Report Progetto Dati Funzionali

Matteo Ceola, Paolo Magagnato, Marco Piccolo, Pietro Stangherlin

Indice

1 Introduzione

2 Obbiettivi

- Analisi esplorative funzionali
- ANOVA funzionale: confronto tra spettri di frequenze per diverse specie di uccelli
- Modello funzione su funzione: si è interessati a valutare se esistano delle relazioni tra ciascun suono emesso ed il suono precedente

3 Dati

I dati considerati sono presenti sul portale xeno-canto. Per ogni audio è disponibile la specie di uccello e le coordinate geografiche del rilevamento.

4 Operazioni preventive

- Passaggio al dominio della frequenza tramite spettro medio
- Normalizzazione delle ampiezze

4.1 Rappresentazione funzionale

4.1.1 Spline penalizzate e vincolate

Per le curve di Ampiezza in funzione della frequenza si è scelta una rappresentazione in base bspline di grado 3. Inizialmente si sono considerate due penalizzazioni: la prima sul numero di basi e la seconda sull'integrale del quadrato della derivata seconda (per un numero di basi abbastanza alto fissato), per entrambi i casi si è considerato come riferimento il parametro che minizzasse il criterio di GCV. Tuttavia questi due criteri non permettono il rispetto dei vincoli: 1) di non negatività della curva 2) di ampiezza non superiore a 1 (a causa della normalizzazione).

Per ciascuno dei due criteri sopra menzionati si sono quindi introdotti i vincoli nel problema di ottimizzazione che può essere scritto come un programma di programmazione quadratica (A1) per cui sono disponibili delle routine.

4.1.2 Risultati

Introdurre i vincoli non dà luogo ad uno stimatore lineare in y, non è quindi possibile usare GCV come criterio per la selezione dei parametri di regolazione, si impiega invece una procedura di convalida incrociata "Leave One Out" (LOOCV). A titolo esemplificativo si riportano le curve relative ai gufi con i quattro metodi: GCV senza vincolo e LOOCV con vincolo; in @tab:representation-selection-df sono riportate le specifiche di ciascun metodo.

animal	constraint	penalty.type	min.error.parameter	domain.unique.points
falchi	FALSE	INT	9.0e+01	125
falchi	FALSE	DIFF	1.2e-06	125
falchi	TRUE	INT	4.8e+01	125
falchi	TRUE	DIFF	2.1e-06	125
gufi	FALSE	INT	9.7e+01	111
gufi	FALSE	DIFF	2.0e-07	111
gufi	TRUE	INT	2.3e+01	111
gufi	TRUE	DIFF	1.0e-07	111

animal	constraint	penalty.type	min.error.parameter	domain.unique.points
gabbiani	FALSE	INT	9.0e+01	111
gabbiani	FALSE	DIFF	4.0e-07	111
gabbiani	TRUE	INT	2.0e+01	111
gabbiani	TRUE	DIFF	6.2e-06	111

Esaminando la Figura 1 si evidenziano diverse problematiche:

- nel caso senza vincoli e con penalizzazione solo sul numero di basi (grafico in alto a sinistra) si osserva che non sono rispettati i vincoli di non negatività e di ampiezza inferiore ad uno, problema presente anche con penalizzazione sull'integrale della derivata seconda al quadrato (grafico in alto a destra).
- i vincoli migliorano chiaramente la rappresentazione funzionale, tuttavia, il numero di basi che minimizza LOOCV (grafico in basso a sinistra) è probabilmente troppo piccolo in quanto alcune funzioni hanno dei picchi troppo bassi, anche qui si potrebbe pensare di aumentare il numero di basi; nell'ultimo caso (penalizzaione sull'integrale della derivata seconda al quadrato) (grafico in basso a destra) una possibile critica è che le funzioni non siano abbastanza lisce.

