dummy:I(time^2)

0.98395224 0.01410756 0.01173395 0.05293471 -0.01223433

C1= [0,0,0,1,0

0,0,0,0,1]

C2= [0,0,1,0,0]

C3 = [0,0,0,0,1]

Ora dobbiamo testare se C\*beta =0 che sarà la nostra ipotesi nulla oppure C\*beta !=0

Possiamo utilizzare un test basato su F0 = (C\*beta)’\*(C(Z’Z)^(-1)C’)^(-1) (C\*beta) / p\*S

S = SSres^2/n-r+1

Z = [1 time time^2 time\*dummy time^2\*dummy]

Il risultato per primo test ci dice che abbiamo una dipendenza dal tipo di lievito. Dal secondo e dal terza test abbiamo evidenza statistica per dire che in entrambi i casi il polinomio legato al tempo sarà di secondo grado

1. Possiamo ridurre il modello dato che attraverso lo stesso test fatto precedentemente utilizzando C= [0,1,0,0,0] non abbiamo evidenza statistica per dire che il solo tempo nel caso di by sia legato alla crescita quindi possiamo proporre un modello ridotto:

rising = b[0] + b[1]\*time +b[2]\* time^2 +epsilon

dove epsilon si suppone siano normali a media nulla indipendenti e identicamente distribuite e b[1] è diversa da zero solo nel caso di utilizzo di sv.

I coefficienti in questo caso saranno:

* Caso lievito by:

Beta0 = 1.01146695

Beta1= 0

Beta 2= 0.01298844

* Caso lievito sv

Beta0 = 1.01146695

Beta1= 0.05665001

Beta 2= 0.0003491497

1. Possiamo decidere quale lievito sia meglio usare basandoci su un intervallo di confidenza per le medie in questo caso quello che andiamo a fare a dato il nuovo tempo formare la nostra nuova matrice disegno chiamata z0 e valutare un intervallo di confidenza del tipo

IC(zo\*beta)= [ zo\*beta +- Ssqrt(zo’\*(Z’Z)^(-1)\*t(1-alpha/2)(n-r-1)]

Nel caso di sv

[1.126164, 1.104025, 1.148302]

Nel caso di by

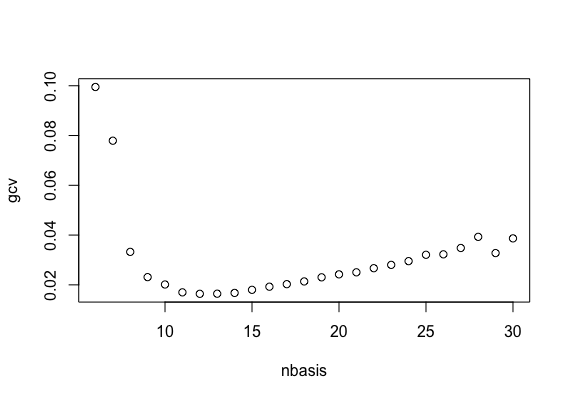
[1.063421, 1.034237, 1.092604]

Consiglierei di utilizzare sv

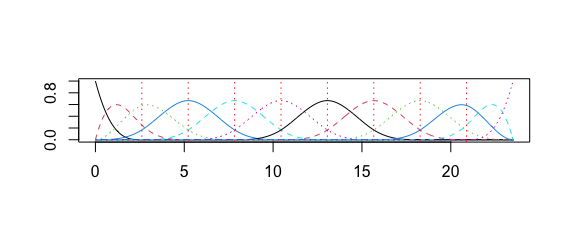
Problema 4

1. Sappiamo di avere dei dati caratterizzati da errore quindi dobbiamo stare attenti a valutarli con un numero di basi che non vada ne a fare undersmoothing né oversmoothing

