Modelli mistura

Vers. 1.1.1

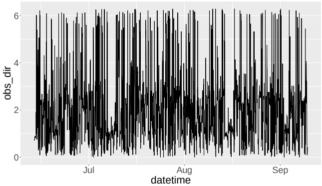
Gianluca Mastrantonio

email: gianluca.mastrantonio@polito.it

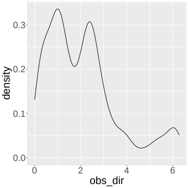
Esempio Introduttivo

Wind data

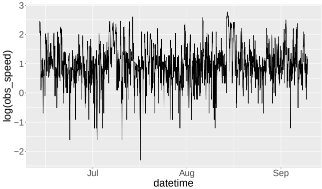
Partiamo analizzando dei dati di direzione e intensità del vento, registrati ogni ora. Vediamo prima la direzione



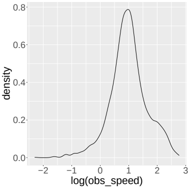
(a)



(b)



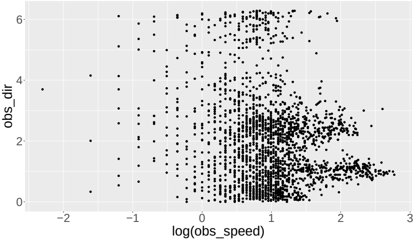
(c)



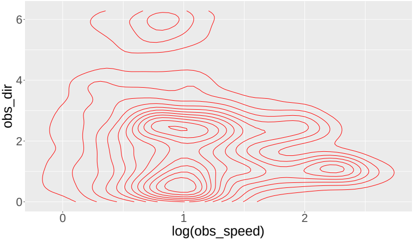
(d)

Wind data

che mostra chiaramente almeno due mode. Nella log-velocità invece sembra essercene una sola, o forse due, anche se la seconda è veramente piccola.
Un modo migliore è di vedere la distribuzione bivariata con uno scatterplot e una stima di densità



(e)

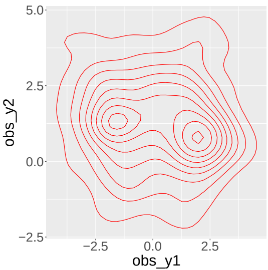


(f)

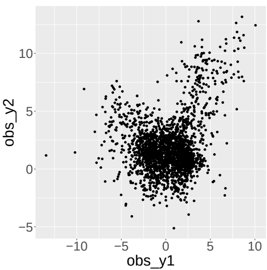
Se indichiamo con η_i la direzione e con x_i la velocità, potremmo anche provare a lavorare con

$$y_{i,1} = x_i \cos(\eta_i) \qquad y_{i,2} = x_i \sin(\eta_i)$$

Wind data



(g)



(h)

Per modellizzare i dati, possiamo assumere che esista un variabile discreta

$$z_i \in \{1, \ldots, L\}$$

che rappresenta lo **stato** del vento. Naturalmente non è nota e è una variabile latente.

Modelli Mistura

Il modello mistura

Si parla di modelli mistura ogni volta che si ha una situazione in cui i parametri di una distribuzioni cambiano in base a una variabile latente discreta $z_i \in \{1, \ldots, L\}$, per esempio

$$y_y | z_i \sim F(\boldsymbol{\theta}_{z_i})$$

e non conoscendo $\boldsymbol{\theta}_{z_i}$, si deve assumere una qualche distribuzioni (le variabili latenti sono sempre variabili aleatorie), per esempio

$$z_i \sim Discrete(\boldsymbol{\pi}_i)$$

dove $Discrete()$ indica che la variabile può assumere valore k con probabilità data da $\boldsymbol{\pi}$. La distribuzione discreta si può scrivere anche come

$$z_i \sim Discrete(\pi_1, \ldots, \pi_k)$$

Il modello mistura

Assumiamo, per semplicità che i dati siano condizionatamente normali. In questo caso, il modello mistura più semplice è

$$
\begin{aligned}
y_i | z_i = k &\sim N(\mu_k, \sigma_k^2) \\
z_i &\sim Discrete(\boldsymbol{\pi})
\end{aligned}
$$

Visto che stiamo modellizzando tutto in un ottica Bayesiana, dobbiamo mettere delle prior sia sui parametri delle verosimiglianza che su $\boldsymbol{\pi}$, per esempio

$$
\begin{aligned}
y_i | z_i = k, \{\mu_j, \sigma_j^2\}_{j=1}^{L} &\sim N(\mu_k, \sigma_k^2) \\
z_i | \boldsymbol{\pi} &\sim Discrete(\boldsymbol{\pi}) \\
\boldsymbol{\pi} &\sim Dir(\boldsymbol{\alpha}) \\
\mu_k &\sim N(0, 10000) \\
\sigma_k^2 &\sim IG(1, 1)
\end{aligned}
$$

$Dir()$ è la distribuzione di Dirichlet, che ha densità

$$f(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{L} \alpha_k) \prod_{k=1}^{L} \pi_k^{\alpha_k - 1}}{\prod_{k=1}^{L} \Gamma(\alpha_k)} \propto \prod_{k=1}^{L} \pi_k^{\alpha_k - 1}$$

con $\alpha_i > 0$, e si può vedere come una generalizzazione della distribuzione Beta. La Dirichlet si può scrivere anche come

$$Dir(\alpha_1, \ldots, \alpha_L)$$

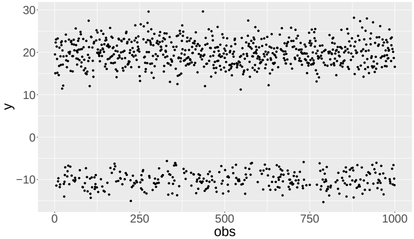
quando si vogliono specificare i parametri. Nel modello mistura, almeno in questa forma, il numero L deve essere noto, e non può essere stimato direttamente, se non facendo particolarmente attenzione. Non lo si può trattare come un qualsiasi parametro dato che la dimensione dello spazio della a posteriori dipende da L. Per ora assumiamo di conoscerlo, poi usando indici quali il WAIC, possiamo decidere il suo valore.