Risultati simili si hanno anche con le altre due specie. Dato che i criteri di selezione automatica proposti hanno mostrato le problematiche sopra descritte si è deciso di adottare un'euristica in maniera tale che le curve mostrassero un discreto adattamento e al contempo fossero abbastanza lisce. Dopo una serie di prove si sono considerate le funzioni vincolate con penalità sulla derivata seconda riducendo il numero di basi.

species	basis_num	lambda
falchi	70	1.0e-07
gufi	70	1.0e-07
gabbiani	70	6.2e-06

5 Medie funzionali

In Figura 2 si riportano le medie e le deviazioni standard funzionali per tutti i gruppi di ciascun animale. Le deviazioni standard sono circa dello stesso ordine delle medie per quasi tutti i gruppi e tutti i punti.

Per i falchi le medie e le deviazioni standard dei gruppi hot e temperate appaiono vicine nel ragne di frequenze tra 3khz e 4.5khz, mentre fuori da questo intervallo presentano differenze sia in media che in varianza, rimandendo comunque sempre sopra le curve del gruppo cold.

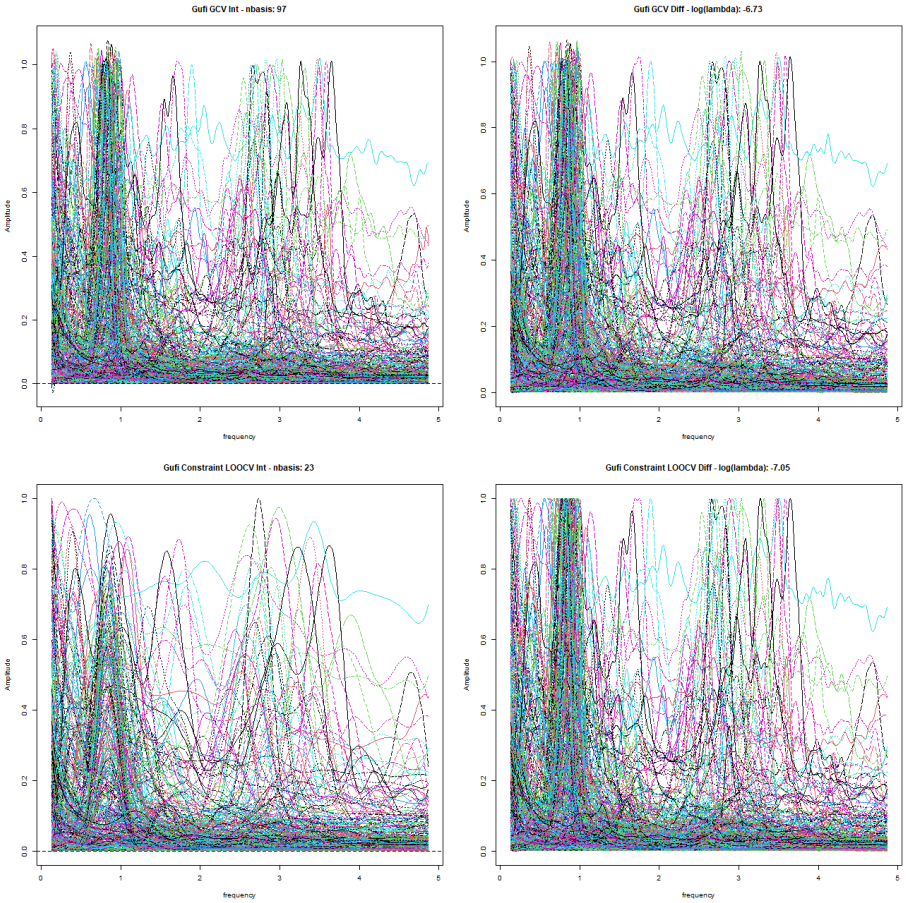


Figura 1: Rappresentazioni funzionali tramite funzioni di base degli spettrogrammi medi dei suoni dei gufi. In alto i criteri non vincolati: a sinistra con il numero di basi selezionate tramite GCV per non penalizzato, a destra con il lambda selezionato con penalità sulla derivata seconda. In basso i criteri vincolati: a sinistra con il numero di basi selezionate tramite convalida incrociata LOOCV per il criterio non penalizzato e a destra con il lambda ottimo con penalità sulla derivata seconda

Come per i falchi anche per i gufi le curve medie nei tre gruppi sembrano seguire un andamente comune, così come le deviazioni standard, qui le due curve più vicine sono invece quelle dei gruppi cold e temperate. Risalta il picco nei pressi dello zero, ciò potrebbe essere semplicemente un artefatto numerico.

