Problemi Inversi e Applicazioni - Parte II

1. Variabili casuali

Def. (σ-algebra): Sia Ω un insieme. Una σ-algebra degli insiemi di Ω è un insieme $\mathcal{A}$ di sottoinsiemi di Ω tale che

- $\Omega \in \mathcal{A}$;
- se $A \in \mathcal{A}$, allora il suo complementare rispetto a Ω é in $\mathcal{A}$;
- ogni unione numerabile di insiemi di $\mathcal{A}$ è in $\mathcal{A}$.

Def. (σ-algebra di Borel): dato $\mathbb{R}^n$, la σ-algebra di Borel $\mathcal{B}$ è la più piccola σ-algebra degli insiemi di $\mathbb{R}^n$ contenente tutti gli aperti di $\mathbb{R}^n$. Gli elementi di questa σ-algebra vengono detti insiemi di Borel.

Def. (misura di probabilità): la mappa $P : \mathcal{A} \to \mathbb{R}$ è detta misura di probabilità se:

- $0 \leq P(A) \leq 1$ per $A \in \mathcal{A}$
- $P(0) = 0$, $P(\Omega) = 1$
- $P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$ dove gli A_k sono disgiunti.

Def. (spazio di probabilità): la tripla $(\Omega, \mathcal{A}, P)$ è detta spazio di probabilità.

Def. (variabile casuale): è dato lo spazio di probabilità $(\Omega, \mathcal{A}, P)$. Una variabile casuale è una mappa $X : \Omega \to \mathbb{R}^n$ misurabile, ovvero tale che $X^{-1}(B) \in \mathcal{A}$ per ogni aperto $B \subset \mathbb{R}^n$.

Def. (distribuzione di probabilità): una variabile casuale genera una misura di probabilità in modo piuttosto naturale. Considero infatti la misura di probabilità $\mu : \mathcal{B} \to \mathbb{R}$ tale che

$$\mu(B) = P(X^{-1}(B)). \tag{1.1}$$

con B un insieme di Borel. Questa è una buona definizione. Infatti: B è un aperto di $\mathbb{R}^n$ e X è misurabile rispetto alla σ-algebra $\mathcal{A}$. Quindi $X^{-1}(B)$ appartiene ad $\mathcal{A}$. P è la misura di probabilità definita su $\mathcal{A}$ e quindi $P(X^{-1}(B))$ è ben definito. La misura di probabilità μ è detta distribuzione di probabilità di X.

Def. (distribuzione di probabilità congiunta): sono date le due variabili casuali $X_1 : \Omega \to R^n$ e $X_2 : \Omega \to R^m$. La distribuzione di probabilità congiunta è la distribuzione di probabilità di $X_1 \times X_2 : \Omega \to \mathbb{R}^n \times \mathbb{R}^m$ tale che $\omega \to (X_1(\omega), X_2(\omega))$.

Def. (densità): data la variabile casuale X, la sua funzione di distribuzione è definita come

$$F(x) := \mu(-\infty, x] = P[X \leq x] = P[X_1 < x_1, \ldots, X_n < x_n], \tag{1.2}$$

dove con x indico una realizzazione di X. Una variabile casuale ha densità π rispetto alla misura di Lebesgue se π è una funzione non negativa su $\mathbb{R}$, assolutamente continua rispetto alla misura di Lebesgue, tale che

$$P[X \in A] = \mu(A) = \int_A \pi(x)dx \quad A \in \mathbb{R} \tag{1.3}$$

dove $P[X \in A]$ indica la probabilità che X assuma valori nell'intervallo A di $\mathbb{R}$.
Def. (espettazione): si dice espettazione $E\{X\}$ della variabile casuale X il numero

$$E\{X\} = \int_\Omega X(\omega) dP(\omega) = \int_{\mathbb{R}^n} x d\mu(x). \tag{1.4}$$

Si dice matrice di correlazione di una variabile casuale $X = (X_1, X_2, \ldots, X_n)$ la matrice in $\mathbb{R}^{n\times n}$

$$corr(X) = E\{XX^T\} = \int_\Omega X(\omega)X^T(\omega) dP(\omega) = \int_{\mathbb{R}^n} xx^T d\mu(x). \tag{1.5}$$

Si dice matrice di covarianza della variabile casuale X la matrice $n \times n$

$$cov(X) = corr(X - E\{X\}). \tag{1.6}$$

2. Formulazione Bayesiana di un problema inverso

Oss. (approccio Bayesiano): l'approccio Bayesiano ai problemi inversi non tenta di rispondere alla domanda: 'quale è la migliore approssimazione dell'incognita?' ma alla domanda: 'quale è il nostro grado di conoscenza dell'incognita?'
Oss. (formulazione generale): l'approccio Bayesiano si basa su due ipotesi generali:

(i) tutte le variabili coinvolte nel problema sono variabili casuali;

(ii) le densità di probabilità codificano il grado di conoscenza di queste variabili.