Il modello mistura

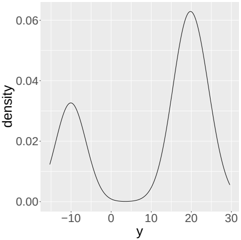
Possiamo simulare dal modello, con dei parametri fissi, per avere un'idea di come si distribuiscono i dati. Assumiamo sempre che i dati siano condizionatamente normali, com

$$L = 2 \qquad \boldsymbol{\pi} = (0.3, 0.7) \qquad \sigma_1^2 = 2^2 \qquad \sigma_2^2 = 3^2$$

e cambiamo i valori di μ_1 e μ_2

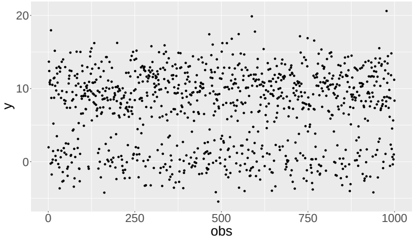


(i) $\mu_1 = -10, \mu_2 = 20$

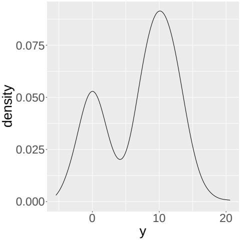


(j) $\mu_1 = -10, \mu_2 = 20$

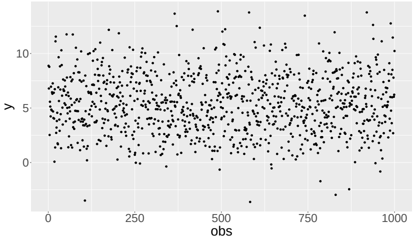
Il modello mistura



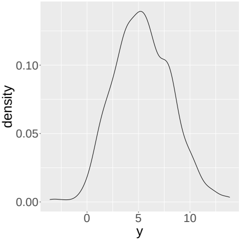
(k) $\mu_1 = 0, \mu_2 = 10$



(l) $\mu_1 = 0, \mu_2 = 10$



(m) $\mu_1 = 4, \mu_2 = 6$



(n) $\mu_1 = 4, \mu_2 = 6$

Visto che $\mathbf{z}$ è una variabile aleatoria, possiamo marginalizzare e vedere quale è la vera distribuzione che stiamo assumendo su $\mathbf{y}$. Visto che

$$f(\mathbf{y}, \mathbf{z}) = \prod_{i=1}^{n} f(y_i, z_i) = \prod_{i=1}^{n} f(y_i|z_i) f(z_i)$$

abbiamo che

$$f(\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{L} f(y_i|z_i = k) f(z_i = k) =$$

$$\prod_{i=1}^{n} \left(\sum_{k=1}^{L} \pi_k f(y_i|z_i = k) \right) = \prod_{i=1}^{n} f(y_i)$$

e, se $y_i|z_i = k \sim N(\mu_k, \sigma_k^2)$, abbiamo che

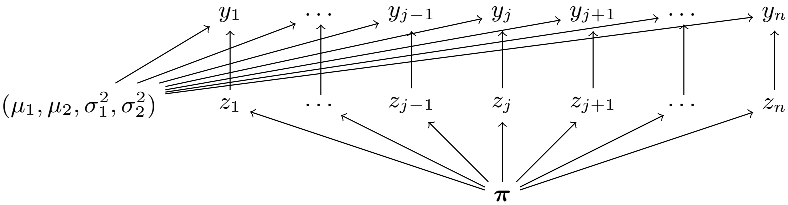
$$f(y_i) = \sum_{k=1}^{L} \pi_k (2\pi\sigma_k^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right)$$

Stima Bayesiana

Da un punto di vista Bayesiano, la stima de modelli mistura è generalmente semplice. Il grafo del modello Bayesiano

$$
\begin{aligned}
y_i | z_i = k, \{\mu_j, \sigma_j^2\}_{j=1}^{L} &\sim N(\mu_k, \sigma_k^2) \\
z_i | \boldsymbol{\pi} &\sim Discrete(\boldsymbol{\pi}) \\
\boldsymbol{\pi} &\sim Dir(\boldsymbol{\alpha}) \\
\mu_k &\sim N(m, v) \\
\sigma_k^2 &\sim IG(a, b)
\end{aligned}
$$

in questo caso è



che ha posteriori (con un L qualsiasi)

$$f(\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2, \pi_1, \ldots, \pi_L, \mathbf{z}|\mathbf{y}) \propto$$
$$\left[\prod_{i=1}^{n}\prod_{k=1}^{L}\left((\sigma_k^2)^{-\frac{1}{2}} \exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right)^{\mathbf{1}_k(z_i)}\right]\left[\prod_{i=1}^{n}\prod_{k=1}^{L}\pi_k^{\mathbf{1}_k(z_i)}\right] \times$$
$$\left[\prod_{k=1}^{L}\pi_k^{\alpha_k-1}\right]\left[\prod_{k=1}^{L}\exp\left(-\frac{(\mu_k-m)^2}{2v}\right)\right]\left[\prod_{k=1}^{L}(\sigma_k^2)^{-a-1}\exp\left(-\frac{b}{\sigma_k^2}\right)\right]$$

Full conditional of μ_k: Se definiamo

$$I_k = \{i : z_i = k\}$$

come il set di indici i per cui $z_i = k$ che ha cardinalità n_k, la full conditional dipende da

$$\left[\prod_{i \in I_k}(\sigma_k^2)^{-\frac{1}{2}} \exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right]\left[\exp\left(-\frac{(\mu_k-m)^2}{2v}\right)\right]$$

che è simile alla full conditional di un modello con dati indipendente, e è proporzionale a

$$\propto \exp\left(-\frac{1}{2}\left(\mu_k^2\left(\frac{n_k}{\sigma_k^2}+\frac{1}{v}\right)-2\mu_k\left(\frac{\sum_{i\in I_k} y_i}{\sigma^2}+\frac{m}{v}\right)\right)\right)$$

e quindi

$$\mu_k|\ldots \sim N\left(\left(\frac{n_k}{\sigma_k^2}+\frac{1}{v}\right)^{-1}\left(\frac{\sum_{i\in I_k} y_i}{\sigma^2}+\frac{m}{v}\right),\left(\frac{n_k}{\sigma_k^2}+\frac{1}{v}\right)^{-1}\right)$$

Full conditional of σ_k^2: Dipende solo da

$$\left[\prod_{i\in I_k}(\sigma_k^2)^{-\frac{1}{2}}\exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right]\left[(\sigma_k^2)^{-a-1}\exp\left(-\frac{b}{\sigma_k^2}\right)\right]$$

e quindi

$$\sigma_k^2|\cdots \sim IG\left(a+\frac{n_k}{2}, b+\frac{\sum_{i\in I_k}(y_i-\mu_k)^2}{2}\right)$$

