UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Misura della distanza tra comunità ecologiche attraverso il principio di massima entropia e la geometria dell'informazione

Relatore

Prof. Samir Suweis

Laureando

Pietro Pecchini

Anno Accademico 2022/2023

Ringraziamenti

Ringrazio chi mi ha permesso di vivere questa esperienza con tranquillità, chi mi ha fatto sorridere e ridere di gioia, chi mi ha aiutato a capirmi e spinto ad interrogarmi; ringrazio per la fortuna di aver persone a cui voler bene.
Un ringraziamento particolare a mio nonno Dino che mi ha sempre supportato.

Abstract

In questa tesi svilupperemo un framework teorico per misurare la "distanza " tra comunità ecologiche. Per farlo si utilizzerà prima il principio di massima entropia per inferire la distribuzione di probabilità delle popolazioni delle specie nella comunità ecologica. Poi si sfrutteranno dei risultati di geometria dell'informazione per studiare le proprietà della geometria indotta dalle distribuzioni di probabilità inferite attraverso il relativo tensore metrico. Infine, definendo opportunamente la distanza in questo spazio, si potrà calcolare quanto vicino o lontane sono due comunità ecologiche. La tesi prevede una parte consistente di calcoli analitici e di simulazioni per testare i risultati.

Indice

Introduzione

La fisica dei sistemi complessi è un ambito della fisica che vuole investigare la dinamica e le proprietà emergenti di sistemi composti da molti gradi di libertà e interagenti tra loro, come ad esempio sistemi interconnessi quali i sistemi economici, quelli sociali o biologici.
In questa tesi il focus centrale è quello di studiare i sistemi ecologici, composti da specie interagenti tra loro. Lo faremo attraverso approcci di fisica statistica e di geometria differenziale, attraverso cui saremo in grado di fornire interpretazioni e previsioni statistiche sulla composizione e la popolazione delle specie all'interno di una data comunità ecologica e sulle interazioni specie-specie o specie-ambiente [1]. In particolare in questa tesi ci limiteremo ad investigare le proprietà stazionarie di tali comunità, mentre lo studio della loro dinamica viene rimandato ad un futuro lavoro.

Le teorie ecologiche principali che descrivono la nascita e la struttura delle comunità sono due [2] [3].
La teoria di "costruzione per nicchia" si basa sul principio che le comunità ecologiche siano composte da un numero limitate di nicchie, ognuna delle quali è occupata da una specie; immaginando che i fattori necessari per la sopravvivenza di una specie siano n (spazio, disponibilita di nutrienti, presenza di prede o mancanza di predatori,...) si può pensare che ogni specie occupi un volumetto n-dimensionale le cui dimensioni variano in base al numero di fattori necessari per la sopravvivenza della specie [2]. Il numero di specie che possono instaurarsi in un dato luogo, e quindi la biodiversità, dipende quindi da quanti volumetti n-dimesionali il luogo può ospitare prima di saturarsi, ovvero terminare, ad esempio, lo spazio fisico disponibile o la disponibilità di nutrienti presenti nel terreno.
Nel caso in cui i volumetti di due specie diverse si sovrappongano si instaura una situazione di competizione tra di esse, che può portare all'estinzione di una delle specie considerate o al riadattamento di una o entrambe le specie per permettere la coesistenza.
Questo modello si basa quindi su un'analisi di ogni singola specie nel dettaglio, andando a vedere come questa, insieme ad altre specie possono distribuirsi le risorse di un dato luogo e deducendo così la biodiversità possibile all'interno di una comunità ecologica [2].
Benchè questo modello sia ampiamente utilizzato ha come limite quello di essere eccessivamente meccanicistico in un campo (l'ecologia) in cui l'enormità di variabili da tenere in considerazione e l'impossibilità di tenerne conto di tutte, rende necessario fare un ragionamento di tipo statistico [3].
Il secondo modello viene definito "costruzione per dispersione" e si basa su una visione meno dettagliata dell'interazione di ogni singola specie con il luogo e la comunità, ma guarda invece a scale spaziali e temporali molto più grandi [3]. Inoltre tale approccio non intende modellizzare la dinamica per lo sfruttamento delle risorse, ma assume che le proprietà dell'ecosistema a queste scale dipendano da processi con una forte componente stocastica quale il "*drift ecologico*", ovvero le fluttuazioni casuali degli eventi di nascita e di morte delle specie all'interno di una comunità ecologica. Altri processi determinanti la costruzione per dispersione sono gli eventi di migrazioni, e di *"estinzione locale"*, ovvero la scomparsa locale di una specie da una data comunità [2].

1.1 La diversità delle specie

Quando si parla di diversità tra comunità ecologiche, ed in particolare per comunità microbiche, si guardano tre tipi di diversità [4]: la *diversità* α che quantifica il numero medio di specie a livello locale, ovvero il numero di specie presenti in ogni comunità, mediato sul numero di comunità; la *diversità* β che quantifica l'intersezione della biodiversità tra due comunità, ed infine la *diversità* γ che è determinata dal numero totale di specie distinte considerando tutte le comunità.

Nonostante l'esistenza di questi indicatori, fare un confronto sulla diversità tra due comunità ecologiche rimane un problema aperto. Infatti, ogni comunità può essere affetta da errori di campionamento e questo può generare dei bias nelle stime di diversità α, β e γ [5].
Un esempio di possibile metrica per misurare la distanza tra le diversità di due specie è dato dalla *DOA, dissimilarity-overlap analysis* [4], che pone l'accento su due aspetti principali: la misura di dissimilarità misura la differenza delle abbondanze relative delle specie all'interno della comunità, mentre l'overlap misura la probabilità che, dato un individuo appartente ad una delle due comunità, questo appartenga anche all'altra.

1.2 Framework teorico

Due tra i principali modelli matematici utilizzati per descrivere l'evoluzione delle popolazioni in una comunità ecologica sono il processo stocastico di nascita/morte" [1] ed il metodo basato sul "principio di massima entropia" [6].

Il primo si basa sull'ipotesi che il rate di nascita e di morte pro-capite per la specie i-esima determini (statisticamente) la popolazione della specie secondo la cosìdetta *Birth-Death Master Equation* [7]

$$\frac{\partial P}{\partial t}(n,t) = b(n-1)\cdot P(n-1,t) - [b(n)+d(n)]\cdot P(n,t) + d(n+1)\cdot P(n+1,t)\ , \tag{1.1}$$

dove $P(n,t)$ è la probabilità che la specie abbia una popolazione di n individui al tempo t, e dove $b(n)$ e $d(n)$ sono rispettivamente i rates di nascita e morte. I parametri $b(n)$ e $d(n)$ possono in molti casi essere ricavati sperimentalmente da dati demografici. Lo studio di questa equazione permette quindi di studiare l'andamento di comunità ecologiche e le loro fluttauzioni.

Il secondo metodo si basa sul principio di massima entropia, definito anche in modo più compatto come "MaxEnt", secondo cui a partire da un set di dati, su cui si hanno alcune informazioni, come ad esempio si conosce la media o la matrice di covarianza, si può stimare la distribuzione di probabilità che meglio descive i dati, e che massimizza l'entropia, ovvero l'incertezza, della distribuzione.

Il concetto di massimizzare l'entropia corrisponde ad una richiesta di costruire un modello che usi solamente alcuni vincoli dati dal sistema e sia per il resto il meno biased possibile. Ad esempio, un modello in cui non si impone nessuna condizione risulta tramite il MaxEnt in una distribuzione uniforme, in cui quindi nessun risultato è più probabile di un'altro, coerentemente con il fatto di non aver fornito alcuna informazione sul sistema. Questo modello permette di fare uno studio di tipo statistico sull'abbondanza delle varie specie presenti in una comunità ecologica, e tramite l'analisi di esso dedurre informazioni riguardanti ad esempio l'interazione tra due specie [8]. Questo approccio sarà utilizzato nel proseguo di questa tesi e verrà approfondito in seguito.

L'idea di questa tesi è quella di testare un nuovo tipo di metrica [9] [10] applicabile nello studio della diversità di comunità ecologiche, basato sulla distanza delle probabilità $P(\vec{n})$ dell'abbondanza delle popolazioni delle specie date dal principio di massima entropia, che permette di ricavare da un set di dati e delle ipotesi sperimentali, la distribuzione che meglio descrive i dati. In particolare $P(\vec{n})$ apparterrà ad una famiglia di funzioni la cui forma è fissata dai vincoli ricavati sperimentalmente; in questo caso si tratterà di distribuzioni gaussiane multivariate generalizzate (MGGDs), ovvero di gaussiane multidimensionali in cui non si impongono vincoli sulla matrice di covarianza.

Studiando queste distribuzioni su una varietà differenziale naturale per lo studio delle distribuzioni di probabilità, e con una metrica naturale indotta, sarà possibile calcolare la distanza tra due distribuzioni rappresentative di due comunità ecologiche diverse, ovvero quantificarne la differenza [9] [10].

Lo studio riportato di seguito è stato eseguito su comunità sintetiche, ovvero comunità generate attraverso delle simulazioni numeriche e per cui possiamo quindi controllare la distribuzione stazionaria delle popolazioni delle specie.

Il Principio di Massima Entropia

2.1 L'entropia dell'informazione

Il concetto di entropia dell'informazione riguarda non tanto l'informazione che abbiamo su sistema, quanto invece l'informazione che è possibile ricavare dal sistema oltre a quella che già si ha; di conseguenza un sistema di cui non si conosce niente avrà una alta entropia, in quanto tutto è ancora da scoprire. Quando si parla di entropia il termine informazione assume il significato di "riduzione della ambiguità", inteso come un aumento delle capacità di discriminare o caratterizzare un certo sistema. Ora illustriamo due esempi che ci aiutano a capire come siano gli eventi rari o poco probabili a fornirci le maggiori informazioni su un sistema.

- Il DNA umano è composto da una sequenza ordinata di nucleotidi comprendenti 4 diverse basi azotate: T A G C. Se si identificasse ognuna di esse tramite un numero in base binaria, ad esempio T=00, A=01, G=10, e C=11, considerando che il numero di nucleotidi nel nostro DNA è di circa 6 miliardi sarebbe possibile scrivere l'intero DNA umano in 1,5 GB. Tuttavia il 99,9% di questo è identico per ogni umano, di conseguenza confrontare sezioni di DNA che rientrano in questa frazione non porta informazione al fine di voler distinguere due persone, ovvero non aiuta a ridurre l'ambiguità. Diversamente, se individuassimo sequenze di nucleotidi diverse tra due individui, esse porterebbero con se una grande quantità di informazione. In questo caso quindi analizzare frazioni di DNA appartenente al 99,9% porta una quantità di informazione quasi nulla, diversamente analizzando sequenze appartenenti allo 0,01% si può ridurre enormemente la ambiguità.
- Un altro esempio risulta essere il gioco dell'impiccato: in questo gioco infatti possiamo scegliere delle lettere da cui partire per indovinare la parola. Se le lettere di partenza sono ad esempio ' _ a _ t e _ a ' si hanno molte possibilità per completare la parola; diversamente se avessimo avuto come lettere iniziali ' z _ t t _ r _ ' , probabilmente avremmo indovinato immediatamente la parola "zattera". Ciò è dovuto al fatto che le consonanti utilizzate, oltre che la loro disposizione, rappresentano un evento estremamente più improbabile che non le vocali proposte nel primo caso, e quindi riducono la ambiguità su come si può concludere la parola.

Si nota quindi come siano proprio gli eventi più rari a portare una maggiore informazione, e di conseguenza a ridurre l'entropia del sistema.
D'altronde ciò risulta essere coerente anche con l'esempio a cui siamo più familiari riguardante le particelle di un gas libere di muoversi tra due contenitori identici: l'evento più probabile, ovvero quelle delle particelle equidistribuite, porta poca informazione ed ha una entropia alta, contrariamente al caso in cui tutte le particelle sono contenute in un unico contenitore, evento molto improbabile e con entropia bassa.