La formulazione Bayesiana di un problema inverso è la seguente: sono date la variabile casuale misura (o dato) Y, la variabile casuale incognita X e l'operatore (in generale, non lineare) A tale che

$$A(X) = Y. \tag{2.7}$$

Siano y e x due generiche realizzazioni di Y e X. Il problema inverso da risolvere consiste nel determinare la densità di probabilità $\pi(x|y)$ per X condizionata sulla realizzazione y del dato, quando

$$A(x) = y. \tag{2.8}$$

Teo. (Teorema di Bayes): Hp.: X e Y sono due variabili casuali a valori in $\mathbb{R}^n$ e $\mathbb{R}^m$ rispettivamente, legate da

$$A(X) = Y. \tag{2.9}$$

Considero le seguenti funzioni:

$$\pi_{pr}(x) := \int_{\mathbb{R}^m} \pi(x, y) dy; \tag{2.10}$$

$$\pi(y|x) := \frac{\pi(x, y)}{\pi_{pr}(x)}; \tag{2.11}$$

$$\pi(y) := \int_{\mathbb{R}^n} \pi(x,y)dx; \tag{2.12}$$

$$\pi(x|y) := \frac{\pi(x,y)}{\pi(y)}. \tag{2.13}$$

Sia $\pi(y) \neq 0$. Th:

$$\pi(x|y) = \frac{\pi_{pr}\pi(y,x)}{\pi(y)}. \tag{2.14}$$

Dim.: dalla (2.11)

$$\pi(x,y) = \pi(y|x)\pi_{pr}(x). \tag{2.15}$$

Sostituendo la (2.15) nella (2.13) si ha la tesi.

Oss. (interpretazione): la formula di Bayes e' un risultato molto semplice. Ciò che è realmente interessante è l'interpretazione delle densità di probabilità che entrano in gioco. In particolare:

- la prior $\pi_{pr}(x) = \int_{\mathbb{R}^m} \pi(x,y)dy$ contiene tutte le informazioni sull'incognita note a priori, ovvero prima che la misura sia disponibile. Può trattarsi di vincoli fisici o matematici, oppure, nel caso di problemi di imaging medico, di informazione fornita da altre modalità di acquisizione. Nel caso in cui nessuna informazione sia disponibile, sarà oppurtuno scegliere una prior poco informativa, come una distribuzione uniforme o una Gaussiana con deviazione standard molto grande (rispetto alle dimensioni lineari in gioco nel problema);
- la likelihood $\pi(y|x) := \frac{\pi(x,y)}{\pi_{pr}(x)}$ è la densità per il dato, condizionata su uno specifico valore x dell'incognita. La likelihood contiene due modelli: il problema diretto $\overline{y} = A(x)$, dove $\overline{y}$ è il dato senza rumore, e il modello del rumore $y = y(\overline{y}, w)$, dove w è la realizzazione della variabile casuale rumore W. Ad esempio, se il rumore è additivo e W ha densità $\pi_{noise}(w)$, allora $y = A(x) + w$ e la likelihood assume forma $\pi(y|x) = \pi_{noise}(y - A(x))$;
- la posterior $\pi(x|y) := \frac{\pi(x,y)}{\pi(y)}$ è la soluzione del problema inverso Bayesiano e contiene tutte le informazioni che si stanno cercando sull'incognita.
- la densità $\pi(y) := \int_{\mathbb{R}^n} \pi(x,y)dx$ al denominatore della (2.14) è un semplice fattore di normalizzazione.

3. Stime

Oss. (soluzione del problema): sul piano formale, la soluzione del problema inverso Bayesiano è la posterior. Però, sul piano pratico, $\pi(x|y)$ è tipicamente una funzione definita su uno spazio di grande dimensione e quindi è di difficile interpretazione. È quindi assai più pratico fornire soluzione puntuali del problema, attraverso formule che

coinvolgano la posterior.

Def. (conditional mean o cm):

$$x_{cm} = \int_{\mathbb{R}^n} x\pi(x|y). \tag{3.16}$$

Il calcolo del conditional mean è un problema di integrazione numerica. Può comportare difficoltà di tipo numerico (soprattutto nel caso di n grande) ma è sempre ben definito, indipendentemente dalla forma della posterior.

Def. (maximum a posteriori o map):

$$x_{map} = arg \max_{x\in\mathbb{R}^n} \pi(x|y). \tag{3.17}$$

Il calcolo del map è un problema di ottimizzazione e può essere mal posto, nel senso che la soluzione del problema di massimo può non esistere o non essere unica. Il calcolo viene tipicamente effettuato applicando algoritmi iterativi tipici della teoria dell'ottimizzazione numerica.

4. Metodo di Tikhonov

Def. (regolarizzazione): considero l'operatore lineare compatto $A : X \to Y$, con X e Y spazi di Hilbert, e il problema inverso mal posto

$$g = Af. \tag{4.18}$$

Assumo che il nucleo di A sia banale (non è un'ipotesi troppo restrittiva, in quanto, se non valesse, ciò che segue può essere immediatamente applicato all'approssimazione dell'operatore inverso generalizzato). Si definisce algoritmo di regolarizzazione la famiglia a un parametro $\{R_\alpha\}_{\alpha>0}$ tale che

- $R_\alpha : Y \to X$ è limitato per ogni $\alpha \in (0, \infty)$
- $\lim_{\alpha\to 0} \|R_\alpha Af - f\| = 0$.

Questa definizione dice che gli operatori R_α sono approssimazioni limitate dell'operatore inverso A^{-1} (che non è limitato). La cosa interessante, tuttavia, è che nelle applicazioni con dati reali (e quindi rumorosi) la scelta ottimale del parametro α non è zero. Infatti, si g_δ una misura di g caratterizzata da $\|g_\delta - g\| \leq \delta$. La versione rumorosa del problema (4.18) è

$$g_\delta = Af + w_\delta, \tag{4.19}$$

dove w_δ è il rumore tale che $\|w_\delta\| \leq \delta$. Applicando R_α ai due membri della (4.19) e usando la disuguaglianza triangolare si ottiene

$$\|R_\alpha g_\delta - f\|_X \leq \|R_\alpha Af - f\| + \delta\|R_\alpha\|. \tag{4.20}$$

Il termine al primo membro e' l'errore di ricostruzione quando si usa l'algoritmo di regolarizzazione. A destra, il primo termine va a zero per $\alpha \to 0$ mentre il secondo

termine diverge. La (4.20) suggerisce che il valore ottimale di α, ovvero quello che minimizza il primo membro, dipende dal livello di noise, ovvero $\alpha = \alpha(\delta)$. Il problema cruciale nella regolarizzazione è quello di fornire una ricetta (un altro algoritmo) per la scelta ottimale della legge $\alpha = \alpha(\delta)$.