Stima Bayesiana

Full conditional of $\boldsymbol{\pi}$: dipende solo da

$$\left[\prod_{i=1}^{n}\prod_{k=1}^{L}\pi_k^{\mathbf{1}_k(z_i)}\right]\left[\prod_{k=1}^{L}\pi_k^{\alpha_k-1}\right] = \left[\prod_{k=1}^{L}\pi_k^{n_k}\right]\left[\prod_{k=1}^{L}\pi_k^{\alpha_k-1}\right] = \prod_{k=1}^{L}\pi_k^{\alpha_k+n_k-1}$$

e quindi

$$\boldsymbol{\pi}|\cdots \sim Dir(\alpha_1+n_1,\ldots,\alpha_L+n_L)$$

Full conditional of z_i: dipende solo da

$$\left[\prod_{k=1}^{L}\left((\sigma_k^2)^{-\frac{1}{2}}\exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right)^{\mathbf{1}_k(z_i)}\right]\left[\prod_{k=1}^{L}\pi_k^{\mathbf{1}_k(z_i)}\right]$$

e per semplicità indico come

$$g_k = \left((\sigma_k^2)^{-\frac{1}{2}}\exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right)$$

allora

$$z_i|\cdots \sim Discrete\left(\frac{\pi_1 g_1}{\sum_{k=1}^{L}\pi_k g_k},\ldots,\frac{\pi_L g_L}{\sum_{k=1}^{L}\pi_k g_k}\right)$$

Stima Bayesiana

Campionare dalla full conditional di z_i è complicato dal punto di vista numerico, perchè' molti dei termini $\pi_k g_k$ potrebbero essere particolarmente piccoli/grandi e il calcolo di

$$\pi_j g_j / \sum_{k=1}^{L} \pi_k g_k$$

può avere problemi numerici.
Abbiamo due soluzioni, che si basano sull'assunzione che sappiamo o abbiamo calcolato $\log(g_k)$, che è generalmente facile

- **Log-Sum-Exp Trick**:

$$\frac{\pi_j g_j}{\sum_{k=1}^{L} \pi_k g_k} = exp\left(\log(\pi_j g_j) - \log\left(\sum_{k=1}^{L} \pi_k g_k\right)\right) =$$

$$exp\left(\log(\pi_j g_j) - \left(c + \log\left(\sum_{k=1}^{L} \exp\left(\log(\pi_k g_k) - c\right)\right)\right)\right)$$

dove $c = \max\{\log(\pi_1 g_1) \ldots, \log(\pi_L g_L)\}$.

Stima Bayesiana

- **Gumbel Trick**: Il secondo metodo non richiede il calcolo della costante di normalizzazione, ma campiona direttamente da

$$z_i | \dots$$

conoscendo solo il vettore di probabilità in scala logaritmica non normalizzato: $(\log(\pi_1 g_1), \dots, \log(\pi_L g_L))$. L'idea è di campionare L variabili $x_k \sim Gumbel(0,1)$ e poi settare

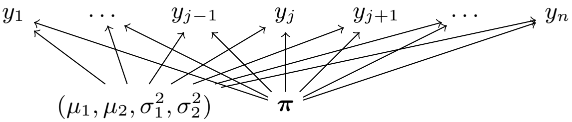
$$z_i = \mathrm{argmax}_k(\log(\pi_k g_k) + x_k)$$

e si può dimostrare che $\mathrm{argmax}_k(\log(\pi_k g_k) + x_k)$ proviene da

$$Discrete\left(\frac{\pi_1 g_1}{\sum_{k=1}^{L} \pi_k g_k}, \dots, \frac{\pi_L g_L}{\sum_{k=1}^{L} \pi_k g_k}\right)$$

Stima Bayesiana - Marginalizzazioni

In un modello mistura si può marginalizzare sia su $\mathbf{z}_i$ che su $\boldsymbol{\pi}$, che su entrambe. Se marginalizziamo su $\mathbf{z_i}$, abbiamo il seguente modello



Abbiamo già visto che la verosimiglianza diventa una media pesata delle singole componenti, e la a posteriori è quindi

$$
\begin{aligned}
&f(\mu_1, \dots, \mu_L, \sigma_1^2, \dots, \sigma_L^2, \pi_1, \dots, \pi_L | \mathbf{y}) \propto \\
&\left[\prod_{i=1}^{n} \sum_{k=1}^{L} \pi_k \left((\sigma_k^2)^{-\frac{1}{2}} \exp\left(\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right) \right) \right] \times \\
&\left[\prod_{k=1}^{L} \pi_k^{\alpha_k - 1}\right] \left[\prod_{k=1}^{L} \exp\left(-\frac{(\mu_k - m)^2}{2v} \right)\right] \left[\prod_{k=1}^{L} (\sigma_k^2)^{-a-1} \exp\left(-\frac{b}{\sigma_k^2} \right)\right]
\end{aligned}
$$

Stima Bayesiana - Marginalizzazioni

Adesso marginalizziamo su $\boldsymbol{\pi}$, mantenendo $\mathbf{z}$

La marginalizzazione di $\boldsymbol{\pi}$ è più complicata, ma si può fare prima di tutto notando che se vogliamo marginalizzare (integrare) rispetto a $\boldsymbol{\pi}$, dobbiamo calcolare solo

$$\int_S f(\mathbf{z}|\boldsymbol{\pi})f(\boldsymbol{\pi})d\boldsymbol{\pi} = \int_S \left[\prod_{i=1}^{n}\prod_{k=1}^{L} \pi_k^{\mathbf{1}_k(z_i)}\right]\left[\prod_{k=1}^{L} \pi_k^{\alpha_k - 1}\right] d\boldsymbol{\pi} = \int_S \prod_{k=1}^{L} \pi_k^{n_k+\alpha_k-1} d\boldsymbol{\pi}$$

dove S è il simplesso. Abbiamo già visto che quello dentro l'integrale è il kernel di una Dirichlet, e allora il valore dell'integrale è la costante di normalizzazione

$$\int_S \prod_{k=1}^{L} \pi_k^{n_k+\alpha_k-1} d\boldsymbol{\pi} = \frac{\prod_{k=1}^{L}\Gamma(\alpha_k + n_k)}{\Gamma(\sum_{k=1}^{L}(\alpha_k + n_k))}$$

Stima Bayesiana - Marginalizzazioni

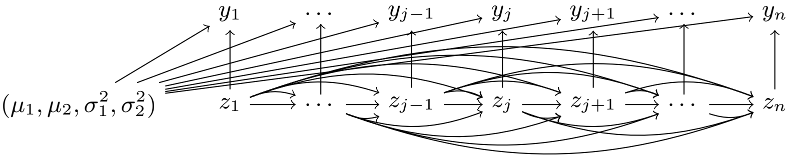
e quindi la a posteriori è

$$f(\mu_1,\ldots,\mu_L,\sigma_1^2,\ldots,\sigma_L^2,\mathbf{z}|\mathbf{y}) \propto$$

$$\left[\prod_{i=1}^{n}\prod_{k=1}^{L}\left((\sigma_k^2)^{-\frac{1}{2}}\exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right)^{\mathbf{1}_k(z_i)}\right]\left[\frac{\prod_{k=1}^{L}\Gamma(\alpha_k+n_k)}{\Gamma(\sum_{k=1}^{L}(\alpha_k+n_k))}\right]\times$$

$$\left[\prod_{k=1}^{L}\exp\left(-\frac{(\mu_k-m)^2}{2v}\right)\right]\left[\prod_{k=1}^{L}(\sigma_k^2)^{-a-1}\exp\left(-\frac{b}{\sigma_k^2}\right)\right]$$

questo corrisponde a un modello



dove si vede che abbiamo introdotto dipendenza tra le variabili latenti. Anche se la distribuzione di $\mathbf{z}$ è complessa, possiamo ancora campionare dalla full conditional facilmente.