Mostriamo ora, come da considerazioni qualitative possiamo ricavare che la forma funzionale dell'entropia deve essere un logaritmo. Si ipotizzi ora di giocare all'impiccato, facendo tuttavia l'ipotesi semplicistica che ogni lettera sia indipendente dalle altre, si definisca I la variabile che rappresenta la quantità di informazione e si ipotizzi di indovinare una lettera; dal momento che la quantità di informazione portata da un evento è legata alla probabilità dell'evento stesso, possiamo affermare che l'informazione è ora $I(p_1)$, dove p_1 rappresenta la probabilità di uscita di quella lettera specifica. Se si ipotizza ora che venga trovata una seconda lettera si può affermare che all'informazione posseduta in precedenza si debba sommare quella ottenuta tramite la seconda lettera, ma d'altra parte poiche gli eventi sono indipendenti secondo la ipotesi iniziale, si ottiene che la informazione totale è anche funzione della probabilità congiunta di trovare queste due lettere, quindi in termini matematici:

$$I(p_1) + I(p_2) = I(p_1 p_2) \ . \tag{2.1}$$

Questa forma suggerisce che la forma di $I(p)$ sia del tipo $I(p) = Kln(p)$, dove K è una costante moltiplicativa che ci si aspetta essere negativa in accordo con il fatto che eventi poco probabili portano molta informazione mentre eventi probabili ne aggiungono poca; per semplicità si pone $K = -1$.

2.2 Il principio di massima entropia

Il principio di MaxEnt si basa sull'idea di creare una distribuzione di probabilità a partire da un set di dati massimizzando l'incertezza, ovvero l'entropia dell'informazione, avendo tuttavia tenuto in considerazione alcune informazioni che già possediamo sul sistema che verranno espresse sottoforma di vincoli.
Abbiamo visto come, data la probabilità di un evento, l'informazione che questo porta risulta essere $I(p_{ev}) = -ln(p_{ev})$, si può quindi affermare che l'entropia di una distribuzione di probabilità sia data dal valor medio di tale grandezza.

$$I_p = -\sum_{i=1}^{n^o eventi} p_i ln(p_i) = -\int_D p(\mathbf{x}) ln(p(\mathbf{x}))\, d\mathbf{x} \ . \tag{2.2}$$

Per il caso rispettivamente discreto e continuo, dove $D \subseteq \mathbb{R}^N$ rappresenta il dominio della variabile.

I vincoli accennati sopra sono nel caso da noi trattato:

- vincolo di normalizzazione: dal momento che si trattano distribuzioni di probabilità si richiede che la distribuzione che si ricava da MaxEnt sia normalizzata, ovvero $\int_D p(\mathbf{x})\, d\mathbf{x} = 1$
- vincolo di media: si fissa la media della distribuzione uguale a quella ricavata dai dati, ovvero $\int_D \mathbf{x} p(\mathbf{x})\, d\mathbf{x} =< \mathbf{x} >$
- vincolo del valore di aspettazione di $x_i x_j$: si fissa il valore di aspettazione del prodotto tra x_i e x_j uguale a quello ricavato dai dati, ovvero $\int_D x_i x_j p(\mathbf{x})\, d\mathbf{x} =< x_i x_j >$

Questi vincoli vengono introdotti nell'equazione da massimizzare sottoforma di moltiplicatori di Lagrange $(\nu, \mu_i, \Omega_{ij})$, di conseguenza l'equazione finale da massimizzare risulta essere [8]

$$\begin{aligned} -\int_D p(\mathbf{x}) ln(p(\mathbf{x}))\, d\mathbf{x} - \nu\left(\int_D p(\mathbf{x})\, d\mathbf{x} - 1\right) - \sum_{i=1}^{dim.spazio} \mu_i \left(\int_D x_i p(\mathbf{x})\, d\mathbf{x} - < x_i >\right) \\ -\frac{1}{2}\sum_{i,j}^{dim.spazio} \Omega_{ij}\left(\int_D x_i x_j p(\mathbf{x})\, d\mathbf{x} - < x_i x_j >\right) = \alpha(p(\mathbf{x})) \ . \end{aligned} \tag{2.3}$$

Per massimizzarla e trovare la distribuzione di probabilità che massimizza l'entropia è necessario derivare per p(**x**), e porre a zero la derivata. In questo modo risulta che la funzione che massimizza l'entropia è del tipo

$$p(\mathbf{x}) = e^{-1-\nu-\mu\mathbf{x}-\frac{1}{2}\mathbf{x}\Omega\mathbf{x}} = \frac{1}{Z} e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y}} \ , \tag{2.4}$$

dove $y_i = x_i + \Omega_{ij}^{-1}\mu_j$, $Z = e^{\frac{1}{2}\mu_i\Omega_{ij}^{-1}\mu_j + 1 + \nu}$ costante di normalizzazione e μ vettore. Il secondo vincolo si presenta nella forma

$$< x_i >= \int_D x_i p(\mathbf{x})\, d\mathbf{x} = \frac{1}{Z}\int_{D'} e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y}}\left(y_i - \Omega_{ij}^{-1}\mu_j\right) d\mathbf{y} = -\Omega_{ij}^{-1}\mu_j \ . \tag{2.5}$$

Allo stesso modo si ha che la terza condizione diventa

$$< x_i x_j >= \int_D x_i x_j p(\mathbf{x})\, d\mathbf{x} = \frac{1}{Z} \int_{D'} e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y}} [< x_i >< x_j > + y_i < x_j > + y_j < x_i > + y_i y_j]\, d\mathbf{y} = \\ < x_j >< x_i > + \frac{1}{Z} \int_{D'} e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y}} y_i y_j\, d\mathbf{y} \tag{2.6}$$

e portando il primo addendo del membro di destra a sinistra si riconosce la componente ij della matrice di covarianza

$$\varepsilon_{ij} =< x_i x_j > - < x_j >< x_i > . \tag{2.7}$$

Introducendo l'integrale

$$\Gamma(\mathbf{J}) = \int_{D'} e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y} + \mathbf{J}\ \mathbf{y}}\, d\mathbf{y} \tag{2.8}$$

è possibile esprimere il secondo termine del membro di destra di eq. (2.6) a meno della costante moltiplicativa come

$$\left.\frac{\partial^2 \Gamma(\mathbf{J})}{\partial J_i \partial J_j}\right|_{\mathbf{J}=\mathbf{0}} = \int_{D'} e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y}} y_i y_j\, d\mathbf{y} \ . \tag{2.9}$$

Facendo ora la sostituzione:

$$\mathbf{z} = \mathbf{y} - \Omega^{-1}\mathbf{J} \tag{2.10}$$

si trova che

$$\Gamma(\mathbf{J}) = e^{\frac{1}{2}\mathbf{J}\Omega^{-1}\mathbf{J}} \int e^{-\frac{1}{2}\mathbf{z}\Omega\mathbf{z}}\, d\mathbf{z} = e^{\frac{1}{2}\mathbf{J}\Omega^{-1}\mathbf{J}} \left[\Gamma(\mathbf{J})|_{\mathbf{J}=\mathbf{0}}\right] \tag{2.11}$$

ottenendo così [8]

$$\frac{1}{Z} \int e^{-\frac{1}{2}\mathbf{y}\Omega\mathbf{y}} y_i y_j\, d\mathbf{y} = \frac{1}{\Gamma(\mathbf{J})} \left.\frac{\partial^2 \Gamma(\mathbf{J})}{\partial J_i \partial J_j}\right|_{\mathbf{J}=\mathbf{0}} = \Omega^{-1}_{ij} \ . \tag{2.12}$$

Confrontandola con l'eq. (2.6) appare chiaro che $\Omega^{-1} = \varepsilon$ è proprio la matrice di covarianza.

Riepiloghiamo quello che abbiamo fatto: partendo da un set di dati abbiamo imposto tre diverse condizioni:

- la condizione di normalizzazione, ovvero che la distribuzione integrata su tutto lo spazio dia esattamente 1, questa condizione è regolata dal parametro libero ν che, in accordo con il principio dei moltiplicatori dovrà essere scelto tale da rendere vera questa condizione

- la condizione che la media della distribuzione sia esattamente uguale alla media del set di dati da cui la deriviamo, questa condizione è regolata dal moltiplicatore di lagrange vettore μ, che, se si ragiona su uno spazio N-dimensionale, avrà anch'esso dimensione N

- la condizione che il valore atteso del prodotto tra due componenti dello spazio calcolato tramite la densità di probabilità ottenuta sia uguale a quello ricavato dai dati, e questa condizione è regolata dagli $\frac{n(n+1)}{2}$ parametri liberi di Ω.

Ci si potrebbe chiedere come mai, dal momento che sembrano essere "a costo zero", non si prosegua a porre vincoli per rendere la funzione ottenuta tramite MaxEnt sempre più fedele ai dati da cui viene calcolata, tuttavia in realtà ciò pone un problema fondamentale: al momento ci siamo limitati al "secondo ordine" nel calcolo del distribuzione, ovvero abbiamo calcolato la condizione di normalizzazione (ordine 0 che impone un parametro adimensionale), la media delle singole variabili (ordine 1 che pone tanti parametri quanto è la dimensione dello spazio n), la media del prodotto tra due variabili

(ordine due che pone un numero di parametri che scala con il quadrato della dimensione dello spazio n); se si volesse andare oltre e si volesse calcolare il valore atteso del prodotto tra tre parametri si dovrebbe generare un tensore di rango 3 completamente simmetrico, la cui dimensione sarebbe quindi proporzionale a n^3. La qualità, l'affidabilità dei parametri che si deducono dai dati dipende fortemente dalla quantità di dati da cui questi si derivano, di conseguenza se si vuole salire di ordine si dovrebbe aumentare esponenzialmente la quantità di dati nel dataset per mantenere l'affidabilità delle stime dei parametri. Un altro motivo per il quale si è deciso di limitarsi al secondo ordine, è relativo all'uso che ne faremo di queste distribuzioni: esse verranno studiate su una varietà differenziale e si calcolerà la distanza, la "diversità", tra due di queste tramite una metrica "ad hoc"; questo tipo di metrica [9] [10] è estremamente difficile da sviluppare matematicamente ed al momento esiste solo per gaussiane multivariate (che è il tipo di funzione che otterremo, in accordo con eq. (2.4)), se quindi aumentassimo l'ordine non sarebbe più possibile calcolare analiticamente la distanza tra due distribuzioni ed il lavoro fatto finora non potrebbe proseguire.

Dopo aver imposto i vincoli sottoforma di moltiplicatori di Lagrange, si fa la richiesta che la entropia della distribuzione sia massimizzata, ovvero che la distribuzione si distribuisca il più uniformemente possibile, per quanto permesso dai vincoli, e così facendo si ottiene una soluzione della forma eq. (2.4), che tramite un cambio di variabili diviene una Gaussiana multivariata generalizzata centrata nell'origine. Attraverso manipolazioni matematiche è possibile calcolare che valori debbano assumere i parametri di Lagrange affinchè i vincoli siano rispettati, e così facendo si trova che $\Omega^{-1} = \varepsilon$, matrice di covarianza e che il vettore parametro μ è legato alla media delle singole variabili secondo una rotazione ed un dilatamento o accorciamento indotto da Ω^{-1}, mentre si può evitare di calcolare ν considerando Z il nuovo parametro di normalizzazione che, ottenuto da un esponenziale, può assumere qualunque valore positivo.

Pensando ora che ciò viene applicato ai dati reali di comunità, vediamo di dare significato ai parametri finora presentati, in cui la costante di normalizzazione ed il vettore μ hanno interpretazione immediata. Il parametro μ infatti è correlato alla popolosità media di ogni specie all'interno di una comunità, mentre il fattore di normalizzazione è una pura richiesta matematica-statistica.