Def. (metodo di Tikhonov): il metodo di Tikhonov consiste nel risolvere il problema di minimo

$$\|Af - g\|^2 + \alpha\|f\|^2 = \min . \tag{4.21}$$

Si dimostra che risolvere questo problema di minimo corrisponde a risolvere l'equazione di Eulero

$$(A^\star A + \alpha I)f = A^\star g \tag{4.22}$$

e che la famiglia a un parametro

$$R_\alpha = (A^\star A + \alpha I)^{-1} A^\star \tag{4.23}$$

è un algoritmo di regolarizzazione.

Oss. (Bayes e regolarizzazione): applicando l'operatore $-\log$ ai due membri del teorema di Bayes e trascurando il termine che dipende dalla sola y, si ottiene:

$$-\log \pi(x|y) = -\log \pi(y|x) - \log \pi_{pr}(x). \tag{4.24}$$

Questa equazione ha un'interpretazione naturale in termini di regolarizzazione: il primo termine a destra è piccolo quando la soluzione fitta bene il dato, mentre il secondo termine a destra può essere visto come un termine di stabilità o, in teoria dell'informazione, di bassa complessità della soluzione.

Oss. (Bayes e Tikhonov): considero l'equazione (4.19) e assumo che:

- w sia la realizzazione di una variabile casuale noise W Gaussiana e additiva, ovvero
$$\pi_{noise}(w) = \exp(-\frac{\|w\|^2}{\sigma_w^2}). \tag{4.25}$$
- A sia lineare;
- la prior sia Gaussiana, ovvero
$$\pi_{pr}(x) = \exp(-\frac{\|x\|^2}{\sigma_x^2}). \tag{4.26}$$

La likelihood ha forma

$$\pi(y|x) = \pi_{noise}(y - Ax) = \exp(-\frac{\|y - Ax\|^2}{\sigma_w^2}). \tag{4.27}$$

Usando il teorema di Bayes (e trascurando il denominatore) si ha, per la posterior:

$$\pi(x|y) = \exp(-\left(\frac{\|y - Ax\|^2}{\sigma_w^2} + \frac{\|x\|^2}{\sigma_x^2}\right). \tag{4.28}$$

Usando il MAP come stima, si deve risolvere il problema di minimo

$$\|y - Ax\|^2 + \frac{\sigma_w^2}{\sigma_x^2}\|x\|^2 = \min, \tag{4.29}$$

che coincide con il problema di Tikhonov. Nella (4.29) è implicita una stima ottimale del parametro di regolarizzazione, nei termini delle varianze delle due Gaussiane in gioco.

5. Expectation - Maximization (EM)

Def. (distribuzione di Poisson): la distribuzione di Poisson è associata a una variabile discreta che esprime un numero di eventi. Se N è la variabile casuale che descrive il numero di eventi che si verifica in un intervallo di tempo dato, la distribuzione di probabilità di Poisson associata a N è

$$P(n) = e^{-\lambda}\frac{\lambda^n}{n!} \ , \tag{5.30}$$

dove λ è il valor medio di N. Ora consideriamo il problema inverso Bayesiano

$$Y = AX \tag{5.31}$$

dove si assume che Y sia una variabile di tipo Poisson. Se si assume che Ax è una stima affidabile del valor medio di Y, allora la likelihood associata al problema inverso (5.31) è data da

$$\pi(y|x) = e^{-Ax}\frac{(Ax)^y}{y!} \ . \tag{5.32}$$

Suppongo in questa section che X e Y siano due vettori casuali di dimensioni N e A sia una matrice rettangolare $N \times N$. Allora si ha che la likelihood ha forma

$$\pi(y|x) = \prod_{i=1}^{N} e^{-(Ax)_i}\frac{(Ax)_i^{y_i}}{y_i!} \ . \tag{5.33}$$

Def. (Kullbach-Leibler): Si ottiene immediatamente che

$$-\log \pi(y|x) = -\sum_{i=1}^{N}[-(Ax)_i + y_i \log(Ax)_i - \log y_i!] \ . \tag{5.34}$$

Eliminando dalla (5.34) tutto ciò che non dipende da x si ottiene

$$KL(x) = \sum_{i=1}^{N}[(Ax)_i - y_i \log(Ax)_i] \ . \tag{5.35}$$

$KL(x)$ ѐdetta funzione (o metrica, o topologia) di Kullbach-Leibler.
Oss, (maximum-likelihood vincolata): definisco il cono positivo

$$C = \{x \in \mathbb{R}^N, \quad x_i \geq 0 \quad i = 1, \ldots, N\} \ . \tag{5.36}$$

Il problema di EM consiste nel trovare in modo computazionalmente efficiente la soluzione del problema di maximum-likelihood vincolato

$$x_{KL} = \arg\min_{x \in C} KL(x) \ . \tag{5.37}$$

Teo. (KKT): Il teorema di Karush-Kuhn-Tucker dimostra che condizione necessaria e sufficiente affinchè x sia soluzione del problema (5.37) è che valgano le condizioni (condizioni KKT):

$$\begin{cases} x_i \geq 0 \\ \lambda_i \geq 0 \\ \nabla_{x_i} KL(x) - \lambda_i = 0 \\ \lambda_i x_i = 0 \end{cases} \tag{5.38}$$

dove $i = 1, \ldots, N$ e $\lambda = (\lambda_1, \ldots, \lambda_N)$ è un moltiplicatore di Lagrange.
Oss. (gradiente di KL): calcolo la derivata parziale di $KL(x)$:

$$\frac{\partial}{\partial x_k} KL(x) = \sum_{i=1}^{N} [\frac{\partial}{\partial x_k} \sum_{j=1}^{N} A_{ij} x_j - \frac{y_i}{(Ax)_i} \frac{\partial}{\partial x_k} \sum_{j=1}^{N} A_{ij} x_j] \ . \tag{5.39}$$