Ipotizziamo di voler campionare da z_i.

- utilizziamo la proprietà della funzione gamma per cui

$$\Gamma(x+1) = x\Gamma(x)$$

- definiamo n_j^{-i} come la numerosità del gruppo j senza tenere in considerazione il valore di z_i.

Stima Bayesiana - Marginalizzazioni

Abbiamo che

$$f(\mathbf{z}) = f(z_i|\mathbf{z}_{-i})f(\mathbf{z}_{-i}) \propto \prod_{k=1}^{L} \Gamma(\alpha_k + n_k^{-1} + \mathbf{1}_k(z_i)) =$$

$$= \left((n_{z_i}^{-i} + \alpha_{z_i})\Gamma(\alpha_{z_i} + n_{z_i}^{-i})\right) \prod_{\substack{k=1 \\ k \neq z_i}}^{L} \Gamma(\alpha_k + n_k^{-i}) =$$

$$(n_{z_i}^{-i} + \alpha_{z_i}) \prod_{k=1}^{L} \Gamma(\alpha_k + n_k^{-i}) \propto$$

$$n_{z_i}^{-i} + \alpha_{z_i}$$

visto che $\prod_{k=1}^{L} \Gamma(\alpha_k + n_k^{-i})$ è costante per ogni valore di z_i.

Stima Bayesiana - Marginalizzazioni

Quindi

$$P(z_i = 1 | \dots) = \frac{n_1^{-i} + \alpha_1}{\sum_{k=1}^{L} (\alpha_k + n_k^{-i})}$$

$$P(z_i = 2 | \dots) = \frac{n_2^{-i} + \alpha_2}{\sum_{k=1}^{L} (\alpha_k + n_k^{-i})}$$

$$\dots$$

$$P(z_i = L | \dots) = \frac{n_L^{-i} + \alpha_L}{\sum_{k=1}^{L} (\alpha_k + n_k^{-i})}$$

e quindi la full conditional è

$$z_i | \dots \sim Discrete \left(\frac{(n_1^{-i} + \alpha_1) g_1}{\sum_{k=1}^{L} (n_k^{-i} + \alpha_k) g_k}, \dots, \frac{(n_L^{-i} + \alpha_L) g_L}{\sum_{k=1}^{L} (n_k^{-i} + \alpha_k) g_k} \right)$$

Label Switching

Tutti i modelli mistura, stimati tramite una procedura Bayesiana, soffrono del problema del **label-switching**.

Per capire da dove nasce il problema assumiamo che se

$$y_i|z_i \sim F(\theta_{z_i})$$

e $z_i \in \{1, 2\}$ e $i = 1, 2, 3$. La verosimiglianza di

$$z_1 = 1, z_2 = 2, z_3 = 1 \qquad \theta_1 = 0, \theta_2 = 100$$

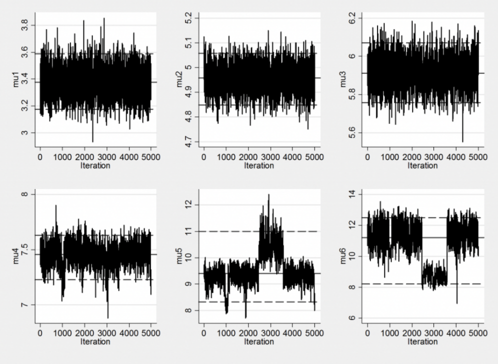
è la stessa di

$$z_1 = 2, z_2 = 1, z_3 = 2 \qquad \theta_1 = 100, \theta_2 = 0$$

Quindi, se cambiamo (facciamo un switch di) tutte le label del processo latente, e i valori dei parametri nello stesso modo, la verosimiglianza non cambia $\Rightarrow$ il modello non è identificabile. Dal punto di vista pratico, capita raramente, me può accadere.

Label Switching

Se c'è del label switching vedrete delle catene che saltano in continuazione tra 2 o più punti. Per esempio



In questo caso ci sono degli approcci per risolverlo, che riordinano i campioni. Per esempio in R possiamo usare il pacchetto **label.switching**.

Dataset Wind

Il modello più semplice che possiamo fare è il seguente

$$\log x_i | z_i \sim N(\mu_{z_i}, \sigma^2_{z_i})$$
$$\eta_i | z_i \sim WC(\rho_{z_i}, \tau_{z_i})$$
$$z_i \sim Disc(\boldsymbol{\pi})$$

dove $WC(\rho, \tau)$ indica la wrapped Cauchy, che è una distribuzioni per variabili circolari, che ha densità

$$f(\eta|\rho, \tau) = \frac{\sinh(\tau)}{2\pi\left(\cosh(\tau) - \cos(\eta - \rho)\right)}$$

con $\mu \in [0, 2\pi)$ e $\tau > 0$.

Dataset Wind

Come prior usiamo

$$\boldsymbol{\pi} \sim Dir(\mathbf{1})$$
$$\mu_k \sim N(0, 10000^2)$$
$$\sigma_k^2 \sim IG(1, 1)$$
$$\tau_k \sim G(1, 1)$$
$$\rho_k \sim U(0, 2\pi)$$

Se stimiamo modelli con $L = 2, 3, 4$ abbiamo che il migliore è $L = 4$ (WAIC = 10715.38, 9730.153, 9180.096).

	μ_1	μ_2	μ_3	μ_4	π_1	π_2	π_3	π_4	ρ_1	ρ_2
Media	1.88	1.11	0.20	0.88	0.13	0.35	0.16	0.36	1.06	2.39
L 95%CI	1.75	1.05	-0.03	0.84	0.11	0.30	0.10	0.30	1.04	2.35
U 95%CI	2.00	1.17	0.41	0.91	0.16	0.40	0.22	0.42	1.08	2.43

Dataset Wind

	ρ_3	ρ_4	σ_1^2	σ_2^2	σ_3^2	σ_4^2	τ_1	τ_2	τ_3	τ_4
Media	1.90	0.50	0.21	0.25	0.44	0.11	0.13	0.32	1.56	0.46
L 95%CI	1.31	0.40	0.15	0.22	0.35	0.09	0.11	0.27	1.11	0.37
U 95%CI	2.55	0.59	0.29	0.29	0.56	0.13	0.15	0.37	2.51	0.54

Per la variabile z_i abbiamo due possibilità per ottenere un valore rappresentativo

- per ogni i, prendiamo il valore più probabile
- calcoliamo per ogni iterazione la verosimiglianza, e prendiamo il vettore $\mathbf{z}$ che la massimizza (in teoria dovremmo marginalizzare rispetto a tutti gli altri parametri)

Dataset Wind

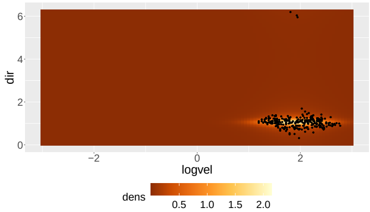
Utilizziamo la prima e otteniamo le seguenti frequenze per i 4 gruppi: 265 791 234 834.