Nei gabbiani la curva media che di discosta dalle altre è quella relativa al gruppo delle Canarie che domina le altre curve alle frequenze basse, mentre per quelle più alta è sotto tutte le altre curve medie. Si nota inolte un picco nella funzione media e un doppio picco per la deviazione standard intorno ai 3.5 khz per il gruppo Centre.

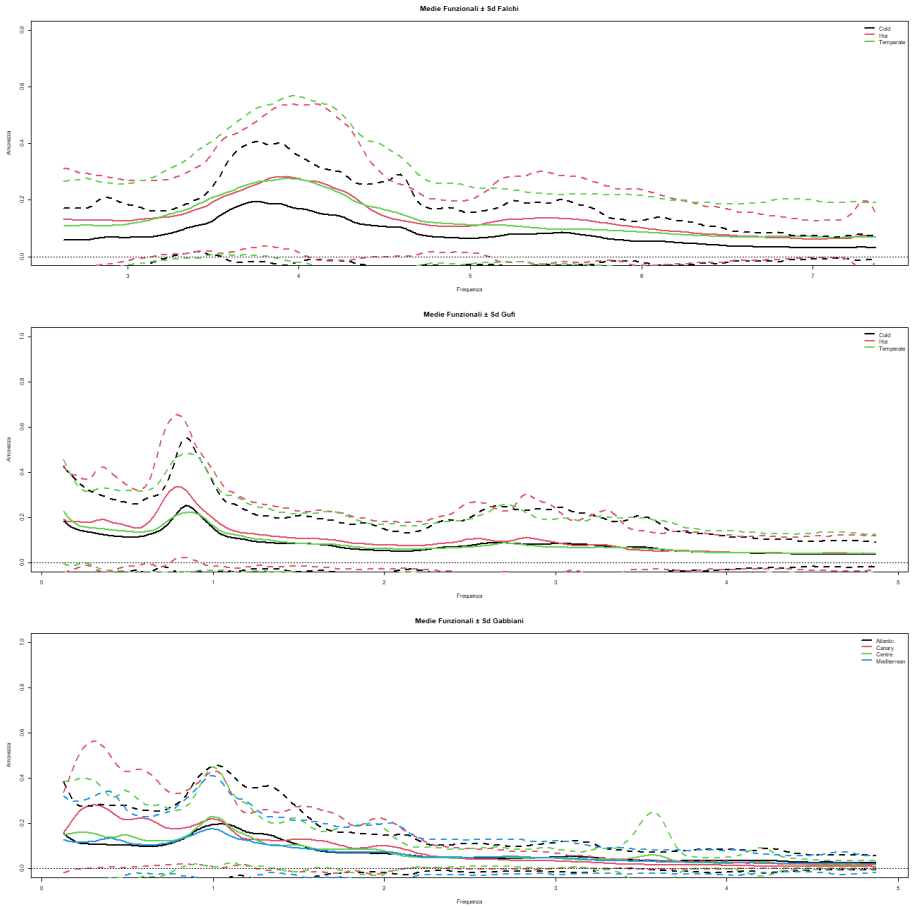


Figura 2: Medie Funzionali degli spettrogrammi medi dei suoni degli animali per i diversi gruppi a cui è aggiunto e sottrato l'errore standard funzionale. Da sinistra verso destra falchi, gufi e gabbiani

6 PCA funzionale

Si effettua un'analisi delle componenti principali funzionali, sia per avere delle ulteriori informazioni descrittive sia per vedere se i punteggi di tali componenti individuano o meno dei cluster di osservazioni potenzialmente diversi da quelli scelti in questa sede.