La matrice Ω invece permette di avere informazioni su come le specie interagiscono tra di loro: in letteratura si introduce la grandezza $M = -\varepsilon^{-1} = -\Omega$, matrice di interazione, il cui elemento M_{ij} quantifica la forza dell'interazione tra la i-esima e la j-esima specie. In particolare $M_{ij} < 0$ indica che le specie tendono a non stare nello stesso stesso posto, il che può essere pensato come se ci sia "competizione" tra di esse, magari le due specie hanno bisogno delle stesse sostanze nutritive e quindi non coesistono pacificamente; diversamente $M_{ij} > 0$ indica che le due specie sono in un rapporto di cosidetto "mutualismo" ovvero si "aiutano", un esempio di ciò può essere rappresentato da una specie di microbi che si nutre di scarti generati da un'altra specie; infine $|M_{ij}|$ quantifica la forza della loro interazione [8].
E' importante tuttavia sottolineare che i ragionamenti fatti non danno alcuna motivazione biologica del perchè queste cose accadano, ma si limitano a descrivere ciò che accade; inoltre spesso in questo tipo di analisi si tiene conto solo delle specie più popolose, trascurando quelle più "rare", di conseguenza il tipo di interazione che si è registrato tra due specie potrebbe essere mediato da una terza specie che non è stata presa in considerazione durante l'analisi.

Geometria dell'Informazione

3.1 Introduzione alla geometria dell'informazione

La "geometria dell'informazione", traduzione in italiano di "Information geometry", è uno strumento estremamente utilizzato nell'ambito della statistica inferenziale che si basa sullo studio delle distribuzioni di probabilità e dell'informazione in esse contenuta tramite gli strumenti della geometria differenziale.
Nel seguito verrà presentata una base della teoria dietro questa branca della matematica e ci si concentrerà poi sulla sua applicazione alle *MGGDs* in quanto fulcro degli argomenti qui trattati.

La condizione di normalizzazione delle distribuzioni di probabilità, che per semplicità di intuizione le consideriamo discrete per ora, si scrive in questo modo

$$\sum_{i=0}^{N} p_i = \sum_{i=1}^{N} \zeta_i^2 = 1 \ , \qquad \text{dove } \zeta_i = \sqrt{p_i}, \tag{3.1}$$

il che suggerisce che ζ possa essere considerata come un vettore la cui punta giace su una sfera N-dimensionale di raggio unitario.

La generalizzazione per distribuzioni di probabilità continue risulta quindi essere

$$\int_D \zeta(\mathbf{x})^2 \, d\mathbf{x} = 1 \tag{3.2}$$

in cui D è il dominio su cui vive la variabile $\mathbf{x}$ e $\zeta(\mathbf{x})$ è raffigurabile come vettore su una sfera ∞-dimensionale immersa nello spazio di Hilbert $L^2(D)$.
Le distribuzioni di probabilità dipendono generalmente da dei parametri θ, quindi $p = p(\mathbf{x}|\theta)$, in cui si intende θ come un insieme di parametri e non uno solo.
All'interno di $L^2(D)$ la forma più comune del prodotto scalare tra due elementi dello spazio è

$$< \eta, \zeta >= \int_D \eta(\mathbf{x})\zeta(\mathbf{x}) \, d\mathbf{x} \ . \tag{3.3}$$

Date due distibuzioni η e ζ appartenenti alla stessa famiglia, ipotizzando che i parametri che le caratterizzano sono inifinitesimamente diversi gli uni dagli altri, possiamo scrivere che $\zeta = \zeta_\theta(\mathbf{x})$ ed in prima approssimazione $\eta = \zeta_\theta + \partial_{\theta_i}\zeta(\mathbf{x})d\theta^i$ in cui si è usata la convenzione di Einstein e la forma compatta $\partial_{\theta_i} = \frac{\partial}{\partial\theta_i}$;
Quindi il modulo quadro della distanza tra esse risulta essere

$$ds^2 =< \eta - \zeta, \eta - \zeta >= \left(\int_D \partial_{\theta_i}\zeta(\mathbf{x})\partial_{\theta_j}\zeta(\mathbf{x}) \right) d\theta^i d\theta^j \tag{3.4}$$

ottenendo così la matrice della metrica $g_{ij} = \int_D \partial_{\theta_i}\zeta_\theta(\mathbf{x})\partial_{\theta_j}\zeta_\theta(\mathbf{x})$.
Tale equazione si può riscrivere introducendo $l_\theta(\mathbf{x}) = \ln p_\theta(\mathbf{x}) = \ln \zeta_\theta^2(\mathbf{x})$ nel seguente modo

$$g_{ij}(\theta) = \frac{1}{4} \int_D p_\theta(\mathbf{x})\partial_{\theta_i} l_\theta(\mathbf{x})\partial_{\theta_j} l_\theta(\mathbf{x}) \, d\mathbf{x} \ . \tag{3.5}$$

Si può scegliere di trascurare la frazione moltiplicativa, essa infatti determina un accorciamento di

tutte le lunghezze, e dato che l'utilità di ricavare delle distanze è quella di poterle confrontare tra loro (un unico valore di distanza non ha significato dato che non avremmo un metro di paragone) questa scelta è legittima.
Si ottiene così la matrice dell'informazione di Fisher, che è alla base dello sviluppo matematico che si utilizzerà per determinare la distanza tra due distribuzioni [11]

$$g_{ij}(\theta) = \int_D p_\theta(\mathbf{x}) \partial_{\theta_i} l_\theta(\mathbf{x}) \partial_{\theta_j} l_\theta(\mathbf{x})\, d\mathbf{x} \ . \tag{3.6}$$

Per trovare il percorso più corto che collega due distribuzioni di probabilità data una metrica si applica il principio variazionale, facendo la richiesta che $\delta \int_a^b ds = 0$, dove nel nostro caso a e b rappresentano le due distribuzioni di partenza e di arrivo all'interno della sfera contenuta in $L^2(D)$ caratterizzate dai loro parametri θ.

Ricordando che la distanza infinitesima data una metrica si calcola come $ds^2 = g_{ij}(\theta) d\theta^i d\theta^j$ possiamo scrivere che

$$\delta ds^2 = 2ds\delta(ds) = d\theta^i d\theta^k \frac{\partial g_{ij}}{\partial \theta^l} \delta\theta^l + 2 g_{qk} d\theta^q d(\delta\theta^k) \ ; \tag{3.7}$$

inserendo $\delta(ds)$ all'interno del principio variazionale si ottiene

$$\int_a^b \left(\frac{1}{2} \frac{d\theta^i}{ds} \frac{d\theta^k}{ds} \frac{\partial g_{ij}}{\partial \theta^l} \delta\theta^l + g_{ql} \frac{d\theta^q}{ds} \frac{d(\delta\theta^l)}{ds} \right) ds = 0 \tag{3.8}$$

ed integrando per parti il secondo termine si ottiene

$$\int_a^b \left[\frac{1}{2} \frac{d\theta^i}{ds} \frac{d\theta^k}{ds} \frac{\partial g_{ij}}{\partial \theta^l} - \frac{d}{ds} \left(g_{ql} \frac{d\theta^q}{ds} \right) \right] \delta\theta^l ds = 0 \ . \tag{3.9}$$

Dato che la condizione deve essere valida per ogni $\delta\theta^l$ si deve imporre che il termine tra le parentesi sia sempre 0.
Inoltre espandendo il secondo membro e considerando solo il termine $\frac{dg_{ql}}{ds}\frac{d\theta^q}{ds}$ può essere riscritto nel modo

$$\frac{dg_{ql}}{ds}\frac{d\theta^q}{ds} = \frac{1}{2}\left(\frac{dg_{lq}}{ds}\frac{d\theta^q}{ds} + \frac{dg_{lk}}{ds}\frac{d\theta^k}{ds} \right) = \frac{1}{2}\left(\frac{dg_{lq}}{d\theta^k} + \frac{dg_{lk}}{d\theta^q} \right) \frac{d\theta^k}{ds}\frac{d\theta^q}{ds} \tag{3.10}$$

Introduciamo quindi i simboli di Christoffel

$$\Gamma^i_{kq} := \frac{1}{2} g^{il} \left(\frac{\partial g_{lq}}{\partial \theta^k} + \frac{\partial g_{lk}}{\partial \theta^q} - \frac{\partial g_{kq}}{\partial \theta^l} \right) \ . \tag{3.11}$$

Si può così finalmente scrivere l'equazione che deve essere rispettata affinche la linea che collega i due punti, le due distribuzioni, sia una geodetica

$$\frac{d^2\theta^i}{ds^2} + \Gamma^i_{kq} \frac{d\theta^k}{ds}\frac{d\theta^q}{ds} = 0 \ . \tag{3.12}$$

3.2 Applicazione alle *MGGDs*

Nel caso affrontato in questa tesi le distribuzioni seguono la forma di *MGGDs*, la cui forma generica è del tipo:

$$f(\mathbf{x}|\mu, \varepsilon) = \frac{1}{(2\pi)^{\frac{s}{2}}} (det\varepsilon)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \varepsilon^{-1} (\mathbf{x}-\mu)} \tag{3.13}$$

dove s indica la dimensione dello spazio della MGGD, che nel nostro caso coincide con il numero di *s*pecie che terremo in considerazione, μ indica il vettore media della gaussiana ed ε ne indica la matrice di covarianza.
Quindi il numero di parametri liberi a questo punto è $n = s + \frac{s(s+1)}{2}$.

Poichè tuttavia al momento non esistono metodi analitici per calcolare la distanza tra due MGGDs con tutti questi gradi di libertà, ci si è concentrati su due tipi principali di queste:

- distribuzioni con matrice di covarianza diagonale
- distribuzioni con vettore medie μ identico

Presentiamo le due casistiche una alla volta.

3.2.1 Specie non interagenti: matrice di covarianza diagonale

Il primo caso è il più semplice, infatti una matrice di covarianza diagonale implica che eq. (3.13) si possa riscrivere come un prodotto di gaussiane indipendenti, si nota quindi che il numero di parametri è $2s$, due per gaussiana.

Per comodità di calcolo, useremo come parametri della matrice di covarianza diagonale non più le varianze quanto invece le deviazioni standard $\sigma_i = \sqrt{\varepsilon_{ii}}$

$$f(\mathbf{x}|\mu,\sigma) = \frac{1}{(2\pi)^{\frac{s}{2}}}(det\varepsilon)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\varepsilon^{-1}(\mathbf{x}-\mu)} = \frac{1}{(2\pi)^{\frac{s}{2}}\Pi_{i=1}^s\sigma_i} e^{-\frac{1}{2}\sum_{i=1}^s \frac{(x_i-\mu_i)^2}{\sigma_i^2}} = \Pi_{i=1}^s \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}} \tag{3.14}$$

Ci aspettiamo quindi che la matrice della metrica sia anch'essa diagonale, il che equivale ad affermare che cambiando i parametri di una sola delle s gaussiane di una MGGD le altre mantengano la stessa distanza che avevano in precedenza rispetto ad una altra MGGD.

Andiamo ora a calcolare la matrice della metrica utilizzando eq. (3.6)

$$g_{\mu_i\mu_i} = \int_D \Pi_{i=1}^s \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}} \left(\frac{x_i-\mu_i}{\sigma_i^2}\right)^2 d\mathbf{x} = \frac{\sigma_i^2}{\sigma_i^4} = \frac{1}{\sigma_i^2} \tag{3.15}$$

in cui si è riconosciuta la definizione di varianza σ_i^2.

$$g_{\sigma_i\sigma_i} = \int_D \Pi_{i=1}^s \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} \left[\partial_{\sigma_i}\left(\sum_{i=1}^s -ln(\sqrt{2\pi}\sigma_i) - \frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)\right]^2 d\mathbf{x} =$$
$$\int \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}} \left(-\frac{1}{\sigma_i} + \frac{(x_i-\mu_i)^2}{\sigma_i^3}\right)^2 dx_i = \frac{1}{\sigma_i^2} - 2\frac{\sigma_i^2}{\sigma_i^4} + \int \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}} \frac{(x_i-\mu_i)^4}{\sigma_i^6} dx_i =$$
$$-\frac{1}{\sigma_i^2} + 3\int \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}} \left(\frac{x_i-\mu_i}{\sigma_i^2}\right)^2 dx_i = \frac{2}{\sigma_i^2} \tag{3.16}$$

in cui tra la terza e la quarta riga si è fatta una integrazione per parti.

$$g_{\mu_i\sigma_i} = \int_D \Pi_{i=1}^s \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}} \left(\frac{x_i-\mu_i}{\sigma_i^2}\right)\left(-\frac{1}{\sigma} + \frac{(x_i-\mu_i)^2}{\sigma_i^3}\right) d\mathbf{x} = 0 \tag{3.17}$$

per motivi di simmetria.