Possiamo vedere le stime delle distribuzioni a posteriori per i singoli cluster. Per far questo dobbiamo calcolare l'integrale

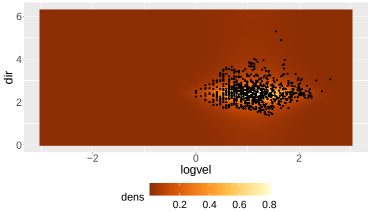
$$f(\log x^*, \eta^* | z = k, \mathbf{y}) = \int f(\log x^*, \eta^* | z = k, \boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) f(\mathbf{z}, \boldsymbol{\theta} | \mathbf{y})$$

dove $\boldsymbol{\theta}$ sono i parametri del modello. Possiamo campionare da questa distribuzione, oppure vedere $f(\log x^*, \eta^* | z = k, \boldsymbol{\theta}, \mathbf{y}, \mathbf{z})$ come una funzione e calcolarlo come integrale MC per diversi punti del dominio. Utilizziamo il secondo metodo e plottiamo anche le osservazioni associate al cluster (che hanno il valore $\hat{z}_i$ uguale a quello del cluster)

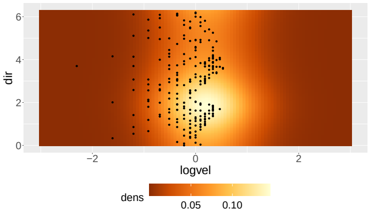
Dataset Wind



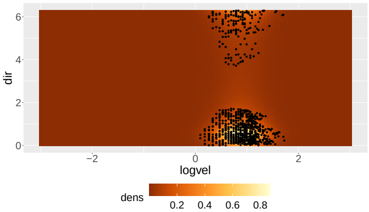
(a) Cluster 1



(b) Cluster 2



(c) Cluster 3



(d) Cluster 4

Hidden Markov Models

Riprendiamo l'esempio precedente, e guardiamo un po' meglio la stima di $\hat{\mathbf{z}}$. Mostriamo la tabella delle frequenze tra $\hat{\mathbf{z}}_{1:n-1}$ e $\hat{\mathbf{z}}_{2:n}$ che viene

	1	2	3	4
1	194	23	5	43
2	22	477	73	219
3	5	70	85	74
4	43	221	71	498

Se prendiamo i valori per riga, per esempio riga 2, questo ci dice quante volte sono passato dallo stato 2 ad altri stati. Per facilitare la lettura, dividiamo ogni elemento per il totale di riga, ottenendo

	1	2	3	4
1	0.73	0.09	0.02	0.16
2	0.03	0.60	0.09	0.28
3	0.02	0.30	0.36	0.32
4	0.05	0.27	0.09	0.60

Introduzione

Posso confrontare questa tabella, che indico con $\mathbf{P}$, con quella che otterrei, teoricamente, se non ci fosse nessuna relazione tra $\hat{\mathbf{z}}_{1:n-1}$ e $\hat{\mathbf{z}}_{2:n}$, i cui valori teorici sono, per ogni cella, pari a totale_riga $\times$ totale_colonna$/n$. Se dividiamo anche questo per il totale di riga, otteniamo

	1	2	3	4
1	0.12	0.37	0.11	0.39
2	0.12	0.37	0.11	0.39
3	0.12	0.37	0.11	0.39
4	0.12	0.37	0.11	0.39

Rispetto al'atteso, c'è una forte persistenza a stare nello stato in cui ci si trovava, visto che $[\mathbf{P}]_{j,j}$ sono molto elevati.

HMM

Vogliamo introdurre della dipendenza temporale tra gli stati latenti. La cosa più facile che possiamo fare è assumere che z_i sia una catena di Markov con una data matrice di transizioni $\mathbf{P}$, con elementi $\pi_{k,j}$, e vettore riga $\boldsymbol{\pi}_k$

Se un modello ha una struttura latente discreta che segue una catena di Markov, è un Hidden Markov Model (HMM).
Per fare un esempio,

$$
\begin{aligned}
y_i | z_i, \{\boldsymbol{\theta}_k\}_{k=1}^{L} &\sim F(\boldsymbol{\theta}_{z_i}) \\
z_i | z_{i-1} &\sim Discrete(\boldsymbol{\pi}_{z_{i-1}}), i = 1, \ldots, n
\end{aligned}
$$

in questo caso z_0 è un parametro. IL DAG è

$$
\begin{tikzcd}
y_1 & \cdots & y_{j-1} & y_j & y_{j+1} & \cdots & y_n \\
z_1 \arrow[u] \arrow[r] & \cdots \arrow[u] \arrow[r] & z_{j-1} \arrow[u] \arrow[r] & z_j \arrow[u] \arrow[r] & z_{j+1} \arrow[u] \arrow[r] & \cdots \arrow[u] \arrow[r] & z_n \arrow[u]
\end{tikzcd}
$$

HMM - Forward-Backward Algorithm

Anche in questo caso potremmo calcolare la densità marginale. Sono interessato a

$$f(\mathbf{y}) = \sum_{k_1=1}^{L} \cdots \sum_{k_n=1}^{L} f(\mathbf{y}|z_1 = k_1, \ldots, z_n = k_n) f(z_1 = k_1, \ldots, z_n = k_n)$$

Il calcolo è complicato, ma possiamo usare il **Forward-Backward Algorithm**

Partiamo definendo

$$a_1(k) = f(y_1, z_1 = k) = \pi_{z_0,k} f(\mathbf{y}|z_1 = k)$$

Calcoliamo poi

$$a_2(k) = f(y_1, y_2, z_2 = k) = \sum_{k_1=1}^{L} f(y_1, y_2, z_2 = k|z_1 = k_1) f(z_i = k_1) =$$

$$\sum_{k_1=1}^{L} f(y_2|z_2 = k, y_1, z_1 = k_1) f(z_2 = k|y_1, z_1 = k_1) f(y_1|z_1 = k_1) f(z_i = k_1)$$

$$\sum_{k_1=1}^{L} f(y_2|z_2 = k) f(z_2 = k|z_1 = k_1) a_1(k_1) = \sum_{k_1=1}^{L} f(y_2|z_2 = k) \pi_{k_1,k} a_1(k_1)$$