Ora,

$$\frac{\partial}{\partial x_k} \sum_{j=1}^{N} A_{ij} x_j = \sum_{j=1}^{N} A_{ij} \delta_{kj} = A_{ik} \ . \tag{5.40}$$

Quindi

$$\frac{\partial}{\partial x_k} KL(x) = \sum_{i=1}^{N} \{A_{ik} - \frac{y_i}{(Ax)_i} A_{ik}\} = \sum_{i=1}^{N} A^T_{ki} \mathbf{1}_i - \sum_{i=1}^{N} A^T_{ki} \frac{y_i}{(Ax)_i} \ , \tag{5.41}$$

dove $\mathbf{1}$ è il vettore colonna fatto con N componenti tutte uguali a 1. Ne consegue che il gradiente di $KL(x)$ è il vettore

$$\nabla_x KL(x) = A^T \left(\mathbf{1} - \frac{y}{Ax} \right) \ , \tag{5.42}$$

dove $\frac{y}{Ax}$ è il vettore di componenti uguali al rapporto tra le componenti di y e di Ax.
Def. (metodo EM): la terza e la quarta condizione KKT implicano

$$x_i (A^T \mathbf{1})_i = x_i \left[A^T \left(\frac{y}{Ax} \right) \right]_i \tag{5.43}$$

per $i = 1, \ldots, N$, da cui

$$x_i = x_i \frac{\left[A^T \left(\frac{y}{Ax} \right) \right]_i}{(A^T \mathbf{1})_i} \ . \tag{5.44}$$

L'equazione (5.44) può essere scritta come l'equazione di punto fisso

$$x = x \frac{A^T \left(\frac{y}{Ax} \right)}{A^T \mathbf{1}} \ , \tag{5.45}$$

dove a destra si ha un vettore le cui componenti sono il prodotto tra le componenti del vettore x e le componenti del vettore $A^T(\frac{y}{Ax})/A^T\mathbf{1}$. Il metodo EM è il metodo iterativo che risolve l'equazione di punto fisso (5.45) attraverso lo schema delle approssimazioni successive

$$x^{k+1} = x^k \frac{A^T\left(\frac{y}{Ax^k}\right)}{A^T\mathbf{1}} \ , \tag{5.46}$$

con inizializzazione

$$x^0 = \mathbf{1} \ . \tag{5.47}$$

6. Filtraggio Bayesiano

Def. (processo stocastico): un processo stocastico è una famiglia ad un parametro $\{X_k\}_{k=1}^{\infty}$ di variabili casuali a valori su $\mathbb{R}^{n_k}$.
Def. (modello evoluzione-osservazione): siano $\{X_k\}$ e $\{Y_k\}$ due processi stocastici a valori in $\mathbb{R}^n$ e $\mathbb{R}^m$ rispettivamente. Gli elementi del processo $\{X_k\}$ vengono denominati vettori di stato mentre gli elementi del processo $\{Y_k\}$ vengono denominati osservazioni. Assumo che valgano le seguenti ipotesi sulle densità di probabilità:

(i) il processo $\{X_k\}$ è un processo di Markov, ovvero

$$\pi(x_{k+1}|x_0, x_1, \ldots, x_k) = \pi(x_{k+1}|x_k); \tag{6.48}$$

(ii) il processo $\{Y_k\}$ è un processo di Markov rispetto alla storia di $\{X_k\}$, ovvero

$$\pi(y_k|x_0, x_1, \ldots, x_k) = \pi(y_k|x_k); \tag{6.49}$$

(iii) il processo $\{X_k\}$ dipende dalle osservazioni passate soltanto attraverso la propria storia, ovvero

$$\pi(x_{k+1}|x_k, y_1, y_2, \ldots, y_k) = \pi(x_{k+1}|x_k); \tag{6.50}$$

(iv) le osservazioni sono indipendenti dalle osservazioni passate, ovvero

$$\pi(y_{k+1}|x_{k+1}, y_1, \ldots, y_k) = \pi(y_{k+1}|x_{k+1}). \tag{6.51}$$

Allora i due processi stocastici vengono detti modello evoluzione-osservazione.
Oss. (interpretazione delle ipotesi di Markov in MEG): se considero il problema di ricostruzione neurale con la MEG, le ipotesi (6.49), (6.50) e (6.51) sono assunzioni ragionevoli. Infatti, considerando la propagazione quasi istantanea del campo magnetico, è ragionevole assumere che la misura al tempo t dipende soltanto dalla corrente al tempo t (equazione (6.49)); è altrettanto ragionevole assumere che la corrente

non dipenda dal campo (equazione (6.50)) e che il campo all'istante t dipenda solo dalla corrente all'istante t, e non dai valori precedenti del campo. La condizione (6.48) è sicuramente più problematica in quanto assume che l'evoluzione della corrente è sostanzialmente un random-walk, il che non è certamente vero.

Def. (filtraggio Bayesiano): è dato il modello evoluzione-osservazione $\{X_k\}$, $\{Y_k\}$. Il problema di determinare la densità a posteriori $\pi(x_k|y_1,\ldots,y_k)$ è detto problema del filtraggio Bayesiano e un algoritmo che realizza tale filtraggio è detto filtro Bayesiano.