Ordinando quindi i parametri come di seguito $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, ..., \mu_s, \sigma_s)$ si ottiene quindi che la metrica è rappresentata dalla seguente matrice [10]:

$$\begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{\sigma_1^2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma_2^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{2}{\sigma_2^2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sigma_s^2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{2}{\sigma_s^2} \\ \end{bmatrix}$$

Da cui si deriva $ds^2 = \sum_{i=1}^{s} \frac{d\mu_i^2 + 2d\sigma_i^2}{\sigma_i^2}$.

Si può notare che, come ci si aspettava, la metrica è diagonale, e ne i parametri che caratterizzano una delle s gaussiane che compongono la MGGD ne una loro variazione hanno alcuna influenza sulla distanza calcolata tra altre coppie di gaussiane di due MGGDs.

In accordo con questo e ricordando che la distanza è definita come $d = \sqrt{\int_\gamma g_{ij} d\theta^i d\theta^j}$ ci si può aspettare che

$$d = \sqrt{\sum_{i=1}^{s} \left(\int_\gamma g_{2i,2i} d\mu^i d\mu^i + \int_\gamma g_{2i+1,2i+1}\, d\sigma^i d\sigma^i \right)} = \sqrt{\sum_{i=1}^{s} d_i} \quad , \tag{3.18}$$

dove d_i rappresenta la distanza tra la coppia i-esima delle gaussiane delle due MGGDs, e γ la geodetica. Questa considerazione permette di focalizzarci sulla distanza tra una sola coppia di gaussiane.
Si può notare che in questo modo la matrice della metrica riportata sopra ricorda quella di un piano iperbolico H^2 dove σ rappresenta l'asse y e μ l'asse x, costruita come

$$\begin{bmatrix} \frac{1}{\sigma_i^2} & 0 \\ 0 & \frac{1}{\sigma_i^2} \end{bmatrix}$$

da cui $ds^2 = \sum_{i=1}^{s} \frac{d\mu_i^2 + d\sigma_i^2}{\sigma_i^2}$.

Confrontando questo risultato con la metrica trovata precedentemente si può notare come le distanze tra due gaussiane caratterizzate dai parametri (μ_{i1}, σ_{i1}) e (μ_{i2}, σ_{i2}) siano legate dalla seguente relazione

$$d\left((\mu_{i1}, \sigma_{i1}), (\mu_{i2}, \sigma_{i2})\right) = \sqrt{2} d_{H^2}\left((\frac{\mu_{i1}}{\sqrt{2}}, \sigma_{i1}), (\frac{\mu_{i2}}{\sqrt{2}}, \sigma_{i2}) \right) \quad . \tag{3.19}$$

A questo punto partendo dalla conoscenza della distanza nel piano iperbolico si può ottenere la distanza nello spazio da noi considerato [10]

$$d\left((\mu_{i1}, \sigma_{i1}), (\mu_{i2}, \sigma_{i2})\right) = \sqrt{2} ln \left(\frac{|(\frac{\mu_{i1}}{\sqrt{2}}, \sigma_{i1}) - (\frac{\mu_{i2}}{\sqrt{2}}, -\sigma_{i2})| + |(\frac{\mu_{i1}}{\sqrt{2}}, \sigma_{i1}) - (\frac{\mu_{i2}}{\sqrt{2}}, \sigma_{i2})|}{|(\frac{\mu_{i1}}{\sqrt{2}}, \sigma_{i1}) - (\frac{\mu_{i2}}{\sqrt{2}}, -\sigma_{i2})| - |(\frac{\mu_{i1}}{\sqrt{2}}, \sigma_{i1}) - (\frac{\mu_{i2}}{\sqrt{2}}, \sigma_{i2})|} \right) \quad , \tag{3.20}$$

dove $|\cdot|$ indica la distanza calcolata tramite la metrica euclidea nello spazio (μ_i, σ_i).

Infine per quanto mostrato in eq. (3.18) possiamo definire la distanza totale tra due MGGDs come

$$d\left((\mu_1,\sigma_1),(\mu_2,\sigma_2)\right) = \sqrt{2\sum_{i=1}^{s}\left[ln\left(\frac{|(\frac{\mu_{i1}}{\sqrt{2}},\sigma_{i1})-(\frac{\mu_{i2}}{\sqrt{2}},-\sigma_{i2})|+|(\frac{\mu_{i1}}{\sqrt{2}},\sigma_{i1})-(\frac{\mu_{i2}}{\sqrt{2}},\sigma_{i2})|}{|(\frac{\mu_{i1}}{\sqrt{2}},\sigma_{i1})-(\frac{\mu_{i2}}{\sqrt{2}},-\sigma_{i2})|-|(\frac{\mu_{i1}}{\sqrt{2}},\sigma_{i1})-(\frac{\mu_{i2}}{\sqrt{2}},\sigma_{i2})|}\right)\right]^2} \quad . \tag{3.21}$$

3.2.2 Specie interagenti: medesimo vettore media e matrice di covarianza senza vincoli

Il secondo metodo permette invece di calcolare la distanza tra due MGGDs del tipo eq. (3.13) la cui forma della matrice di covarianza è senza vincoli, mentre il vettore rappresentante le medie deve essere uguale per entrambe le MGGDs.

In questo caso i parametri che determineranno la forma della funzione sono le entrate della matrice simmetrica di covarianza ε_{ij}, per un totale quindi di $\frac{s(s+1)}{2}$ parametri.

Mettiamo in evidenza la relazione [9]

$$\begin{aligned}
-\int_D p(\mathbf{x})\frac{\partial^2 \ln p(\mathbf{x})}{\partial\theta^i\partial\theta^j}\,d\mathbf{x} &= -\int_D p(\mathbf{x})\frac{\partial}{\partial\theta^i}\left(\frac{1}{p(\mathbf{x})}\frac{\partial p(\mathbf{x})}{\partial\theta^j}\right)d\mathbf{x} = \\
\int_D p(\mathbf{x})\frac{1}{p(\mathbf{x})^2}\frac{\partial p(\mathbf{x})}{\partial\theta^i}\frac{\partial p(\mathbf{x})}{\partial\theta^j}\,d\mathbf{x} &- \int_D p(\mathbf{x})\frac{1}{p(\mathbf{x})}\frac{\partial^2 p(\mathbf{x})}{\partial\theta^i\partial\theta^j}\,d\mathbf{x} = \\
\int_D p(\mathbf{x})\frac{1}{p(\mathbf{x})}\frac{\partial p(\mathbf{x})}{\partial\theta^i}\frac{1}{p(\mathbf{x})}\frac{\partial p(\mathbf{x})}{\partial\theta^j}\,d\mathbf{x} &- \frac{\partial^2}{\partial\theta^i\partial\theta^j}\int_D p(\mathbf{x})\,d\mathbf{x} = \\
\int_D p(\mathbf{x})\frac{\partial \ln p(\mathbf{x})}{\partial\theta^i}\frac{\partial \ln p(\mathbf{x})}{\partial\theta^j}\,d\mathbf{x} &= g_{ij}
\end{aligned} \tag{3.22}$$

in cui nell'ultimo passaggio si è ricordata eq. (3.6).

E' conveniente quindi in questo caso calcolare il valore di aspettazione della derivata seconda lungo i due parametri del logaritmo della distribuzione.

Tale calcolo risulta essere molto complicato e richiede conoscenze avanzate di matematica, si riporta dunque il risultato in forma differenziale [9]

$$-\int_D p(\mathbf{x})\mathrm{d}^2 \ln p(\mathbf{x})\,d\mathbf{x} = \frac{1}{2}tr\left(\varepsilon^{-1}\mathrm{d}\varepsilon\varepsilon^{-1}\mathrm{d}\varepsilon\right) \quad . \tag{3.23}$$

Per proseguire risulta comodo fare una cambio di parametri nel seguente del tipo $\varepsilon \to K^T\varepsilon K$, dove K è una matrice regolare che soddisfa $K^T\varepsilon_1 K = \Phi_1 = I_s$ matrice identità, e $K^T\varepsilon_2 K = \Phi_2$ con Φ_2 diagonale con entrate λ_i gli s autovalori della matrice $\varepsilon_1^{-1}\varepsilon_2$, con ε_1 e ε_2 matrici di covarianza delle distribuzioni di partenza e di arrivo.

Si può notare come la metrica è invariante sotto tale sostituzione, infatti $\varepsilon \to K^T\varepsilon K$ implica $\varepsilon^{-1} \to K^{-1}\varepsilon^{-1}K^{T-1}$ e $d\varepsilon \to K^T d\varepsilon K$, che inserito nel termine di destra di eq. (3.23) restituisce

$$\frac{1}{2}tr\left(K^{-1}\varepsilon^{-1}\mathrm{d}\varepsilon\varepsilon^{-1}\mathrm{d}\varepsilon K\right) = \frac{1}{2}tr\left(\varepsilon^{-1}\mathrm{d}\varepsilon\varepsilon^{-1}\mathrm{d}\varepsilon K^{-1}K\right) = \frac{1}{2}tr\left(\varepsilon^{-1}\mathrm{d}\varepsilon\varepsilon^{-1}\mathrm{d}\varepsilon\right) \quad , \tag{3.24}$$

dove nel primo passaggio si è usata la proprietà della traccia di essere invariante per commutazione.

Si è visto quindi che in questo modo la matrice di partenza è la matrice identica $\Phi_1 = I_s$ e la matrice di arrivo Φ_2 è diagonale.

Si fa ora un altro cambio $\Phi \to H^T \Phi H = \Lambda$, dove H è una matrice ortogonale, Λ è una matrice diagonale con autovalori λ_i.

Ricapitolando gli ultimi passaggi: la matrice K esegue un cambio di coordinate che diagonalizza le matrici di inizio e di arrivo contemporaneamente, rendendo in particolare la matrice iniziale una matrice identità, tuttavia nel percorso dalla matrice iniziale a quella finale la matrice potrebbe subire variazioni che la fanno discostare dalla forma diagonale, di conseguenza si introduce la matrice H ortogonale che si occupa di mantenere la matrice Λ diagonale durante tutto il percorso, "assorbendo" eventuali variazioni che altrimenti renderebbero Λ non diagonale.

Si fa inoltre notare che nell'istante iniziale e finale $H = I_s$ in quanto Φ_1 e Φ_2 sono già diagonali.
La metrica dipenderà ora dai parametri di H e Λ, e dalle relative variazioni.

Applicando questi ragionamenti si ottiene una forma della metrica del tipo [9]

$$-\int_D p(\mathbf{x}) \mathrm{d}^2 \ln p(\mathbf{x})\, d\mathbf{x} = \frac{1}{2}\sum_{i=1}^{s}\left(\frac{\mathrm{d}\lambda_i}{\lambda_i}\right)^2 + \sum_{i<j}\frac{(\lambda_i-\lambda_j)^2}{\lambda_i\lambda_j}(\mathrm{d}\eta_{ij})^2 \quad , \tag{3.25}$$

dove $d\eta_{ij}$ è l'elemento di matrice di $H^T dH$ e λ_i si ricordano essere gli autovalori della matrice diagonale Λ.

Dal momento che Φ è definita positiva, e quindi $\lambda_i > 0\ \forall i$, si ha che la curva è minima per $d\eta_{ij} = 0$, e poichè H è la matrice identità nel punto di partenza e nel punto di arrivo ne consegue che sarà la matrice identità lungo tutto il percorso e Φ, che quindi sarà sempre uguale a Λ, sarà diagonale lungo il percorso più breve, ovvero lungo la geodetica.

Possiamo introdurre ora un altro termine $r_i = \ln(\lambda_i)$, di modo tale che, usando r_i come parametro e utilizzando la relazione $dr_i = \frac{d\lambda_i}{\lambda_i}$, la metrica diventi

$$g_{ii} = \frac{1}{2} \qquad\qquad g_{ij} = 0 \tag{3.26}$$

Dal momento che i simboli di Christoffel si calcolano secondo eq. (3.11), si trova che questi sono tutti nulli, e per l'equazione delle geodetiche (eq. (3.12)), si trova che $\frac{d^2 r_i}{dt^2}(s) = 0$ e quindi $r_i = a_i s + b_i$, in cui $r_i(0) = 0$ punto di partenza e $r_i(1) = r_{i,2} = ln(\lambda_{i,2})$. dove si ricorda che $\lambda_{i,2}$ è autovalore i-esimo di $\varepsilon_1^{-1}\varepsilon_2$.