HMM - Forward-Backward Algorithm

Abbiamo quindi che in generale

$$a_i(k) = f(y_1, \dots, y_i, z_i = k) = \sum_{k_{i-1}=1}^{L} f(y_1, \dots, y_i, z_i = k | z_{i-1} = k_{i-1}) f(z_{i-1} = k_{i-1}) =$$

$$\sum_{k_{i-1}=1}^{L} f(y_i | z_i = k, z_{i-1} = k_{i-1}, y_1, \dots, y_{i-1}) f(z_i = k | z_{i-1} = k_{i-1}, y_1, \dots, y_{i-1}) \times$$

$$f(y_1, \dots, y_{i-1} | z_{i-1} = k_{i-1}) f(z_{i-1} = k_{i-1}) =$$

$$\sum_{k_{i-1}=1}^{L} f(y_i | z_i = k) \pi_{k_{i-1},k} a_{i-1}(k_{i-1})$$

Possiamo poi calcolare

$$f(\mathbf{y}) = \sum_{k=1}^{L} a_n(k) = \sum_{k=1}^{L} f(y_1, \dots, y_n, z_n = k)$$

Nell congiunta marginale abbiamo dipendenza tra le osservazioni

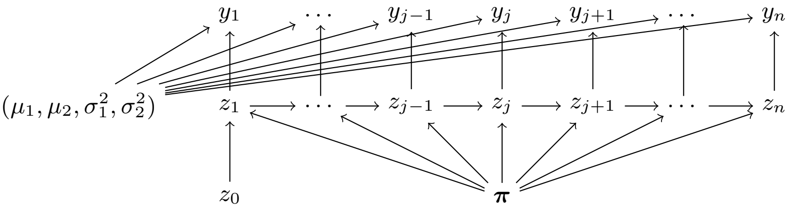
HMM - Stima Bayesiana

Se vogliamo stimarlo come un modello Bayesiano dobbiamo mettere delle prior e definire una verosimiglianza. Possiamo assumere

$$
\begin{aligned}
y_i|z_i = k, \{\mu_j, \sigma_j^2\}_{j=1}^L &\sim N(\mu_k, \sigma_k^2) \\
z_i|z_{i-1}, \boldsymbol{\pi}_{z_{i-1}} &\sim Discrete(\boldsymbol{\pi}_{z_{i-1}}), i = 1, \ldots, n \\
z_0 &\sim Discrete(\boldsymbol{\rho}) \\
\boldsymbol{\pi}_k &\sim Dir(\boldsymbol{\alpha}) \\
\mu_k &\sim N(m, v) \\
\sigma_k^2 &\sim IG(a, b)
\end{aligned}
$$

Il DAG di questo modello diventa

HMM - Stima Bayesiana



che ha posteriori (con un L qualsiasi)

$$f(\mu_1, \dots, \mu_L, \sigma_1^2, \dots, \sigma_L^2, z_0, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_L, \mathbf{z}|\mathbf{y}) \propto$$

$$\left[\prod_{i=1}^{n}\prod_{k=1}^{L}\left((\sigma_k^2)^{-\frac{1}{2}}\exp\left(\frac{(y_i-\mu_k)^2}{2\sigma_k^2}\right)\right)^{\mathbf{1}_k(z_i)}\right]\left[\prod_{j=1}^{L}\prod_{k=1}^{L}\pi_{j,k}^{n_{j,k}}\right]\times$$

$$\left[\prod_{k=1}^{L}\rho_k^{\mathbf{1}_k(z_0)}\right]\left[\prod_{j=1}^{L}\prod_{k=1}^{L}\pi_{j,k}^{\alpha_k-1}\right]\left[\prod_{k=1}^{L}\exp\left(-\frac{(\mu_k-m)^2}{2v}\right)\right]\left[\prod_{k=1}^{L}(\sigma_k^2)^{-a-1}\exp\left(-\frac{b}{\sigma_k^2}\right)\right]$$

dove $n_{k,j}$ indica il numero di tutte le transizioni da k a j e quindi

$$\prod_{j=1}^{L}\prod_{k=1}^{L}\pi_{j,k}^{n_{j,k}} = \prod_{i=1}^{n}\prod_{k=1}^{L}\pi_{z_{i-1},k}^{\mathbf{1}_k(z_i)} = \prod_{i=1}^{n}\prod_{k=1}^{L}\pi_{z_{i-1},z_i}$$

Abbiamo che

Full conditional of μ_k: La stessa del modello mistura

Full conditional of σ_k^2: La stessa del modello mistura

Full conditional of z_0: le uniche componenti che dipendono da z_0 sono

$$\left[\prod_{k=1}^{L}\rho_k^{\mathbf{1}_k(z_0)}\right][\pi_{z_0,z_1}]$$

e quindi

$$P(z_0 = k|\dots) \propto \rho_k\pi_{k,z_1} \Rightarrow P(z_0 = k|\dots) = \frac{\rho_k\pi_{k,z_1}}{\sum_{j=1}^{L}\rho_j\pi_{j,z_1}}$$

HMM - Stima Bayesiana

Full conditional of $\boldsymbol{\pi}_j$: con calcoli simili a quelli del modello mistura, possiamo vedere che la full conditional è proporzionale a

$$\left[\prod_{k=1}^{L} \pi_{j,k}^{n_{j,k}}\right] \left[\prod_{k=1}^{L} \pi_{j,k}^{\alpha_k - 1}\right]$$

e quindi

$$\boldsymbol{\pi}_j | \cdots \sim Dir(\alpha_1 + n_{j,1}, \ldots, \alpha_L + n_{j,L})$$

Full conditional of z_i: come prima chiamo

$$g_k = (\sigma_k^2)^{-\frac{1}{2}} \exp\left(\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\right)$$

e quindi ho che la full conditional è proporzionale a

$$\left[\prod_{k=1}^{L} g_k^{\mathbf{1}_k(z_i)}\right] \left[\prod_{k=1}^{L} \pi_{z_{i-1},k}^{\mathbf{1}_k(z_i)} \pi_{k,z_{i+1}}^{\mathbf{1}_k(z_i)}\right]$$

e quindi

$$P(z_i = k | \ldots) = \frac{g_k \pi_{z_{i-1},k} \pi_{k,z_{i+1}}}{\sum_{j=1}^{L} g_j \pi_{z_{i-1},j} \pi_{j,z_{i+1}}}$$

HMM - Marginalizzazione

Come per il modello mistura, possiamo marginalizzare. Potremmo, in teoria, farlo anche per $\mathbf{z}$, ma sappiamo farlo solo in maniera algoritmica (Forward Backward) e non sappiamo realmente come è scritta la verosimiglianza. Quindi è inutile per la creazione dell'MCMC.