Teo. (equazioni di Kolmogorov e Bayes): Hp.: $\{X_k\}$ e $\{Y_k\}$ sono un modello evoluzione-osservazione. Sia $D_k = \{y_1,\ldots,y_k\}$. Allora, Th.:

$$\pi(x_{k+1}|D_k) = \int \pi(x_{k+1}|x_k)\pi(x_k|D_k)dx_k; \tag{6.52}$$

$$\pi(x_{k+1}|D_{k+1}) = \frac{\pi(y_{k+1}|x_{k+1})\pi(x_{k+1}|D_k)}{\pi(y_{k+1}|D_k)}. \tag{6.53}$$

Dim.: Per la (6.50)

$$\pi(x_{k+1},x_k,D_k) = \pi(x_{k+1}|x_k,D_k)\pi(x_k,D_k) = \tag{6.54}$$

$$= \pi(x_{k+1}|x_k)\pi(x_k|D_k)\pi(D_k). \tag{6.55}$$

Usando la (6.55):

$$\pi(x_{k+1}|D_k) = \frac{\pi(x_{k+1},D_k)}{\pi(D_k)} = \tag{6.56}$$

$$= \frac{1}{\pi(D_k)}\int \pi(x_{k+1},x_k,D_k)dx_k = \int \pi(x_{k+1}|x_k)\pi(x_k|D_k)dx_k, \tag{6.57}$$

e quindi l'equazione di Kolmogorov è dedotta. Ora deduco l'equazione di Bayes.

$$\pi(x_{k+1}|D_{k+1}) = \frac{\pi(x_{k+1},D_{k+1})}{\pi(D_{k+1})} = \frac{\pi(y_{k+1}|x_{k+1},D_k)\pi(x_{k+1},D_k)}{\pi(D_{k+1})} = \tag{6.58}$$

Per la (6.51):

$$= \frac{\pi(y_{k+1}|x_{k+1})\pi(x_{k+1},D_k)}{\pi(D_{k+1})} = \frac{\pi(y_{k+1}|x_{k+1})\pi(x_{k+1}|D_k)\pi(D_k)}{\pi(D_{k+1})} = \tag{6.59}$$

$$= \frac{\pi(y_{k+1}|x_{k+1})\pi(x_{k+1}|D_k)\pi(D_k)}{\pi(y_{k+1},D_k)} = \tag{6.60}$$

$$= \frac{\pi(y_{k+1}|x_{k+1})\pi(x_{k+1}|D_k)\pi(D_k)}{\pi(y_{k+1}|D_k)\pi(D_k)} \tag{6.61}$$

e quindi si ha l'equazione di Bayes.

Oss. (filtro Bayesiano): il precedente teorema indica una ricetta generale per costruire un filtro Bayesiano. È anzitutto necessario avere

- una stima iniziale di $\pi(x_0|y_0)$ che è la posterior al tempo zero, ma, visto che al tempo zero non vi è alcuna misura, può essere vista come la prior al tempo zero;
- un modello per il nucleo di transizione $\pi(x_{k+1}|x_k)$;
- un modello per la likelihood $\pi(y_k|x_k)$, ovvero il problema diretto.

A questo punto è possibile innescare la coppia di equazioni (6.52) e (6.53) e, almeno formalmente, ottenere la posterior a ogni k. Va anche osservato che in questa procedura non si fa alcuna ipotesi di stazionarietà sul nucleo di transizione o sulla likelihood per cui è ammesso che queste due densità varino nel tempo.

Oss. (scelta della likelihood): in termini generali, cioè tenendo conto del noise sulla misura, il modello che descrive il problema inverso è

$$Y = f(X, E). \tag{6.62}$$

Possiamo avere vari casi particolari:

- noise additivo con X ed E variabili casuali indipendenti (ovvero se X assume il valore x, la pdf di E rimane inalterata). Allora Y condizionata sul valore x di X è distribuita come E e quindi
$$\pi(y|x) = \pi_{noise}(y - f(x)); \tag{6.63}$$
- noise additivo con X ed E dipendenti. In questo caso la pdf di E dipende dal valore assunto da X. Allora
$$\pi(y|x) = \int \pi(y|x, e)\pi_{noise}(e|x)de. \tag{6.64}$$
D'altra parte, l'equazione (6.62) dice che fissato $X = x$ ho probabilità uno che $Y = f(x) + e$ e zero per ogni altro caso, ovvero
$$\pi(y|x, e) = \delta(y - f(x) - e) \tag{6.65}$$
e quindi
$$\pi(y|x) = \pi_{noise}(y - f(x)|x); \tag{6.66}$$
- noise moltiplicativo. Nel caso di amplificatori un tipico modello è
$$Y = Ef(X). \tag{6.67}$$
In questo caso, in (6.64) si ha
$$\pi(y|x, e) = \delta(y - ef(x)) \tag{6.68}$$
per cui
$$\pi(y|x) = \int \delta(y - ef(x))\pi_{noise}(e|x)de = \frac{1}{f(x)}\pi_{noise}(y/f(x)), \tag{6.69}$$
dove, nell'integrale, ho effettuato il cambio di variabile $\nu = ef(x)$;
- blind deconvolution. Suppongo che il modello sia dato da
$$Y = A(V)X + E \tag{6.70}$$
dove V è un'altra variabile casuale con pdf $\pi_{param}(v)$. Se E è indipendente da X e

V allora

$$\pi(y|x) = \int \pi_{noise}(y - A(v)x)\pi_{param}(v)dv. \tag{6.71}$$

Oss. (scelta di $\pi(x_0)$): la scelta più tipica è quella Gaussiana:

$$\pi(x_0) = \exp(-\|x\|^2/\sigma^2). \tag{6.72}$$

Altre due scelte piuttosto tipiche sono la prior L^1

$$\pi(x_0) = \exp(-\alpha \sum_{j=1}^{n} |x_0^{(j)}|) \tag{6.73}$$

e la prior 'maximum entropy'

$$\pi(x_0) = -\sum_{j=1}^{n} \alpha x_0^{(j)} \log\left(\frac{x_0^{(j)}}{a}\right). \tag{6.74}$$

Queste ultime due forme vengono scelte quando si sa a priori che la funzione che si vuole ricostruire è caratterizzata da un supporto molto piccolo.