In questo modo otteniamo

$$d(\varepsilon_1,\varepsilon_2) = \int_{\varepsilon_1}^{\varepsilon_2} ds = \int_{\Phi_1=I_s}^{\Phi_2} ds = \int_{r_i=0}^{r_i=r_{i,2}} \left[\frac{1}{2}\sum_{i=1}^{s}(dr_i)^2\right]^{\frac{1}{2}} = \int_{t=0}^{t=1}\left[\frac{1}{2}\sum_{i=1}^{s}\left(\frac{\mathrm{d}r_i}{\mathrm{d}t}\right)\right]^{\frac{1}{2}} dt = \left[\frac{1}{2}\sum_{i=1}^{s}(r_{i,2})^2\right]^{\frac{1}{2}} \tag{3.27}$$

e ricordando che $r_{i,2} = ln(\lambda_{i,2})$ otteniamo il risultato finale [9]

$$d(\varepsilon_1,\varepsilon_2) = \left\{\frac{1}{2}\sum_{i=1}^{s}[\ln(\lambda_{i,2})]^2\right\}^{\frac{1}{2}} \tag{3.28}$$

Applicazione a comunità sintetiche

Il costrutto matematico sviluppato nelle sezioni precedenti è stato utilizzato per determinare la "distanza" tra diverse comunità la cui distribuzione è approssimabile ad una MGGD.

In particolare il fine è quello di valutare se sussistano differenze evidenti tra le distanze calcolate tra comunità sane-sane, malate-malate e sane-malate: ad esempio ci si potrebbe aspettare che le distribuzioni di comunità sane siano più simili tra di loro rispetto a quella di comunità malate in cui ognuno è malato a modo suo e di conseguenza le distribuzioni sono più disparate; se così fosse allora ci aspetteremmo che le distanze tra comunità sane-sane siano mediamente inferiori di quelle malati-malati.

Inoltre ci si potrebbe aspettare che le distanze tra comunità sane-malate siano mediamente maggiori delle distanze tra comunità sane-sane o malate-malate.

I dati analizzati non sono dati reali relativi al microbioma umano, ma sono una riproduzione numerica di possibili dati reali; ciò è motivato dal fatto che il processo usato per il calcolo delle distanze non è un processo utilizzato e diffuso, ma si propone essere un'alternativa ai metodi attuali, di conseguenza è necessario testare i limiti del processo con distribuzioni che conosciamo e che siamo in grado di controllare, di modo tale da vedere come il modello reagisce a variazioni dei parametri delle distribuzioni.

Inoltre un altro scopo di questa tesi è confrontare quale dei due metodi per calcolare le distanze sviluppati precedentemente sia più efficace per il tipo di distribuzioni che vengono trattate, dove per efficacia si intende la capacità di distinguere una comunità sana da una malata.

4.1 Generazione dei parametri

Lo studio di questi modelli è stato effettuato su MGGDs di dimensione 10, capaci di descrivere quindi 10 specie diverse e le interazioni delle une con le altre, il che implica che per ogni comunità sono stati generati $n_{param} = 10 + \frac{10(10+1)}{2}$ parametri, dove i primi 10 rappresentano il vettore media μ ed i restanti $\frac{10(10+1)}{2}$ appartengono alla matrice di covarianza.

Sono state generate un totale di 30 comunità sane e 30 comunità malate.

Il tipo di generazione dei parametri è differente per comunità sane e malate.

Si parte dal calcolo del vettore medie, e si fa l'ipotesi che questo vettore sia simile sia tra comunità sane che tra comunità malate, con la differenza però che le comunità malate abbiano più "libertà di movimento" attorno a questo valore; ciò è giustificato dal fatto che una malattia può far aumentare o diminuire la popolosità di una specie, tuttavia comunità malate diverse saranno soggette a malattie diverse e quindi si può ipotizzare che per una malattia che provoca la diminuzione della popolosità di una specie ne possa esistere un'altra che invece ne determina un aumento, quindi la media della popolosità di una data specie sarà circa la stessa sia per le comunità sane che per quelle malate, ma in quelle malate varierà maggiormente rispetto a quelle sane per quanto detto sopra.

Il processo è stato strutturato come segue:

- Si parte da una distribuzione di probabilità log-normale della forma

$$p(x) = \frac{e^{\frac{-(\ln(x)-m_m)^2}{2s_m^2}}}{x\sqrt{2\pi}s_m} \quad , \tag{4.1}$$

dove i parametri m_m e s_m danno la forma alla distribuzione, ed il pedice m indica che si sta considerando la generazione del vettore $m$edia.

- Da essa si estraggono 10 valori (1 per specie) che corrisponderanno ai centroidi delle gaussiane dalle quali si genereranno gli effettivi vettori medie delle comunità.
- Qui entra in gioco la distinzione tra comunità sane e comunità malate: si generano 20 gaussiane, due per specie, centrate nei valori estratti; La deviazione standard, σ_{ms} per le $m$edie delle comunità $s$ane e σ_{mm} per quelle $m$alate, la deviazione standard di queste gaussiane sarà più piccola del valore del centroide e sarà diversa tra comunità sane e malate, nello specifico sarà maggiore per le comunità malate.
- Da queste gaussiane si estraggono i valori delle medie per ogni spece per ogni comunità.

A titolo esemplificativo viene mostrata una immagine in cui il numero di specie è ridotto a 2:

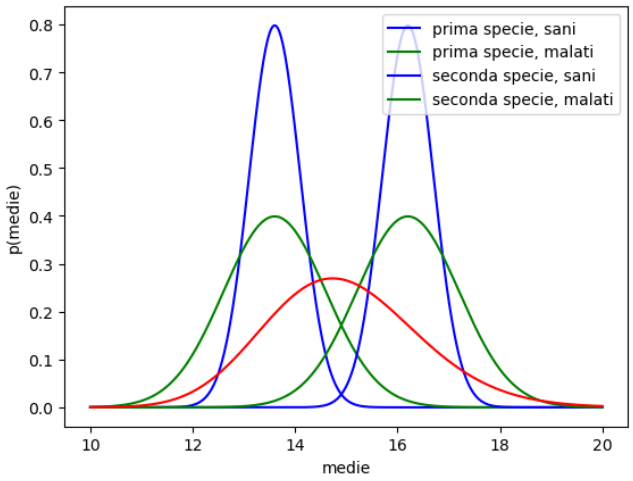


Figura 4.1: generazione medie

Dalla distribuzione log-normale raffigurata in rosso sono stati estratti due valori, che rappresentano i centroidi delle gaussiane, dopodichè si sono generate due gaussiane per ogni centroide, una per le comunità sane raffigurata in blu e caratterizzata da una deviazione standard ridotta, ed una per le comunità malate raffigurata in verde e con deviazione standard maggiore.

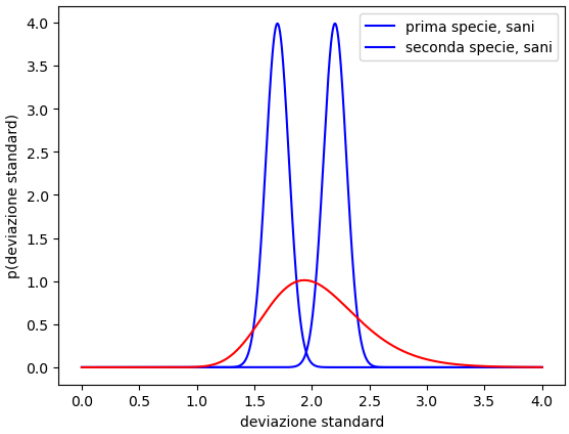
Infine da ogni gaussiana rappresentativa delle comunità sane si estraggono tanti valori quante sono il numero comunità sane, e lo stesso viene fatto per le comunità malate.

Al termine del processo quindi ogni comunità ha associato un vettore s-dimensionale che rappresenta le popolosità delle s specie.

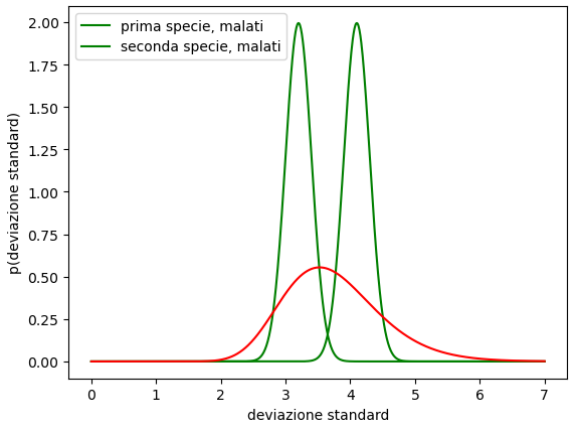
Per la generazione delle deviazioni standard delle specie (e di conseguenza della loro varianza e quindi della diagonale della matrice di covarianza), sono state generate due distribuzioni log-normali differenti: una per le comunità sane ed una per le comunità malate con valori mediamente più alti e più dispersi. Questo è giustificato dal fatto che in una comunità sana quasi ogni persona ha gli stessi valori di popolosità delle specie, di conseguenza la deviazione standard sarà ridotta, diversamente in una comunità malata, dal momento che le malattie in gioco possono essere differenti, c'è chi potrebbe avere una malattia che riduce la popolosità di una specie, e chi invece ne ha un'altra che la aumenta, determinando quindi allontanamenti dalla media della comunità molto più importanti che non nel caso di comunità sana.

La generazione dei parametri data la distribuzione log-normale avviene nello stesso modo descritto precedentemente, con la differenza che questa volta per ogni valore estratto dalla distribuzione log-normale ci sarà una sola gaussiana centrata in quel valore e non due.

Per rendere più chiaro si fa ancora un esempio per due sole specie:



(a) generazione deviazioni standard sani



(b) generazione deviazioni standard malati

Figura 4.2

Dal momento che le specie sono generalmente lievemente interagenti si sono generati i parametri fuori dalla diagonale di ogni matrice di covarianza casualmente tramite una gaussiana centrata nel valore nullo e con deviazione standard ridotta rispetto a quella delle specie di cui ne rappresenta l'interazione. Infine si è controllato che la matrice di covarianza così risultante fosse effettivamente definita positiva.

4.2 Generazione dei valori e calcolo delle distanze

Ora che le distribuzioni di probabilità per ogni comunità sono ben definite si possono generare i dati: ogni misurazione è rappresentata da un vettore s-dimensionale chiamato anche realizzazione che nel caso del microbioma umano rappresenterebbe la popolosità delle dieci specie batteriche estratte da un campione fecale di una singola persona.
La generazione avviene tramite il modello di Montecarlo esteso a s-dimensioni.
Sui dati così generati, il cui numero per comunità è di circa 10000, si applica il principio di massima entropia, ovvero si cerca la distribuzione che meglio li descrive dati i vincoli presentati nella sezione 2.2, ottenendo così una MGGD per ogni comunità il cui vettore media e matrice di covarianza sono uguali a quelli generati.
Infine si calcola la distanza tra le distribuzioni così generate tramite i due metodi differenti presentati in sezione 3.2.1 e 3.2.2. Per il metodo interagente si trascura il fatto che le MGGDs abbiano media differente, mentre per il caso non interagente si trascurano tutti i termini fuori dalla diagonale.

4.3 Analisi dell'andamento dei dati

Per avere una idea grafica di come ci può immaginare i parametri mostriamo come essi si distribuiscono in un piano (μ, σ) che è in analogia con il piano H^2 di cui si è fatta menzione precedentemente.