Il caso più interessante è la marginalizzazione di tutti i $\boldsymbol{\pi}_j$. In questo caso dobbiamo calcolare solo

$$f(\mathbf{z}) = \int_{S^k} f(\mathbf{z}|\boldsymbol{\pi}) \prod_{j=1}^{L} f(\boldsymbol{\pi}_j) d\boldsymbol{\pi}_1 \dots d\boldsymbol{\pi}_L = \prod_{j=1}^{L} \int_S \prod_{k=1}^{L} \pi_{j,k}^{n_{j,k}+\alpha_k-1} d\boldsymbol{\pi}_j$$

Abbiamo già visto che quello dentro l'integrale è il kernel di una Dirichlet, allora

$$f(\mathbf{z}) = \prod_{j=1}^{L} \frac{\prod_{k=1}^{L} \Gamma(\alpha_k + n_{j,k})}{\Gamma(\sum_{k=1}^{L} (\alpha_k + n_{j,k}))}$$

Utilizzando ancora una volta la proprietà della gamma

$$\Gamma(x+1) = x\Gamma(x)$$

HMM - Marginalizzazione

Assumiamo che $z_{i-1} = k_{i-1}$ e $z_{i+1} = k_{i+1}$, con $n_{j,k}^{-1}$ il numero di transizione da j a k senza considerare quelle che coinvolgono z_i e con

$$n_k^{-i} = \sum_{j=1}^{J} n_{j,k}^{-i}$$

che indica quante volte siamo in stato k senza contare cosa è successo al tempo i. Allora per l'ultima variabile abbiamo che

$$f(z_n = k | \mathbf{z}_{-n}) \propto (n_{k_{i-1},k}^{-n} + \alpha_k) g_k$$

Per tutte le altre dobbiamo fare due casi, questo perchè se

- $z_{i-1} = z_i = z_{i+1} = k$ allora

$$\begin{cases} n_{k,k}^{-i} & = n_{k,k} - 2 \\ n_{j,j'}^{-i} & = n_{j,j'} \text{ if } (j,j') \neq (k,k) \end{cases}$$

- altrimenti, con $z_i = k$,

$$\begin{cases} n^{-i}_{k_{i-1},k} & = n_{k_{i-1},k} - 1 \\ n^{-i}_{k,k_{i+1}} & = n_{k,k_{i+1}} - 1 \\ n^{-i}_{j,j'} & = n_{j,j'} \text{ negli altri casi} \end{cases}$$

Calcoliamo le full conditional

Caso 1: $k_{i-1} \neq k_{i+1}$. In questo caso abbiamo che

$$f(z_i = k|\mathbf{z}_{-i}) \propto \frac{(n^{-i}_{k_{i-1},k} + \alpha_k)(n^{-i}_{k,k_{i+1}} + \alpha_{k_{i+1}})}{\sum_{k=1}^{L}(\alpha_k + n^{-i}_{k_{i-1},k})} g_k$$

Caso 2: $k_{i-1} = k_{i+1}$.

- Se $k \neq k_{i-1}$ allora

$$f(z_i = k|\mathbf{z}_{-i}) \propto \frac{(n^{-i}_{k_{i-1},k} + \alpha_k)(n^{-i}_{k,k_{i+1}} + \alpha_{k_{i+1}})}{\sum_{k=1}^{L}(\alpha_k + n^{-i}_{k_{i-1},k})} g_k$$

HMM - Marginalizzazione

- Se $k = k_{i-1}$ allora

$$f(z_i = k|\mathbf{z}_{-i}) \propto \frac{(n_{k,k}^{-i} + 1 + \alpha_k)(n_{k,k}^{-i} + \alpha_{k_{i+1}})}{\sum_{k=1}^{L}(\alpha_k + n_{k_{i-1},k}^{-i})} g_k$$

HMM - Algoritmo di Viterbi

invece di simulare uno stato alla volta nell'MCMC, potremmo campionare direttamente tutto il vettore. Questo si può fare usando una versione dell'algoritmo di **Viterbi**.
Lo scopo finale è simulare da $f(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P})$ ($\boldsymbol{\theta}$ sono tutti i parametri del modello) simulando da (fate attenzione che parte da n e arriva a 1)

- z_n da $f(z_n|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P})$
- z_{n-1} da $f(z_{n-1}|z_n, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P})$
- z_{n-2} da $f(z_{n-2}|z_n, z_{n-1}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P})$
- $\dots$
- z_1 da $f(z_1|z_n, z_{n-1}, \dots, z_2, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P})$

HMM - Algoritmo di Viterbi

Se riprendiamo l'algoritmo forward-backward, abbiamo che per l'ultima osservazione

$$f(z_n = k|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P}) \propto f(z_n = k, \mathbf{y}|\boldsymbol{\theta}, \boldsymbol{P}) = a_n(k)$$

e quindi possiamo simulare z_n. Fate attenzione che il forward-backward ha assunto che z_0 fosse nota (è dentro $\boldsymbol{\theta}$), ma possiamo estenderlo e marginalizzare su z_0 se definiamo

$$a_1(k) = f(y_1, z_1 = k) = \sum_{j=1}^{L} f(y_1, z_1 = k|z_0 = j) f(z_0 = j) = \sum_{j=1}^{L} \rho_j \pi_{j,k} f(\mathbf{y}|z_1 = k)$$

Per simulare le altre, definiamo

$$\mathbf{z}^{i+1} = (z_{i+1}, z_{i+2}, \ldots z_n)$$

e

$$\mathbf{y}^{i+1} = (y_{i+1}, y_{i+2}, \ldots y_n)$$

e notiamo che

$$f(z_i|\mathbf{z}^{i+1}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P}) \propto$$

$$f(z_i, \mathbf{z}^{i+1}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P}) = f(\mathbf{y}|, \mathbf{z}^{i+1}, z_i, \boldsymbol{\theta}, \boldsymbol{P}) f(\mathbf{z}^{i+1}|z_i, \boldsymbol{\theta}, \boldsymbol{P}) f(z_i|\boldsymbol{\theta}, \boldsymbol{P})$$