Oss. (scelta del transition kernel): la scelta più tranquilla è quella di un random walk, per cui

$$\pi(x_{k+1}|x_k) = \mathcal{U}(x_{k+1} - x_k), \tag{6.75}$$

dove $\mathcal{U}(x)$ è ,a distribuzione uniforme. In MEG, informazioni più dettagliate possono essere ottenute da altre modalità di tipo funzionale (per esempio, fMRI o EEG o PET o SPECT).

7. Filtro di Kalman

Oss. (modello evoluzione-osservazione): assumiamo di avere a disposizione un modello di evoluzione che riguarda il processo stocastico $\{X_k\}$ e un modello di osservazione (ovvero un problema diretto) che lega il processo dei dati $\{Y_k\}$ con il processo delle soluzioni $\{X_k\}$. Esplicitamente:

$$X_{k+1} = F(X_k, W_{K+1}) \tag{7.76}$$

$$Y_k = G(X_k, V_k). \tag{7.77}$$

W_{k+1} viene detto rumore di stato ed è una variabile casuale a valori in $\mathbb{R}^n$. Invece V_k viene detto rumore di osservazione ed è una variabile casuale a valori in $\mathbb{R}^n$. Assumo che

$$(W_k, W_l) = 0 \quad k \neq l \tag{7.78}$$

$$(V_k, V_l) = 0 \quad k \neq l \tag{7.79}$$

$$(W_k, V_l) = 0 \quad \forall k \neq l \tag{7.80}$$

$$(W_l, X_0) = 0 \quad \forall l. \tag{7.81}$$

Da queste ipotesi le quattro proprietà di Markov seguono automaticamente e quindi si può dire che $\{X_k\}$, $\{Y_k\}$ formano un modello evoluzione-osservazione.
Def. (filtro di Kalman): considero un modello evoluzione-osservazione

$$X_{k+1} = F_{k+1}X_k + W_{k+1} \quad k = 0, 1, 2, \ldots \tag{7.82}$$

$$Y_k = G_k X_k + V_k \quad k = 1, 2, \ldots \tag{7.83}$$

con F_{k+1} e G_k matrici note. Inoltre: W_{k+1} e V_k sono variabili Gaussiane di media zero; X_0 è una Gaussiana a media zero; i W_k e V_k sono mutualmente indipendenti. Infine $(W_l, X_0) = 0 \quad \forall l$. Il problema di determinare $\pi(x_k|D_k)$ si dice problema di filtraggio di Kalman e la soluzione di tale problema è detta filtro di Kalman.
Oss. (notazioni):

$$N(x_0, \Gamma) := \left(\frac{1}{2\pi|\Gamma|}\right)^{n/2} \exp\left[-\frac{1}{2}(x - x_0)^T\Gamma^{-1}(x - x_0)\right]; \tag{7.84}$$

$$x_{k|l} := E(X_k|D_l) := \int_{\mathbb{R}^n} x_k \pi(x_k|D_l)dx_k; \tag{7.85}$$

$$\Gamma_{k|l} := cov(X_k|D_l) = \tag{7.86}$$

$$= \int_{\mathbb{R}^n} x_k x_k^T \pi(x_k|D_l)dx_k - E(X_k|D_l)E(X_k|D_l)^T. \tag{7.87}$$

Teo. (Kalman): Hp.:

$$X_{k+1} = F_{k+1}X_k + W_{k+1} \quad k = 0, 1, \ldots \tag{7.88}$$

$$Y_k = G_k X_k + V_k \quad k = 1, 2, \ldots \tag{7.89}$$

W_{k+1} ha pdf $N(0, \Gamma_{W_{k+1}})$; V_k ha pdf $N(0, \Gamma_{V_k})$; X_0 ha pdf $N(0, \Gamma_{X_0})$; $W_k \perp W_l \quad k \neq l$; $V_k \perp V_l \quad k \neq l$; i W_k e i V_k sono mutualmente indipendenti e i W_k sono indipendenti da X_0.
Th.: se la posterior all'istante k è Gaussiana, ovvero $\pi(x_k|D_k) = N(x_{k|k}, \Gamma_{k|k})$, allora

$$\pi(x_{k+1}|D_k) = N(x_{k+1|k}, \Gamma_{k+1|k}) \tag{7.90}$$

con $x_{k+1|k} = F_{k+1}x_{k|k}$ e $\Gamma_{k+1|k} = F_{k+1}\Gamma_{k|k}F_{k+1}^T + \Gamma_{w_{k+1}}$.
Se la prior all'istante $k+1$ è Gaussiana, ovvero $\pi(x_{k+1}|D_k) = N(x_{k+1|k}, \Gamma_{k+1|k})$, allora

$$\pi(x_{k+1}|D_{k+1}) = N(x_{k+1|k+1}, \Gamma_{k+1|k+1}) \tag{7.91}$$

con

$$x_{k+1|k+1} = x_{k+1|k} + K_{k+1}(y_{k+1} - G_{k+1}x_{k+1|k}), \tag{7.92}$$

$$\Gamma_{k+1|k+1} = (1 - K_{k+1}G_{k+1})\Gamma_{k+1|k}, \tag{7.93}$$

$$K_{k+1} = \Gamma_{k+1|k}G_{k+1}^T(G_{k+1}\Gamma_{k+1|k}G_{k+1}^T + \Gamma_{V_{k+1}})^{-1}. \tag{7.94}$$

K_{k+1} è detta matrice di guadagno di Kalman.