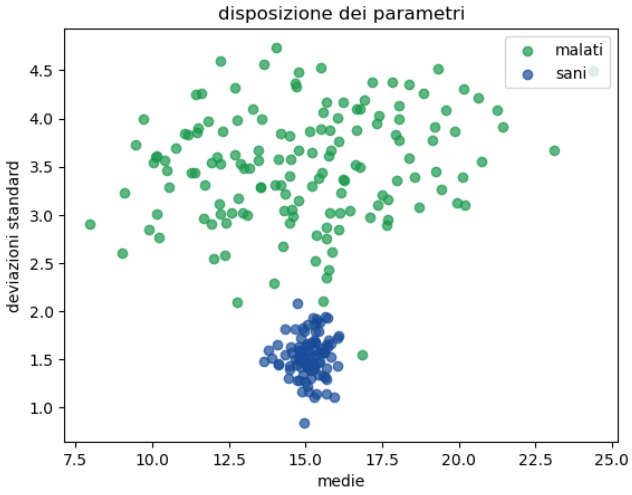


Figura 4.3: disposizione dei parametri

Questa immagine rappresenta i parametri di 100 comunità sane e 100 comunità malate (1 punto per comunità), tutte di dimensione 1 (descritte quindi da una unica gaussiana); questi dati servono solo per una visualizzazione grafica e non sono utilizzati nelle nostre analisi, le quali come detto in precendenza si concentrano su gaussiane 10-dimensionali.
In particolare:

- i punti verdi, punti "malati", sono tanto più alti rispetto ai punti blu, punti "sani", tanto maggiore è la media della log-normale che descrive la deviazione standard dei malati rispetto alla media della log-normale che descrive quella dei sani.

- i punti verdi(blu) sono tanto più dispersi in altezza tanto maggiore è il valore del parametro σ_{dm} (σ_{ds}) che rappresenta la deviazione standard della gaussiana da cui vengono generate le $d$eviazioni standard delle comunità $m$alate (sane).

- i punti verdi(blu) sono tanto più dispersi in larghezza tanto più alto è il valore del parametro σ_{mm} (σ_{ms}) che rappresenta la deviazione standard della gaussiana da cui vengono generate le $m$edie delle comunità $m$alate(sane).

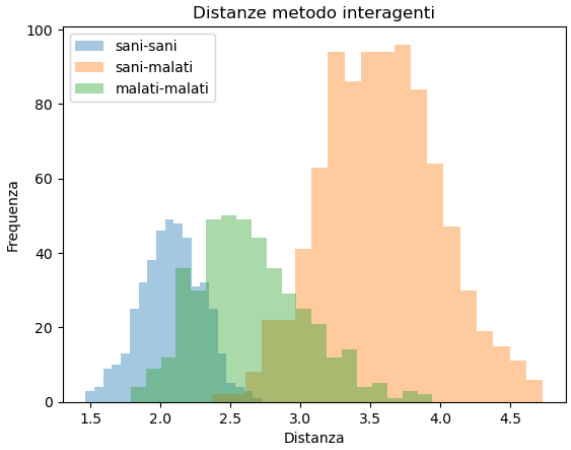
Infine si nota che la media del valore di μ delle disposizioni verde e blu è circa la stessa, ciò è dovuto al fatto che, a differenza delle deviazioni standard delle comunità, le medie sono generate dalla stessa log-normale.

Date le popolazioni generate per ogni comunità si è provveduto al calcolo delle distanze come chiarito precedentemente, in particolare le distanze che verranno rappresentate negli istogrammi e nei grafici da ora in poi saranno relative a distanze tra comunità sane-sane indicate con il colore blu, distanze tra comunità malate-malate indicate con il colore verde ed infine distanze tra comunità sane-malate indicate con il colore arancio.
In particolare dato che le comunità sane sono 30 e le comunità malate sono 30, ci si aspetta che il numero di conteggi delle distanze tra comunità sane-sane sia di $\frac{29(29+1)}{2} = 435$, ed equivalentemente per le distanze tra comunità malate-malate il numero di conteggi sia $\frac{29(29+1)}{2} = 435$ mentre il numero di conteggi delle distanze sane-malate è $30 \cdot 30 = 900$.

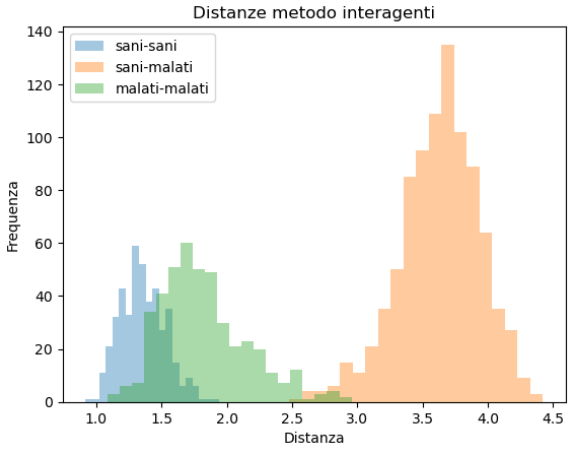
4.3.1 Variazione del numero di realizzazioni

Il primo trend che si è voluto controllare è l'evoluzione delle medie e delle deviazioni standard degli istogrammi rappresentanti le distanze al variare del numero di realizzazioni in una comunità da un minimo di 1000 realizzazioni ad un massimo di 100000.

I grafici di seguito mostrano l'andamento per il caso di distanze calcolate tramite il metodo interagente:

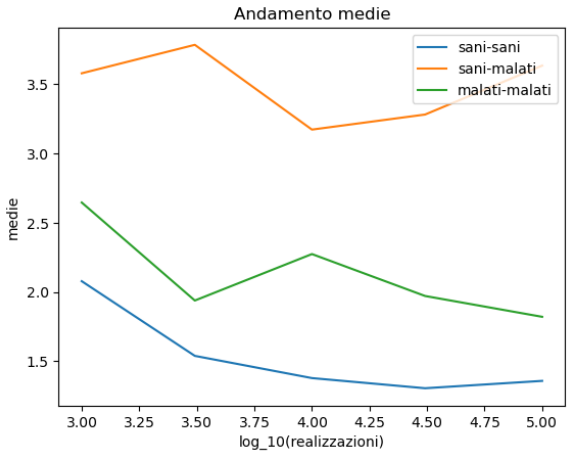


(a) distanze per 1000 realizzazioni per comunità

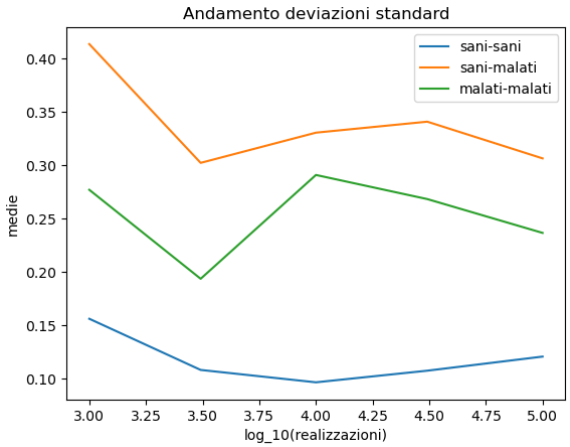


(b) distanze per 100000 realizzazioni per comunità

Figura 4.4



(a) andamento delle medie

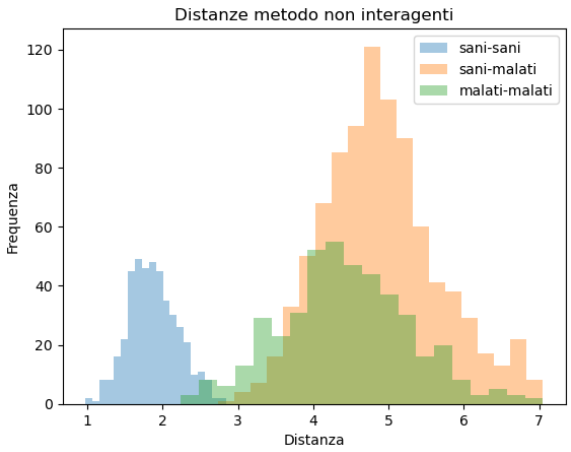


(b) andamento delle deviazioni standard

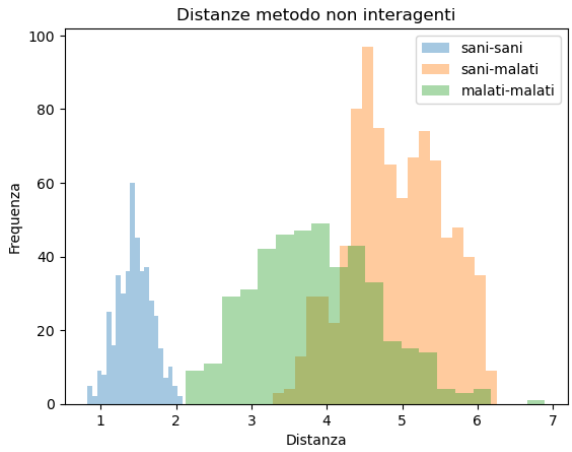
Figura 4.5

I primi due grafici in Figura 4.4 mostrano come si distribuiscono le distanze per il numero minimo ed il numero massimo di realizzazioni, mentre i grafici in Figura 4.5, il cui asse delle ordinate è in scala logaritmica, mostrano l'andamento lungo 5 valori del numero di realizzazioni: 1000, 3100, 10000, 31000, 100000.
Un dato che appare sin da subito evidente è come la distanza tra comunità sane-sane sia inferiore che non nelle altre due casistiche.
Si può notare come non sia presente alcun evidente trend che faccia pensare ad errori sistematici al variare del numero di realizzazioni, le distanze sembrano infatti essere soggette principalmente ad oscillazioni statistiche.

Lo stesso processo è stato fatto anche per il metodo delle distanze non-interagenti:

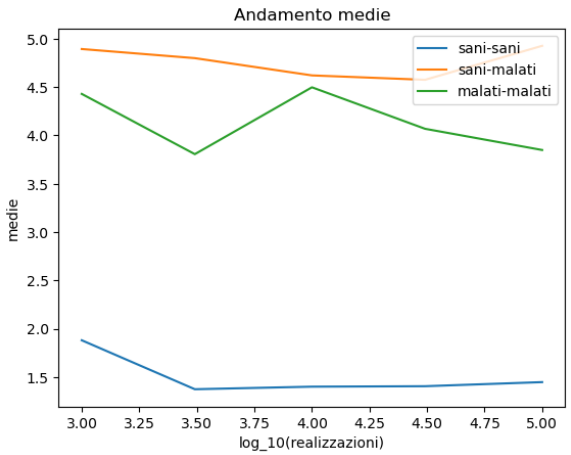


(a) distanze per 1000 realizzazioni per comunità

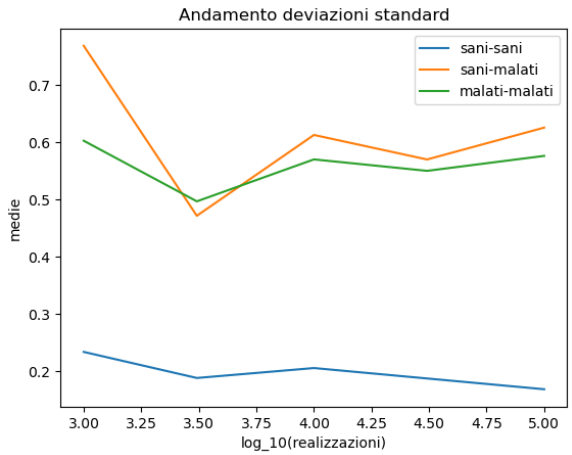


(b) distanze per 100000 realizzazioni per comunità

Figura 4.6



(a) andamento delle medie



(b) andamento delle deviazioni standard

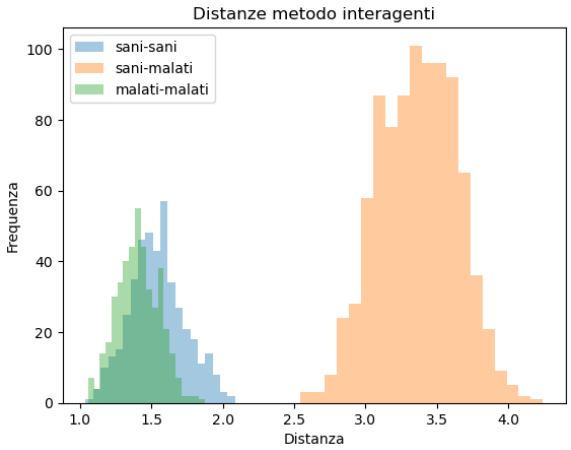
Figura 4.7

In questo caso ancora più che in quello precedente è evidente come la distribuzione delle distanze appaia indipendente dal numero di realizzazioni per ogni comunità.
Questi dati suggerirebbero che già un campione di sole 1000 realizzazioni, ovvero la raccolta di campioni fecali da 1000 individui, rappresenti un pool sufficiente per fare una analisi statistica di questo tipo, ed un aumento del numero di realizzazioni non permetterebbe di ottenere stime significativamente più precise e accurate.

