Caratterizzazione e Clustering di Matrici di Correlazione ADNI-2 Tramite Distribuzione di Wishart

Carlo Mengucci

ALMA MATER STUDIORUM · UNIVERSITÁ DI BOLOGNA

28 Novembre 2017

Obiettivi dello Studio

- Organizzazione ed esplorazione preliminare della sezione del DataBase ADNI-2 contenente le matrici di correlazione di *Resting State* di soggetti sani e affetti da Alzheimer
- Giustificazione analitica dell'utilizzo della *Distribuzione di Wishart* come ipotesi nulla per la caratterizzazione
- Sviluppo di un algoritmo in grado di separare i soggetti sani dai malati e di individuare un gruppo di significativo di *features* per il clustering

ADNI-2

- 404 soggetti totali
- Ad ogni soggetto è associata una matrice di correlazione $N \times N$, $N = 549$
- Ognuno degli N elementi rappresenta un *Macrovoxel* di cui è estratta la correlazione *topologica* rispetto a tutte le altre componenti del sistema
- Ogni *Macrovoxel* è definito su un insieme di $3 \cdot 10^3$ Voxel
- Dei 404 soggetti sono stati utilizzati soltanto gli ultimi 300 a causa di discrepanze tra procedure di normalizzazione tra due sezioni distinte individuabili all'interno del DataBase

Forma Analitica [a]

[a] *Hardle, Wolfgang and Leopold Simar. 2012. Applied Multivariate Statistical Analysis. Heidelberg: Springer Berlin Heidelberg*

- La distribuzione di Wishart consiste in una famiglia di distribuzioni per *matrici simmetriche definite positive*
- Siano $X_1 ... X_n$ vettori indipendenti $N_p(0, \Sigma)$ e tali da formare una matrice di dati $p \times n$, $X = [X_1 ... X_n]$. La distribuzione di *matrici random* $p \times p$, $M = XX' = \Sigma_{i=1}^{n} X_i X_i'$ è una distribuzione di Wishart.

Forma Analitica [a]

[a] *Hardle, Wolfgang and Leopold Simar. 2012. Applied Multivariate Statistical Analysis. Heidelberg: Springer Berlin Heidelberg*

- La matrice random $M_{p\times p} = \Sigma_{i=1}^{n} X_i X_i'$ segue una distribuzione di Wishart a n gradi di libertà e *matrice di covarianza* Σ ed è definita $M \sim W_p(n, \Sigma)$. Per $n \geq p$ la *pdf* di M assume la forma :

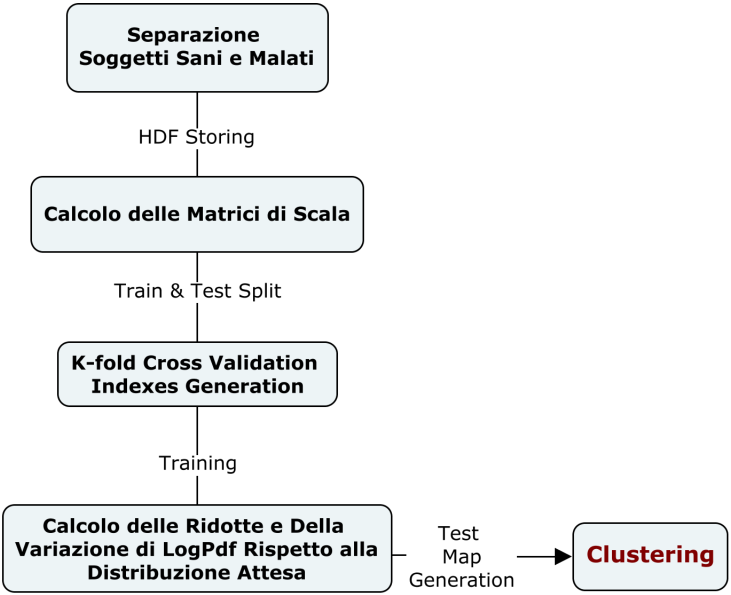
$$f(M) = \frac{1}{2^{\frac{np}{2}}\Gamma_p(\frac{n}{2})\|\Sigma\|^{\frac{n}{2}}}\|M\|^{\frac{n-p-1}{2}} exp[-\frac{1}{2} trace(\Sigma^{-1}M)] \tag{1}$$

- La Wishart può essere interpretata come l'estensione multivariata di una distribuzione χ^2

Definizione delle Condizioni di Applicazione

- Le matrici di correlazione sono per definizione simmetriche definite positive
- Il numero n di gradi di libertà del sistema è dato dal campionamento ($n = 3 \cdot 10^3$) del singolo Macrovoxel
- Utilizzando come matrice di scala la matrice data dalla media delle matrici di correlazione delle due categorie di soggetti, è possibile ricostruire la Wishart attesa per le categorie stesse.

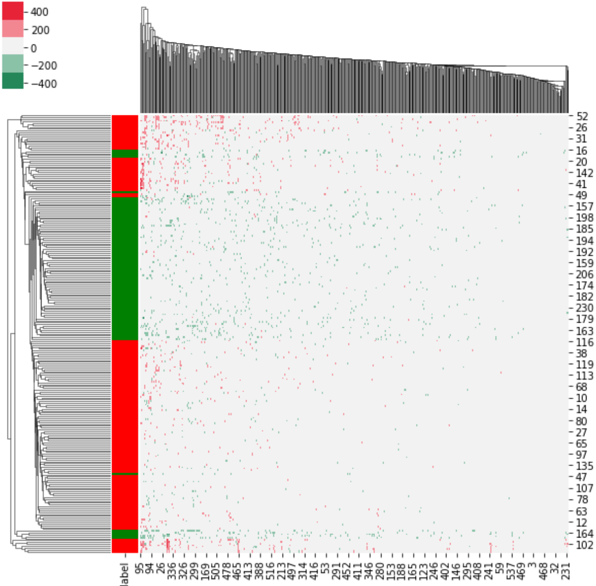
Pipeline


Separazione
Soggetti Sani e Malati
HDF Storing
Calcolo delle Matrici di Scala
Train & Test Split
K-fold Cross Validation
Indexes Generation
Training
Calcolo delle Ridotte e Della
Variazione di LogPdf Rispetto alla
Distribuzione Attesa
Test
Map
Generation
Clustering

Calcolo delle Ridotte: Approfondimento

- *Definizione*: sia A una matrice $n \times n$, una sottomatrice $k \times k$ di A ottenuta eliminando $n-k$ righe e le stesse $n-k$ colonne di A è detta *sottomatrice principale*.
- *Teorema*: Se A è simmetrica definita positiva, ogni sottomatrice principale di A è anch'essa simmetrica definita positiva.
- Queste proprietà permettono di calcolare il peso che ogni componente del sistema possiede, in termini di variazione di LogPdf rispetto alla matrice di scala generatrice della distribuzione attesa.
- Vengono cioè valutate N variazioni di LogPdf su N sottomatrici principali di dimensioni $(N-1) \times (N-1)$.

Clustering


400
200
0
−200
−400
52
26
31
16
20
142
41
49
157
198
185
194
192
159
206
174
182
230
179
163
116
38
119
113
68
10
14
80
27
65
97
135
47
107
78
63
12
164
102
label
95
94
26
336
326
299
169
505
478
465
413
388
516
213
497
314
416
53
291
452
411
346
280
153
188
165
123
246
402
146
295
408
241
59
537
469
3
468
32
231

LogPdf Distribution