8. Monte Carlo sampling

Oss. (idea generale): se $\pi(x|y)$ è nota, la stima di X è il conditional mean

$$x_{CM} = \int x\pi(x|y)dx, \tag{8.95}$$

e, in generale, la stima di $f(X)$ è

$$I = \int f(x)\pi(x|y)dx. \tag{8.96}$$

Il problema è che spesso la forma analitica di $\pi(x|y)$ non è nota. L'idea alla base dei metodi Monte Carlo è questa: si cerca un insieme $\{x_i\}_{i=1}^{\alpha}$ di valori assunti da X (e quindi $\{x_i\}_{i=1}^{\alpha}$ è nel dominio di $\pi(x|y)$) e un insieme di pesi w_i tali che

$$\hat{I}_\alpha = \sum_{i=1}^{\alpha} w_i f(x_i) \tag{8.97}$$

approssimi I, ovvero tali che $\lim_{\alpha\to\infty} \hat{I}_\alpha = I$. Ci sono tre modi per costruire i due insiemi $\{x_i\}$ e $\{w_i\}$: (a) il random sampling; (b) l'importance sampling; (c) una catena di Markov.

Oss. (random sampling): campiono il set $\{x_i\}$ usando proprio $\pi(x|y)$. Allora la legge dei grandi numeri mi garantisce che

$$\hat{I}_\alpha = \sum_{i=1}^{\alpha} \frac{1}{\alpha} f(x_i) \tag{8.98}$$

è una buona approssimazione. In termini di approssimazione della pdf invece che dell'integrale, si ha

$$\pi(x|y) \simeq \sum_{i=1}^{\alpha} \frac{1}{\alpha}\delta(x - x_i) \tag{8.99}$$

Lo svantaggio di questo approccio è che non sempre è possibile (o computazionalmente ragionevole) campionare secondo $\pi(x|y)$. Questo perchè $\pi(x|y)$ non è esplicitamente nota o perchè non è gestibile.

Oss. (importance sampling): considero una densità di probabilità $p(x)$ tale che

$supp(p(x)) \subset supp(\pi(x|y))$. Sia $\{x_i\}$ un set di punti estratti con $p(x)$, detta *importance density*. Applico il random sampling a

$$f(x)\frac{\pi(x|y)}{p(x)} \tag{8.100}$$

e ottengo

$$\hat{I}_\alpha = \sum_{i=1}^{\alpha} \frac{1}{\alpha} f(x_i)\frac{\pi(x_i|y)}{p(x_i)}. \tag{8.101}$$

Ancora per la legge dei grandi numeri $\hat{I}_\alpha \to I$. Anche qui posso vedere l'importance sampling come un'approssimazione della distribuzione $\pi(x|y)$, ovvero:

$$\pi(x|y) \simeq \sum_{i=1}^{\alpha} \frac{1}{\alpha}\frac{\pi(x_i|y)}{p(x_i)}\delta(x - x_i). \tag{8.102}$$

L'importance sampling è una buona strategia se l'importance density è una densità facile da utilizzare per fare estrazioni. Una buona scelta per $p(x)$ è che sia un buon compromesso tra la maneggevolezza computazionale e la somiglianza con $\pi(x|y)$. Un altro vantaggio è che anche nel caso in cui $\pi(x|y)$ abbia massimi locali, $p(x)$ non ne risente. D'altra parte, se si sceglie una $p(x)$ inadeguata, si può esplorare male lo spazio degli stati.

Def. (particle filter): considero il modello evoluzione-osservazione con le corrispondenti equazioni di Bayes e di Kolmogorov. Un particle filter è una soluzione approssimata delle equazioni calcolata utilizzando un metodo Monte Carlo.

Def. (Sampling Importance Resampling particle filter): il SIR particle filter è un particle filter realizzato con la seguente procedura: Suppongo che $\pi(x_k|D_k)$ sia noto. Quindi:

(i) applico un random sampling per calcolare $\pi(x_{k+1}|D_k)$ dall'equazione di Kolmogorov. Ovvero: estraggo $\{x_i\}$ con $\pi(x_k|D_k)$. Allora:

$$\pi(x_{k+1}|D_k) \simeq \sum_{i=1}^{\alpha} \frac{1}{\alpha}\pi(x_{k+1}|x_k^i). \tag{8.103}$$

(ii) Applico un importance sampling per calcolare $\pi(x_{k+1}|D_{k+1})$ usando $\pi(x_{k+1}|D_k)$ in (8.103) come importance density. Ovvero: estraggo $\{\tilde{x}_{k+1}^i\}$ con (8.103) e ottengo

$$\pi(x_{k+1}|D_{k+1}) \simeq \sum_{i=1}^{\alpha} w_{k+1}^i \delta(x_{k+1} - \tilde{x}_{k+1}^i) \tag{8.104}$$

con

$$w_{k+1}^i = \frac{1}{\alpha}\frac{\pi(\tilde{x}_{k+1}^i|D_{k+1})}{\pi(\tilde{x}_{k+1}^i|D_k)}. \tag{8.105}$$

Dal teorema di Bayes segue immediatamente

$$\pi(x_{k+1}|D_{k+1}) \simeq \sum_{i=1}^{\alpha} \frac{1}{\alpha C_{k+1}} \pi(y_k|\tilde{x}_k^i)\delta(x_{k+1} - \tilde{x}_{k+1}^i). \tag{8.106}$$