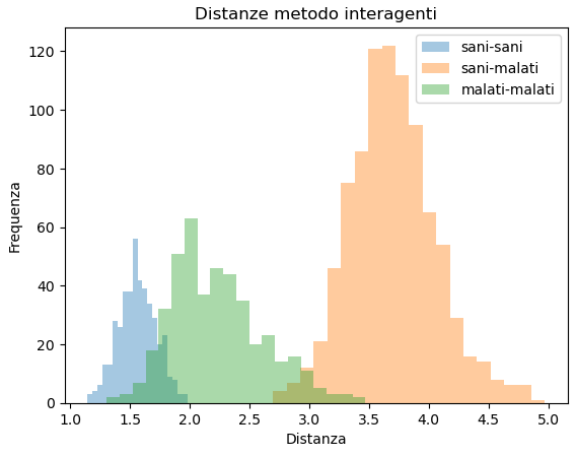
4.3.2 Modifica della variabilità della deviazione standard delle comunità malate σ_{dm}

Un altro tipo di analisi è stato effettuato variando la deviazione standard con cui si generano le deviazioni standard per le distribuzioni delle comunità malate (σ_{dm}), ovvero è stata variata la deviazione standard delle curve verdi rappresentate in Figura 4.2.

Si riportano i grafici che mostrano l'andamento delle distanze calcolate con il metodo interagente:



(a) distanze per $\sigma_{dm} = 0.1$



(b) distanze per $\sigma_{dm} = 1$

Figura 4.8

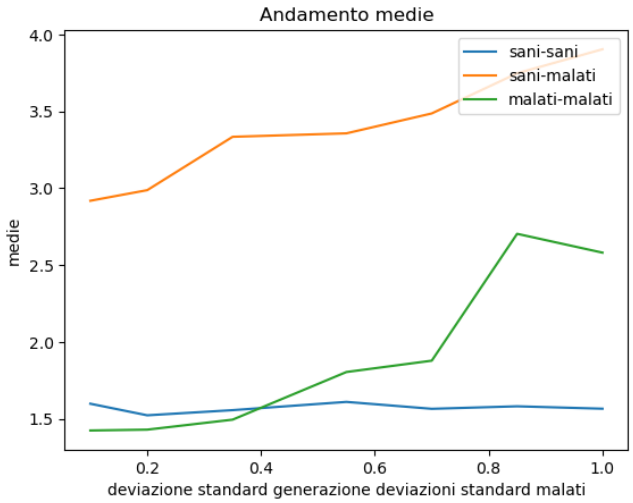
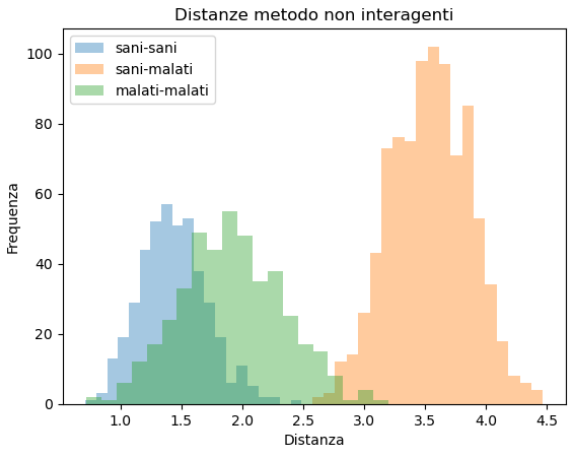


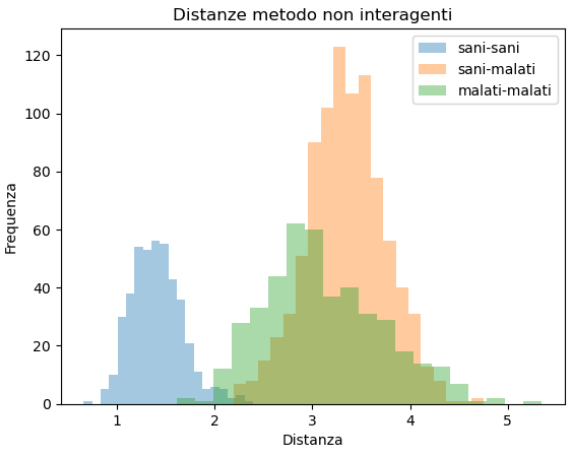
Figura 4.9: medie in funzione di σ_{dm}

Per riferimento, il valore dello stesso parametro per le distribuzioni sane è stato mantenuto fisso a $\sigma_{ds} = 0.2$, dove il pedice indica che ci stiamo riferendo alla *d*eviazione delle distribuzioni *s*ane.
Da questo grafico appare chiaro che le distanze tra comunità sane-sane non sono affette dalla variazione di questo parametro, in accordo con quanto ci aspetteremmo, mentre le distanza tra comunità malate-malate subisce una forte crescita, ed anche la distanza tra comunità sane-malate sebbene meno importante.
Tutto ciò è in accordo con quanto si poteva prevedere: infatti aumentare σ_{dm} determina una maggior variabilità delle deviazioni standard delle distribuzioni delle comunità malate, di conseguenza esse risulteranno più lontante tra loro (con riferimento a Figura 4.3 esse sono più disperse sull'asse y), ed allontanandosi determinano un incremento anche nella distanza tra comunità sane-malate.

Si riportano i grafici per lo stesso test nel caso non interagente:



(a) distanze per $\sigma_{dm} = 0.1$



(b) distanze per $\sigma_{dm} = 1$

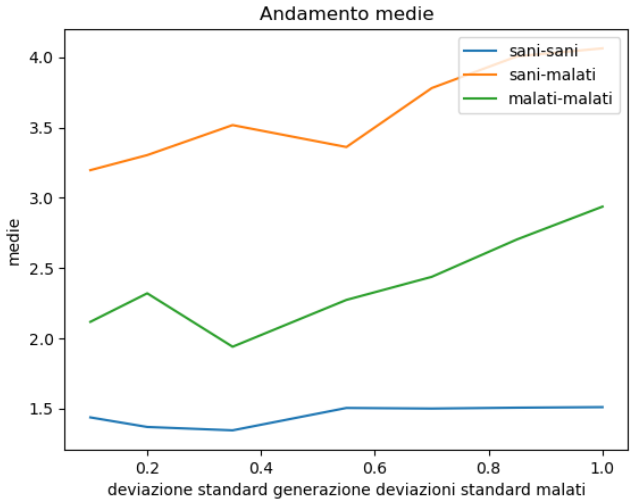


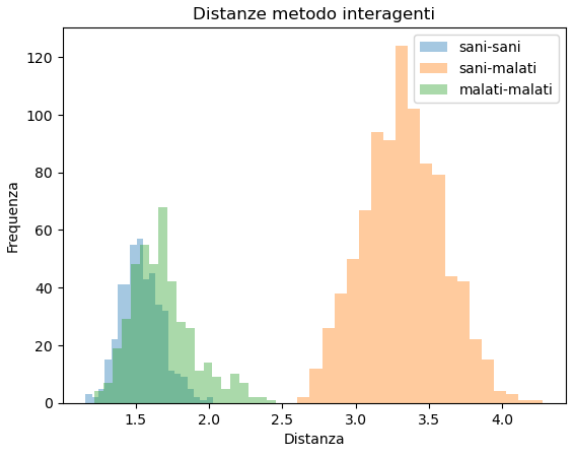
Figura 4.11: medie in funzione di σ_{dm}

Si può notare che per il metodo non-interagente si nota un lieve incremento delle distanze nel caso comunità malate-malate, tuttavia molto meno marcato che nel caso interagente.
Ciò è coerente con ciò che si poteva prevedere, infatti il metodo non interagente tiene conto anche della diversità dei vettori medie mentre effettua approssimazioni sulla differenza delle matrici di covarianza, di conseguenza ridurre σ_{dm}, ovvero rendere le matrici di covarianza delle comunità malate sempre più simili, non determina un effetto così rilevante. Inoltre le distanze tra le comunità malate-malate risulta sempre maggiore rispetto a quelle sane-sane dal momento che le medie delle comunità malate sono più disperse di quelle sane (dispersione orizzontale in riferimento a Figura 4.3)

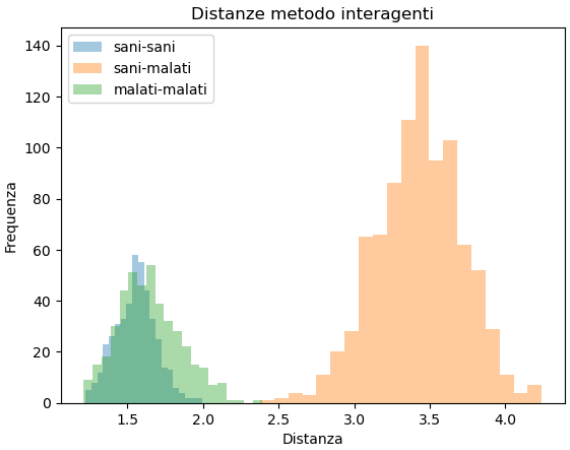
4.3.3 Modifica della variabilità della media delle comunità malate σ_{mm}

Si è deciso infine di indagare l'andamento delle distanze al variare del parametro σ_{mm} che quantifica quanto il vettore medie delle comunità malate si discosta dal vettore centroide della gaussiana da cui viene generato (con riferimento a Figura 4.1 il parametro regola la deviazione standard della curva verde). Questo parametro è stato fatto variare da un valore minimo di 0.5 (minima dispersione) ad uno massimo di 4 (massima dispersione), per confronto il valore per le comunità sane σ_{ms} vale 0.5.

Si ricorda che il metodo interagente non tiene conto della differenza tra i vettori media, quindi ci si può aspettare che questo metodo sia insensibile a questo tipo di variazioni.
Si riportano di seguito i grafici:



(a) distanze per $\sigma_{mm} = 0.5$



(b) distanze per $\sigma_{mm} = 4$

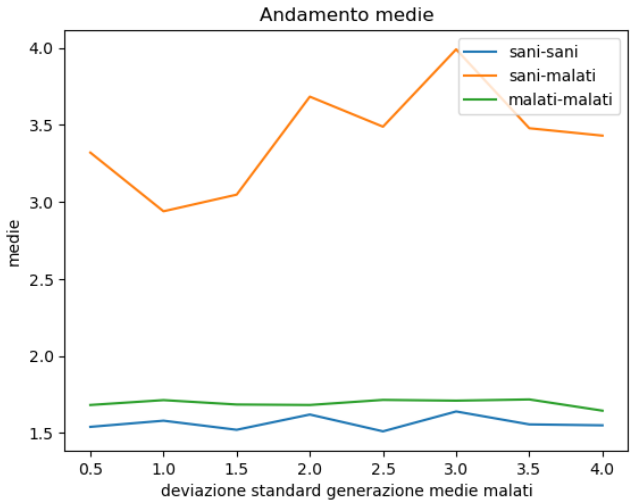
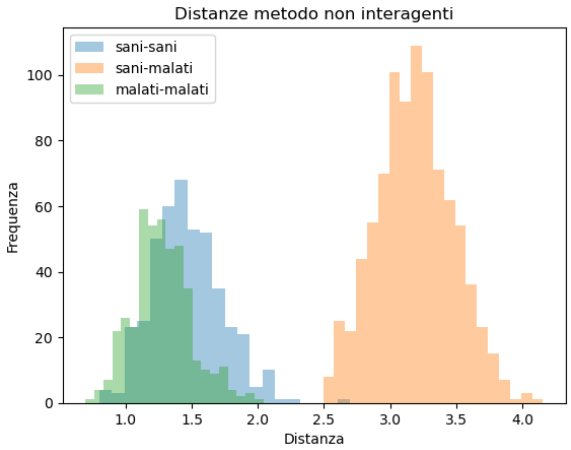


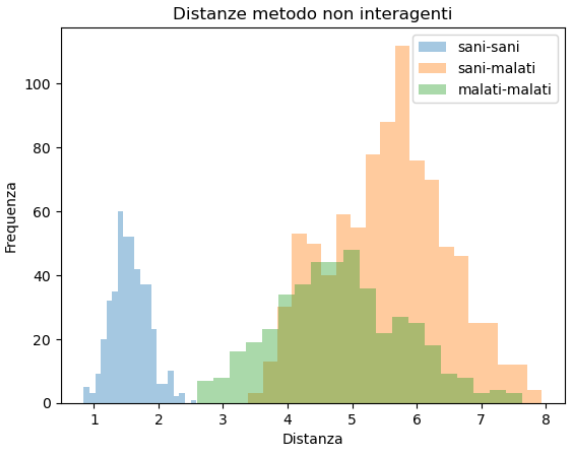
Figura 4.13: medie in funzione di σ_{mm}

Si può notare come i grafici mostrano che accade ciò che era stato ipotizzato, ovvero la distanza con il metodo interagente è completamente impassibile a cambi del vettore medie rimanendo costante al variare di questo.
Le distanze tra comunità sane-sane e sane-malate oscillano a cuasa della stocasticità con cui vengono generati i parametri iniziali delle comunità, tuttavia non presentano trend sistematici rilevanti.