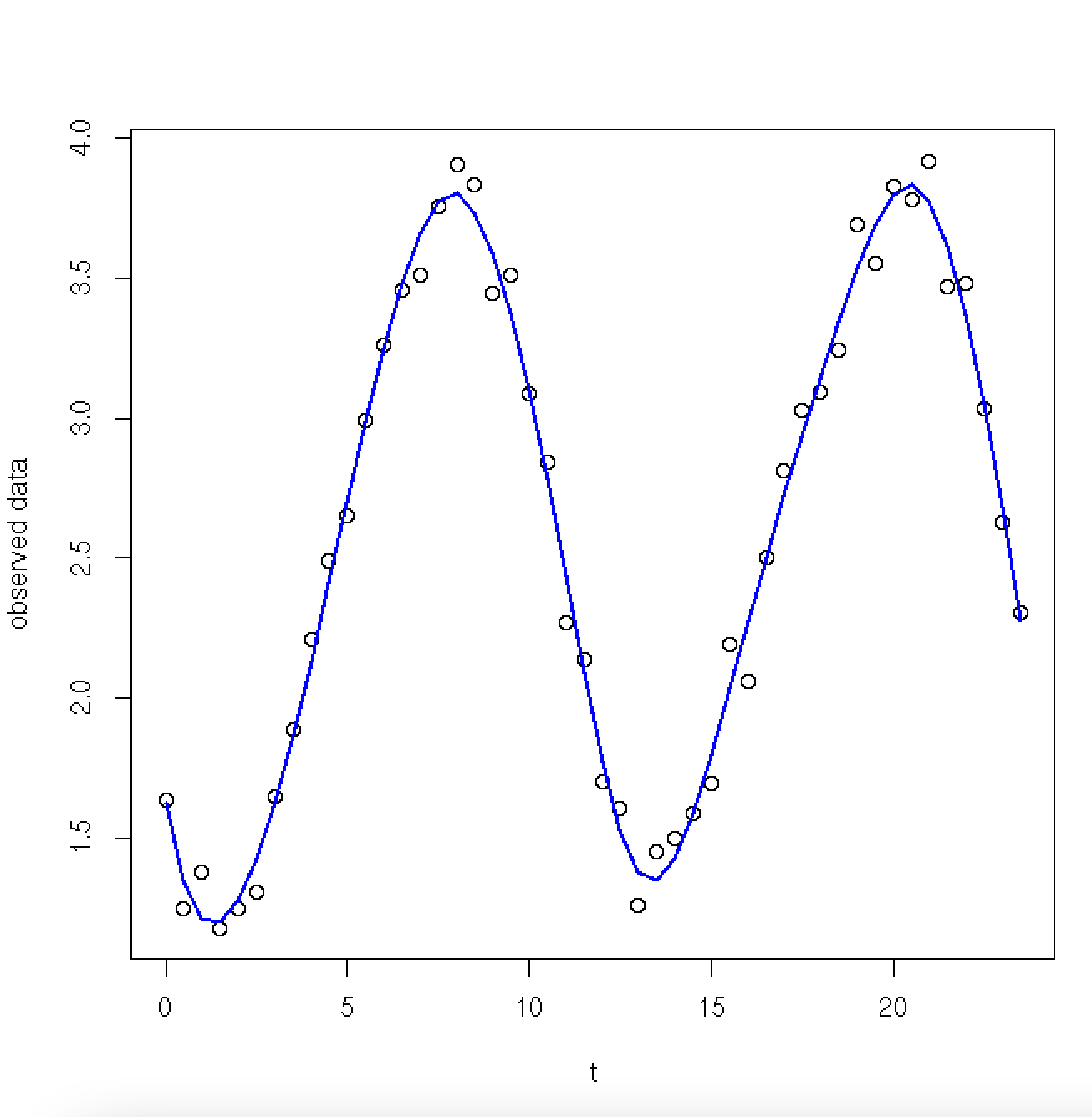
Approcciamo lo smoothing con delle basi B-spline di ordine 3 e per sceglierne in numero migliore andiamo a utilizzare la GCV in particolare possiamo vedere dal grafico



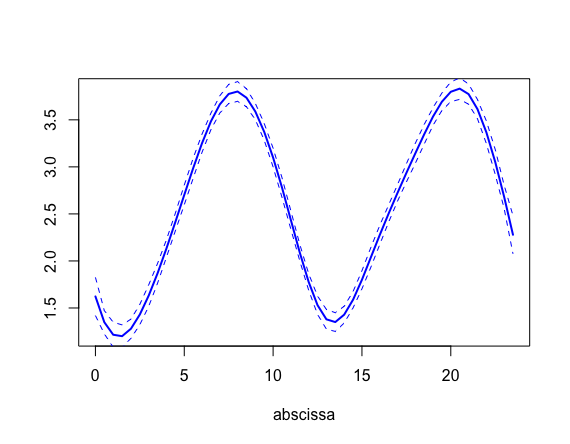
Che il valore migliore per il numero di basi è 12. A questo punto le possiamo rappresentare



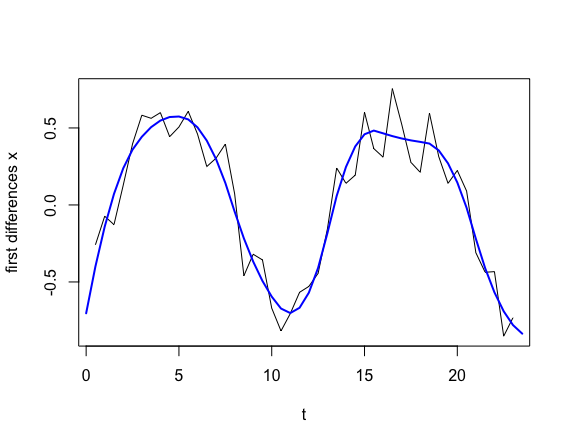
E di conseguenza possiamo rappresentare i nostri dati come funzionali attraverso questa base:



1. Ricordano che il LS smoothing può essere visto come una regressione lineare dove la matrice disegno è data dalla valutazione delle funzioni di base nei punti di osservazione. DI conseguenza trovati i valore in base alle LS possiamo stimarne la varianza e di conseguenza fornire degli intervalli di confidenza sulla funzione scelta



1. Per valutare la derivata prima dai dati utilizziamo un approssimazione che è data dalle central finite difference mentre nel caso degli smooth data ci basterà valutare la derivata della funzione. Le possiamo poi constrontare



1. Non utilizzerei un diverso numero di basi, lo smoothing sia dei dati che della derivata prima sembra essere coerente con l’andamento generale del fenomeno senza incorporare gli errori di misurazione su essi, cambiando numero di basi quindi rischieremmo di incorporare l’errore o perdere parte dell’andamento del fenomeno. D’altro canto utilizzare delle basi di tipo Fourier potrebbe rappresentare meglio il nostro fenomeno essendo periodici e oscillatorio.