Inizialmente si fissa un numero di armoniche pari a 10 (Figura 3), per i falchi le prime 3 componenti spiegano circa l'80% della varianza, per i gufi sono necessarie 4 componenti per spiegare circa l'80% della varianza, mentre per i gabbiani l'85% della varianza è spiegato dalle prime tre componenti.

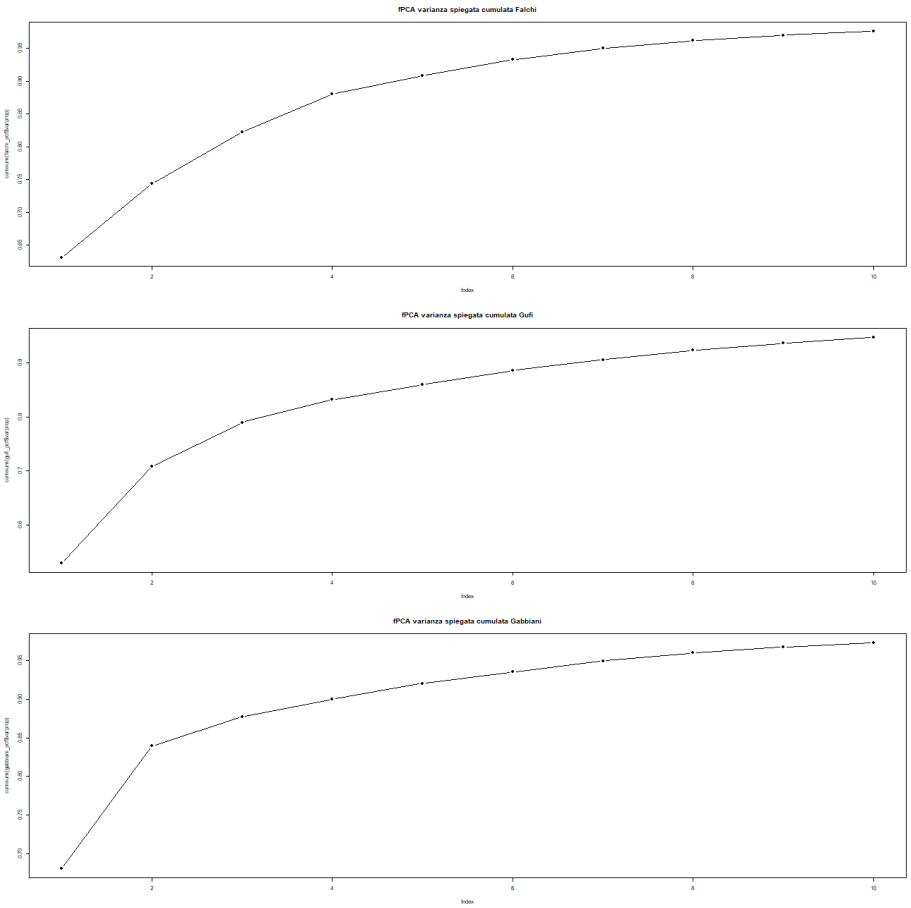
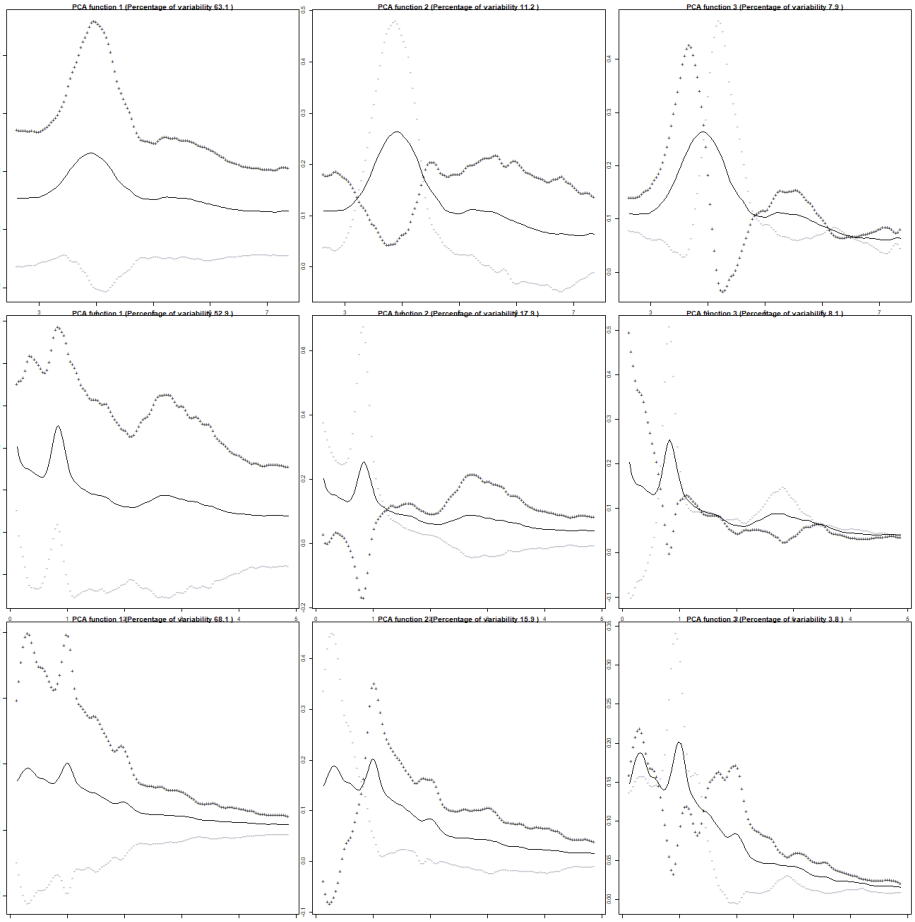