Oss. (pesi): La (8.106) ha un significato fisico immediato: il peso è grande se la likelihood è grande, ovvero vengono privilegiate le estrazioni che si adattano meglio al dato.
Oss. (layered sampling): il layered sampling indica un modo per estrarre i $\{\tilde{x}_{k+1}^i\}$ che consiste nell'estrarre la particella i-esima utilizzando $\pi(x_{k+1}|x_k^i)$.
Oss. (resampling): una volta che la posterior a $k+1$ è stata ottenuta, conviene ricampionare $\{\tilde{x}_{k+1}^i\}$ privilegiando le particelle con peso w_{k+1}^i grande. Ad esempio, posso scegliere Nw_{k+1}^i volte la particella $\tilde{x}_{k+1}^i$.
Oss. (esempio: MEG): il campo magnetico $B(r,t)$ in posizione r esterna alla testa e all'istante t, associato alla corrente neurale $J(r',t)$ in posizione r' sulla corteccia cerebrale all'istante t è dato dall'equazione di Biot-Savart

$$B(r,t) = \frac{\mu_0}{4\pi} \int_V J(r',t) \times \frac{r-r'}{|r-r'|^3} dr' \ . \tag{8.107}$$

Approssimo la testa con una sfera Ω con conducibilità σ costante. Uso per la densità di corrente il modello dipolare

$$J(r') = Q(r_0)\delta(r' - r_0) \ . \tag{8.108}$$

I misuratori del campo magnetico sono posti sulla sfera e misurano la componente radiale del campo magnetico. Si dimostra che il campo magnetico è dato dalla formula di Sarvas

$$B = \frac{\mu_0}{4\pi F^2}(FQ \times r_0 - Q \times r_0 \cdot r\nabla F). \tag{8.109}$$

con

$$F = a(ra + r^2 - r_0 \cdot r), \tag{8.110}$$

e $a = |r - r_0|$ e quindi si ottiene la formula di Sarvas. Da questa formula si deduce che B è lineare rispetto a Q e non-lineare rispetto a r_0.
Oss. (particle filter per MEG): a causa della non-linearità del problema di identificazione di parametri (8.109), il calcolo di Q e r_0 può avvenire utilizzando un filtro a particelle, in cui ciascuna particella è un elemento di $\mathbb{R}^3 \times \mathbb{R}^3$, che contiene le tre coordinate della posizione e le tre coordinate del momento di dipolo.
Oss. (filtro Rao-Blackwell): l'approccio Rao-Blackwell al particle filtering si basa sulla fattorizzazione

$$\pi(j_k|b_{1:k}) = \pi(q_k|r_k, b_{1:k})\pi(r_k|b_{1:k}). \tag{8.111}$$

Si assumono le quattro ipotesi di Markov per la posizione, si assume che il rumore sul dato e $\pi(q_k|r_k, b_{1:k})$ abbiano distribuzione Gaussiana; quindi si può scrivere il seguente

algoritmo: sia $\pi(r_k|b_{1:k})$ noto; sia $\pi(q_k|r_k, b_{1:k})$ noto, funzione Gaussiana di media e covarianza note. Allora:

(i) si estraggono α particelle dalla posterior $\pi(r_k|b_{1:k})$;

(ii) dall'equazione di Kolmogorov e dal random sampling per l'approssimazione di $\pi(r_k|b_{1:k})$:

$$\pi(r_{k+1}|b_{1:k}) \simeq \sum \frac{1}{\alpha}\pi(r_{k+1}|r_k^i); \tag{8.112}$$

(iii) si approssima $\pi(r_{k+1}|b_{1:k+1})$ con un importance sampling avente la (8.112) come importance density:

$$\pi(r_{k+1}|b_{1:k+1}) \simeq \sum \frac{1}{\alpha}\frac{\pi(\tilde{r}_{k+1}^i|b_{1:k+1})}{\pi(\tilde{r}_{k+1}^i|b_{1:k})}\delta(r_{k+1} - \tilde{r}_{k+1}^i); \tag{8.113}$$

(iv) $\pi(q_{k+1}|r_{k+1}, b_{1:k+1})$ è una Gaussiana con media e covarianza fornite dal filtro di Kalman.

Teo. (Rao-Blackwell): Hp.:

$$w(J_k) := \frac{\pi(J_k|b_{1:k})}{\pi(J_k|b_{1:k-1})}; \tag{8.114}$$

$$w(R_k) := \frac{\pi(R_k|b_{1:k})}{\pi(R_k|b_{1:k-1})}; \tag{8.115}$$

Th.:

$$var_{\pi(j_k|b_{1:k-1})}(w(J_k)) \geq var_{\pi(r_k|b_{1:k-1})}(w(R_k)). \tag{8.116}$$

Dim.: usando la definizione di valor medio si ottiene

$$E_{\pi(q_k|r_k,b_{1:k-1})}\{w(J_k)\} = \int \pi(q_k|r_k, b_{1:k-1})w(J_k)dq_k = \tag{8.117}$$

$$= \int \pi(q_k|r_k, b_{1:k-1})\frac{\pi(R_k, q_k|b_{1:k})}{\pi(q_k|r_k, b_{1:k-1})\pi(R_k|b_{1:k-1})}dq_k = w(R_k). \tag{8.118}$$

Utilizzando il precedente risultato e il fatto che la varianza di una variabile casuale è la differenza tra il valor medio del quadrato della variabile e il quadrato del valor medio, si ottiene

$$var_{\pi(j_k|b_{1:k-1})}(w(J_k)) - var_{\pi(r_k|b_{1:k-1})}(w(R_k)) = \tag{8.119}$$

$$= E_{\pi(r_k|b_{1:k-1})}var_{\pi(q_k|r_k,b_{1:k-1})}\{w(J_k)\} \tag{8.120}$$

che è una quantità non negativa.