HMM - Algoritmo di Viterbi

ma abbiamo che (guardate il DAG)

$$f(\mathbf{y}|\mathbf{z}^{i+1}, z_i, \boldsymbol{\theta}, \boldsymbol{P}) = f(y_1, \dots, y_i|z_i, \boldsymbol{\theta}, \boldsymbol{P}) f(\mathbf{y}^{i+1}|\mathbf{z}^{i+1}, \boldsymbol{\theta}, \boldsymbol{P}) \overset{z_i}{\propto} f(y_1, \dots, y_i|z_i, \boldsymbol{\theta}, \boldsymbol{P})$$

e

$$f(\mathbf{z}^{i+1}|z_i, \boldsymbol{\theta}, \boldsymbol{P}) = f(z_n|z_{n-1}, \boldsymbol{\theta}, \boldsymbol{P}) \dots f(z_{i+2}|z_{i+1}, \boldsymbol{\theta}, \boldsymbol{P}) f(z_{i+1}|z_i, \boldsymbol{\theta}, \boldsymbol{P}) \overset{z_i}{\propto} f(z_{i+1}|z_i, \boldsymbol{\theta}, \boldsymbol{P})$$

e quindi abbiamo che

$$f(z_i|\mathbf{z}^{i+1}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P}) \propto f(y_1, \dots, y_i|z_i, \boldsymbol{\theta}, \boldsymbol{P}) f(z_{i+1}|z_i, \boldsymbol{\theta}, \boldsymbol{P}) f(z_i|\boldsymbol{\theta}, \boldsymbol{P}) =$$

$$f(y_1, \dots, y_i, z_i|\boldsymbol{\theta}, \boldsymbol{P}) f(z_{i+1}|z_i, \boldsymbol{\theta}, \boldsymbol{P}) = \alpha_i(z_i) \pi_{z_i, z_{i+1}}$$

e quindi

$$P(z_i = k|\mathbf{z}^{i+1}, \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{P}) \propto \alpha_i(k) \pi_{k, z_{i+1}}$$

HMM - Algoritmo di Viterbi

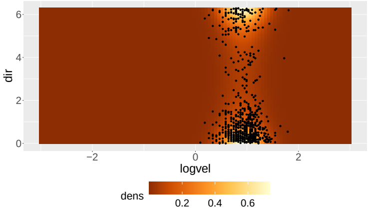
Prendiamo l'esempio del vento e proviamo un HMM.

$$
\begin{aligned}
\log x_i | z_i, \{\mu_k, \sigma_k^2\}_{k=1}^{L} &\sim N(\mu_{z_i}, \sigma_{z_i}^2) \\
\eta_i | z_i, \{\rho_k, \tau_k\}_{k=1}^{L} &\sim WC(\rho_{z_i}, \tau_{z_i}) \\
z_i | z_{i-1} &\sim Disc(\boldsymbol{\pi}_{z_{i-1}}), i \geq 1 \\
z_0 &\sim Disc(1/L) \\
\boldsymbol{\pi}_j &\sim Dir(\mathbf{1}) \\
\mu_k &\sim N(0, 10000^2) \\
\sigma_k^2 &\sim IG(1, 1) \\
\tau_k &\sim G(1, 1) \\
\rho_k &\sim U(0, 2\pi)
\end{aligned}
$$

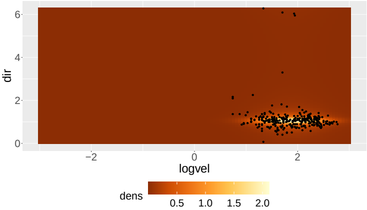
Stimo solo il modello con 4 stati, che era il numero di stati migliore del modello mistura con WAIC 9180.096. L'HMM ha WAIC 8344.04.

HMM - Algoritmo di Viterbi

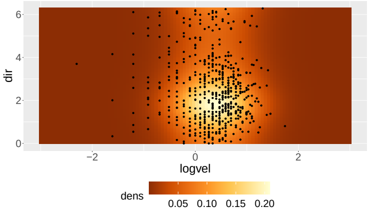
In termini di densità predittive non ci sono enormi differenze, ma alcune sono interessanti



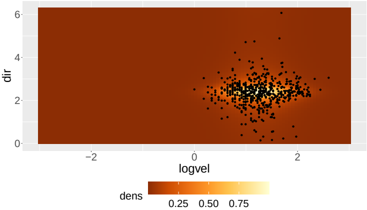
(e) Cluster 1



(f) Cluster 2

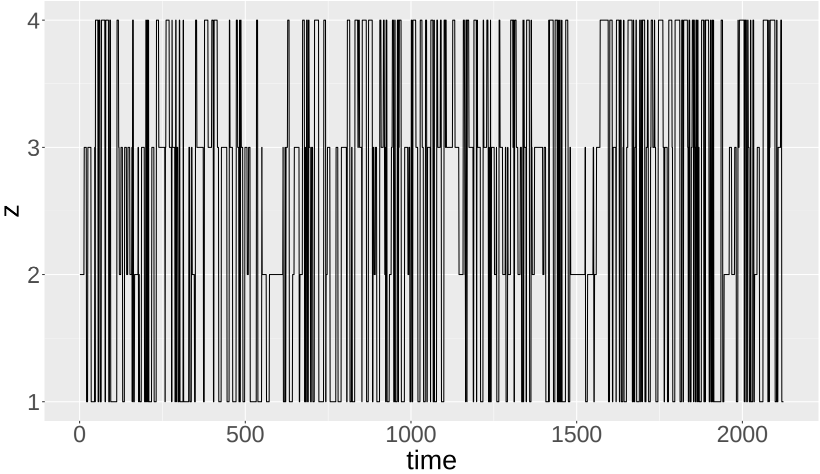


(g) Cluster 3



(h) Cluster 4

Possiamo vedere la serie delle stime di z



e la matrice di transizione

	k = 1	k = 2	k = 3	k = 4
j=1	0.762	0.020	0.073	0.145
	(0.714 0.804)	(0.008 0.036)	(0.044 0.109)	(0.107 0.185)
j=2	0.028	0.878	0.081	0.014
	(0.004 0.062)	(0.835 0.913)	(0.045 0.124)	(0.001 0.036)
j=3	0.116	0.034	0.790	0.060
	(0.075 0.160)	(0.019 0.054)	(0.740 0.837)	(0.029 0.098)
j=4	0.132	0.004	0.087	0.777
	(0.095 0.173)	(0.000 0.013)	(0.055 0.124)	(0.731 0.819)

Change Point Model

Change-Point model

Ci sono dei casi in cui la serie segue diversi regimi, ma una volta che cambia regime, non torna più a regimi visitati nel passato. Questi modelli si chiamano **Change-Point** o **Structural-change** model.

Ci sono vari modi per definirli, ma il più semplice è vederli come un HMM con una particolare matrice di transizione, del tipo.

$$y_i | z_i, \{\boldsymbol{\theta}_k\}_{k=1}^{L} \sim F(\boldsymbol{\theta}_{z_i})$$
$$z_i | z_{i-1} \sim Discrete(\boldsymbol{\pi}_{z_{i-1}}), i = 1, \dots, n$$

dove $\boldsymbol{\pi}_j$ con $j < L$ ha solo due valori diversi da zero, che sono

$$\pi_{j,j} > 0 \qquad \pi_{j,j+1} > 0$$

con

$$\pi_{j,j+1} = 1 - \pi_{j,j}$$

mentre $\boldsymbol{\pi}_k = (0, 0, 0 \dots, 0, 0, 1)$. Quindi il modello parte da $z_1 = 1$, che è una costante, e poi, ad ogni tempo, o rimane nello stesso stato o si sposta in quello adiacente.

Dal punto di vista Bayesiano si può mettere una serie di distribuzioni beta su $\pi_{j,j}$.