Figura 3: Varianza spiegata cumulata per le prime 10 componenti funzionali principali degli spettrogrammi dei suoni (dall'alto in basso) di falchi, gufi e gabbiani

Focalizzandosi sulle prime tre armoniche (**?@fig-f_pca_harmonics**) si può

vedere che, per tutti gli animali, la prima armonica segue abbastanza la forma della media funzionale, la seconda presenta un andamento opposto alla media e la terza varia molto tra i diversi animali e non è di facile intepretazione.



Considerando le prime due armoniche (ma un risultato analogo si ottiene anche con le prime tre) e rappresentando graficamente i punteggi (Figura 4) non si distingue nessun cluster di punti.

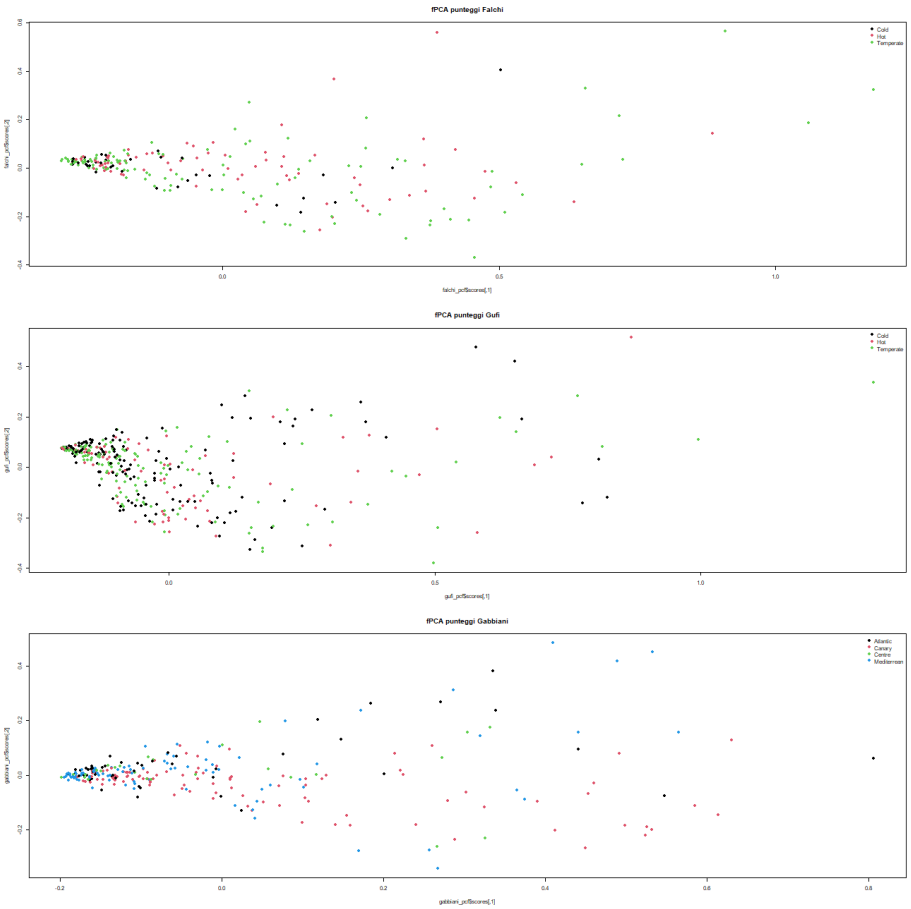


Figura 4: Punteggi delle prime due armoniche relative alle componenti principali funzionali degli spettrogrammi dei suoni (dall'alto in basso) di falchi, gufi e gabbiani

7 ANOVA funzionale

8 Modello funzione su funzione

9 Conclusioni

10 Appendice

10.1 A1

Il programma di ottimizzazione quadratica nella sua forma più generale è definito come

$$\min_b(-d^T b + 1/2 b^T D b) \text{ s.t. } A^T b \geq b_0 \tag{1}$$

Per una singola osservazione funzionale, per il lisciamento tramite scrittura in funzioni di base la funzione da minimizzare rispetto a b è

$$(y - \Phi b)^T (y - \Phi b) + \lambda b^T P b \tag{2}$$