Nel caso del metodo interagente invece ci si aspetta che la situazione cambi dal momento che la distanza calcolata dipende fortemente dalla somiglianza tra i vettori delle medie.
Vengono riportati di seguito i grafici che ne mostrano l'andamento:



(a) distanze per $\sigma_{mm} = 0.5$



(b) distanze per $\sigma_{mm} = 4$

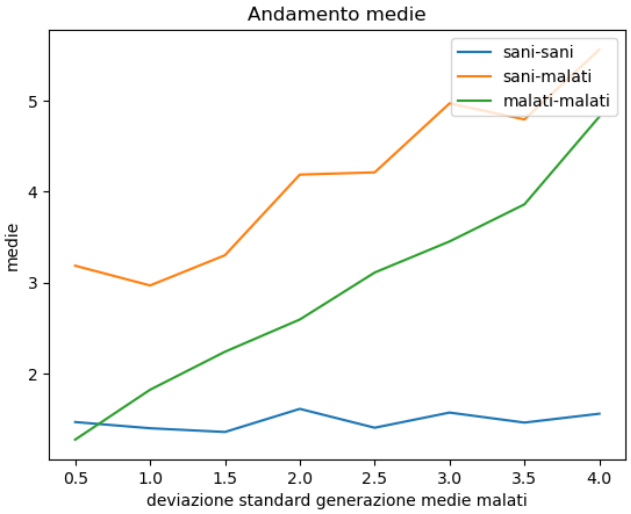


Figura 4.15: medie in funzione di σ_{mm}

In questo caso è evidente come le distanze siano estremamente influenzate dal valore di questo parametro. Esse infatti subiscono una brusca crescita all'aumentare del parametro in accordo con quanto ci aspetteremmo.
Con riferimento a Figura 4.3 le comunità malate si dispongono più compatte in direzione orizzontale al diminuire del valore di σ_{mm}, quindi oltre a diminuire drasticamente la distanza tra di esse diminuisce anche quella con le comunità sane, infatti il trend delle distanze tra comunità malate-malate è accompagnato da un analogo trend per le comunità sane-malate.

Discussione e Conclusioni

L'analisi fatta nella sezione precedente permette di evidenziare tre punti principali emersi in questo lavoro: 1) l'importanza del numero di realizzazioni necessarie, ovvero del numero di campioni, di una comunità ecologica affinché si riesca a distinguere effettivamente una comunità sana da una malata; 2) Le misure delle distanze tra comunità sia nel metodo non interagente 3.2.1 che in quello interagente 3.2.2 variano all'aumentare della variabilità della deviazione standard delle comunità malate; 3) La differenza tra i due metodi aumenta al variare della variabilità del vettore delle medie per le comunità malate.

I risultati ottenuti permettono inoltre di fare una distinzione tra quale metodo sia più efficace in determinati casi. In sezione 4.3.1, si è ricavato che un campione di 1000 individui è sufficiente per procedere ad una catalogazione della comunità come "malata" o "sana" tramite entrambi i metodi. Nonostante ciò il metodo interagente (vedi Figure 4.4 e 4.5) sembra essere in grado di distinguere con più facilità le comunità sane da quelle malate. Le distanze tra comunità sane-malate sono infatti maggiori di quelle sane-sane o malate-malate, mentre con il metodo non interagente (vedi Figure 4.6 e 4.7) le distanze tra comunità malate-malate sembrano sovrapporsi a quelle sane-malate.

La variazione di σ_{dm} analizzata in sezione 4.3.2 permette di distinguere entro quali limiti di σ_{dm} il metodo interagente è più efficace del metodo non interagente. Infatti guardando Figura 4.9 si può notare che le distanze tra malati-malati sono più basse di quelle sani-sani per valori di σ_{dm} sino a circa 2 volte σ_{ds}; ciò è dovuto al fatto che in generale il valore delle deviazioni standard delle comunità malate sono più alti di quelle sane (vedi Figure 4.2b e 4.2a in cui si vede che la log-normale per i malati ha la moda intorno a 3.5, mentre quella per i sani la ha intorno a 2), e dal momento che la distanza scala inversamente con il valore della deviazione standard ne consegue che quelle malate saranno meno diverse nella condizione di $\sigma_{dm} = \sigma_{ds}$. Di conseguenza al diminuire del parametro σ_{dm} il metodo interagente diventa molto più efficace del metodo non interagente, il quale deve tenere conto anche della distanza delle medie.

Infine, nella sezione 4.3.3 abbiamo mostrato che al diminuire di σ_{mm} il metodo non interagente diventa capace di distinguere le comunità sane da quelle malate sempre meglio, ed in particolare per valori di σ_{mm} uguali a σ_{ms} le distanze tra comunità malate-malate sono minori di quelle sane-sane.

5.1 Prospettive Future

Le analisi riportate permettono di considerare l'impatto dei parametri discussi tenendo fissi tutti gli altri parametri, i quali tuttavia sono stati scelti arbitrariamente (vedi appendice in Sezione 6). Se si fossero scelti parametri iniziali diversi l'analisi sarebbe stata leggermente diversa; di conseguenza sarebbe utile ricavare da dati sperimentali quali sono i valori effettivi di questi parametri, di modo da fare una analisi più specifica. Questo è però fattibile solo per comunità con poche specie. Inoltre sarebbe utile investigare come variano le distanze al variare di più parametri contemporaneamente, in particolare si potrebbe confrontare insieme gli effetti delle variazione di σ_{dm} e σ_{mm}, costruendo un grafico tridimensionale con i parametri su asse x e y, e il valore delle distanze sull' asse z. Infine, una volta fatta questa analisi con un confronto con i dati effettivi del microbioma umano e avendo visto effettivamente in che percentuale un metodo è più efficace dell'altro, si potrebbero integrare i due metodi come combinazione lineare pesata tra i due metodi; ad esempio se si verifica che il metodo interagente riesce a distinguere comunità sane da quella malate più efficacemente di quello non interagente nel 70% dei casi si potrebbe formulare una nuova metrica la cui distanza totale tiene conto per il 70% della distanza calcolata nel caso interagente e per il 30% quella calcolata con il caso non

interagente. E' fondamentale però scegliere un metodo ed usare lo stesso su tutte le comunità altrimenti le distanze calcolate dipendendo dal metodo scelto non si potrebbero confrontare. Ad esempio, se una comunità B si riconosce essere malata confrontandola con il metodo interagente con una comunità A che sappiamo essere sana, ed una comunità C invece risulta chiaramente malata tramite il metodo non interagente rispetto ad A, non saremmo in grado di dire quale delle due comunità B o C è *più malata*, ovvero più distante da A, in quanto si sono usati due metodi di misurazione delle distanze (metodo interagente e metodo non interagente) basati su procedure diverse e che si concentrano su parametri diversi, quindi le distanze così calcolate non sono confrontabili.

Appendice

Si riporta una tabella contenente i valori dei parametri fissi con relativa descrizione utilizzati per la generazione delle distribuzioni:

sigla	valore	descrizione
m_m	2.7	è correlata a valor medio e dev. std della log-norm in Figura 4.1
s_m	0.05	è correlata a valor medio e dev. std della log-norm in Figura 4.1
m_{ds}	0.7	è correlata a valor medio e dev. std della log-norm in Figura 4.2a
s_{ds}	0.2	è correlata a valor medio e dev. std della log-norm in Figura 4.2a
m_{dm}	1.3	è correlata a valor medio e dev. std della log-norm in Figura 4.2b
s_{dm}	0.2	è correlata a valor medio e dev. std della log-norm in Figura 4.2b
σ_{ms}	0.5	è la deviazione standard delle gaussiane blu in Figura 4.1
σ_{ds}	0.2	è la deviazione standard delle gaussiane in Figura 4.2a

Tabella 6.1: Parametri fissi

Bibliografia

[1] Sandro Azaele et al. ≪Statistical mechanics of ecological systems: Neutral theory and beyond≫. In: *Reviews of Modern Physics* 88.3 (2016), p. 035003.

[2] Etienne RS Wennekes PL Rosindell J. ≪The neutral-niche debate: a philosophical perspective≫. In: *Acta Biotheor.* (2012). DOI: 10.1007/s10441-012-9144-6. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3440563.

[3] Elizabeth K Costello et al. ≪The application of ecological theory toward an understanding of the human microbiome≫. In: *Science* 336.6086 (2012), pp. 1255–1262.

[4] Silvia Zaoli e Jacopo Grilli. ≪The stochastic logistic model with correlated carrying capacities reproduces beta-diversity metrics of microbial communities≫. In: *PLOS Computational Biology* 18.4 (apr. 2022), pp. 1–14. DOI: 10.1371/journal.pcbi.1010043. URL: https://doi.org/10.1371/journal.pcbi.1010043.

[5] Braden T Tierney et al. ≪Systematically assessing microbiome–disease associations identifies drivers of inconsistency in metagenomic research≫. In: *PLoS biology* 20.3 (2022), e3001556.

[6] Jayanth R Banavar, Amos Maritan e Igor Volkov. ≪Applications of the principle of maximum entropy: from physics to ecology≫. In: *Journal of Physics: Condensed Matter* 22.6 (gen. 2010), p. 063101. DOI: 10.1088/0953-8984/22/6/063101. URL: https://dx.doi.org/10.1088/0953-8984/22/6/063101.

[7] Crispin W Gardiner et al. *Handbook of stochastic methods*. Vol. 3. springer Berlin, 1985.

[8] Igor Volkov et al. ≪Inferring species interactions in tropical forests≫. In: *Proceedings of the National Academy of Sciences* 106.33 (2009), pp. 13854–13859. DOI: 10.1073/pnas.0903244106. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0903244106. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0903244106.

[9] Geert Verdoolaege e Paul Scheunders. ≪On the Geometry of Multivariate Generalized Gaussian Models≫. In: *Journal of Mathematical Imaging and Vision* 43 (lug. 2011), pp. 180–193. DOI: 10.1007/s10851-011-0297-8.

[10] Sueli I. R. Costa, Sandra A. Santos e João E. Strapasson. *Fisher information distance: a geometrical reading*. 2014. arXiv: 1210.2354 [stat.ME].

[11] Dorje C Brody e Daniel W Hook. ≪Information geometry in vapour–liquid equilibrium≫. In: *Journal of Physics A: Mathematical and Theoretical* 42.2 (dic. 2008), p. 023001. DOI: 10.1088/1751-8113/42/2/023001. URL: https://dx.doi.org/10.1088/1751-8113/42/2/023001.