dove y è il vettore dei punti osservati Φ è la matrice delle funzioni di base valutate nei punti osservati del dominio della curva e b è il vettore dei coefficienti, P è una generica matrice di penalità e $\lambda > 0$ indica l'entità della penalizzaione (per un criterio non penalizzato è sufficiente porre $\lambda = 0$). Minizzare Equazione 2 equivale a minizzare

$$-y^T \Phi b + b^T \frac{1}{2}(\Phi^T \Phi + \lambda P) b \tag{3}$$

Da cui $d = y^T \Phi$ e $D = \Phi^T \Phi + \lambda P$. Usando la definizione di scrittura in basi il vincolo è $0 \leq \phi^T(f) b \leq 1 \quad \forall f$, in pratica si discretizza $\phi(f_j)$ per $j = 1, .., J$. Sia Φ_J la matrice di funzioni di base valuate su griglia discretizzata: deve valere $\Phi_J b \leq \mathbb{1}$ o equivalentemente $-\Phi_J b \geq -\mathbb{1}$ dove con questa scrittura si intende che la disuguaglianza deve valere per ogni elemento dei vettori. Similmente, per vincolo di positività si ha $\Phi_J b \geq \mathbb{0}$, tuttavia, poichè per costruzione le basi bspline sono sempre non negative è sufficiente imporre $\mathbb{I} b \geq \mathbb{0}$ con $\mathbb{I}_J$ matrice identità. Combinando le due espressioni si ottiene

$$\begin{pmatrix} -\Phi \\ \mathbb{I} \end{pmatrix} b \geq \begin{pmatrix} -\mathbb{1} \\ \mathbb{0} \end{pmatrix}$$

chiaramente $A = (-\Phi^T \quad \mathbb{I})$ e $b_0 = (-\mathbb{1} \quad \mathbb{0})^T$, è dunque conclusa la scrittura del problema vincolato come programma quadratico.