Monte-Carlo, gestione dell'incertezza e analisi dei rischi

GGC
colabufo@mail.dm.unipi.it
giuseppe.colabufo@polytechnique.edu

2018

Indice

1 Introduzione

Appunti dei corsi di Metodi di Monte-Carlo [2] e Gestione delle incertezze e analisi dei rischi [1]. Ho integrato anche gli appunti del (breve) corso "Bayesian Estimation & Stochastic Optimization" del prof. Vikram Krishnamurthy che ho seguito all'UniMelb a maggio-giugno 2019 (basato principalmente su [3].

Per semplicità saranno in bilingua secondo come mi trovo più comodo a scrivere. Segnalerò con una bandierina le parti in francese e con una bandierina quelle in inglese.

La prima parte contiene nozioni fondamentali di probabilità e statistica, utilizzate nelle prove e nelle osservazioni e che in ogni caso non si può non conoscere. All'inizio di ogni sezione e sottosezione specifico le referenze bibliografiche.

1.1 Motivazioni

Monte-Carlo. L'importanza dei metodi di Monte-Carlo sta nella capacità di esplorare spazi di configurazioni di grosse dimensioni per ricavarne delle informazioni. Tre grandi utilizzi:

1. calcolo di integrali (sfruttando la Legge de Grandi Numeri);
2. simulazione di una distribuzione complessa;
3. ottimizzazione stocastica.

Gestion des incertitudes. Les simulations numériques permettent de limiter le risque et d'éviter le coût d'expériences réelles. Nombreuses sources d'incertitudes existent. La figure 1 décrit les grandes étapes d'une étude d'incertitudes d'un code numérique :

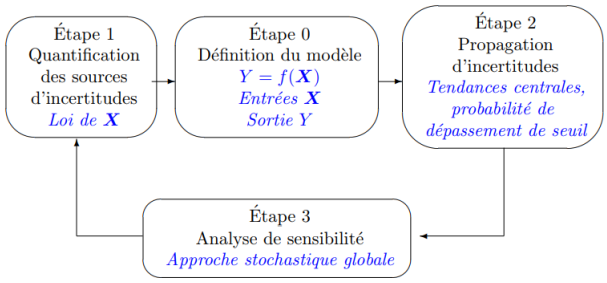

Figura 1: Les étapes d'une étude d'incertitudes

2 Le Basi

Markov.

$$\mathbb{P}\left(X \geq \alpha\right) \leq \frac{\mathbb{E}[X]}{\alpha}$$

Chebyshev.

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \varepsilon\right) \leq \frac{Var(X)}{\varepsilon^2}$$

Chebyshev esponenziale.

$$\mathbb{P}\left(X \geq \varepsilon\right) \leq e^{-t\varepsilon}\mathbb{E}(e^{tX})$$

Disuguaglianza facile.

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \leq \mathbb{E}(X^2)$$

Varianza totale.

$$Var(Z) = Var\left(\mathbb{E}(Z|\mathcal{G})\right) + \mathbb{E}\left(Var(Z|\mathcal{G})\right)$$

Covarianza.

$$Cov(aX + b, Y) = aCov(X, Y)$$

Cauchy-Schwartz. ***

$$(\sum xy)^2 \leq \sum x^2 \sum y^2$$

e in particolare può essere utile $(\frac{1}{N}\sum x)^2 \leq \sum x^2$.

Lemma 2.1 (Fatou)**.** *Se $f_1, f_2, \ldots$ è una successione di funzioni non negative e misurabili definite su uno spazio di misura (S, Σ, μ), allora:*

$$\int_S \liminf_{n\to\infty} f_n \, d\mu \leq \liminf_{n\to\infty} \int_S f_n \, d\mu$$

Lemma 2.2 (Fatou inverso)**.** *Sia $f_1, f_2, \ldots$ una successione di funzioni misurabili con valori appartenenti a $\mathbb{R}$ esteso definita su uno spazio di misura (S, Σ, μ). Se esiste una funzione non negativa g, misurabile e con $\int_S g \, d\mu < \infty$ su S, tale che $f_n \leq g$ per ogni n, allora:*

$$\int_S \limsup_{n\to\infty} f_n \, d\mu \geq \limsup_{n\to\infty} \int_S f_n \, d\mu$$

Teorema 2.3 (Cochran)**.** *Sia $X \sim \mathcal{N}\left(\mu, \sigma^2 Id\right)$ e $F_1 \oplus^\perp \cdots \oplus^\perp F_m = \mathbb{R}^n$. Notiamo P_{F_i} la matrice di proiezione ortogonale su F_i di dimensione d_i. Allora*

- *I vettori aleatori $P_{F_1}X, \ldots, P_{F_m}X$ sono a due a due indipendenti e di legge $\mathcal{N}(P_{F_1}\mu, \sigma^2 P_{F_1}), \ldots, \mathcal{N}(P_{F_m}\mu, \sigma^2 P_{F_m})$.*
- *le variabili aleatorie reali $\dfrac{\|P_{F_1}(X-\mu)\|^2}{\sigma^2}, \ldots, \dfrac{\|P_{F_m}(X-\mu)\|^2}{\sigma^2}$ sono a due a due indipendenti di leggi $\chi^2(d_1), \ldots, \chi^2(d_m)$.*

Furbizie per i conti. Possiamo fare i conti come se $\mathbb{E}(X) = 0$, a meno di sostituire X con $X - \mathbb{E}(X)$ poi risostituire di nuovo alla fine.

2.1 Exponential family

Referenze: Bayesian Estimation & Stochastic Opt - Vikram Krishnamurthy 2019 [3]

Normal, Bernoulli, Gamma, Chi-Squared, Beta, Binomial, Poisson, ...

$$p_\theta(x) = \exp(\theta' T(x) - A(\theta))h(x)$$

where θ is a vector of natural parameters, $T(x)$ is the sufficient statistic, $A(\theta)$ is the log partition function = log of normalization factor and it is convex in θ.

Osservazione x and θ only interact in $\theta' T(x)$ and θ' here stands for the transpose vector.

Properties of Exponential Family

- Exponential family pdfs are closed under multiplication.
- Product of Gaussian densities is Gaussian.
- Stein' Identity: Useful for computing moments. If X belongs to exponential family, and g is differentiable, then

$$\mathbb{E}\left[\left(\frac{\nabla h(X)}{h(X)} + \theta' T(X)\right) g(X)\right] = -\mathbb{E}[\nabla g(X)]$$

  an easy example is for an univariate normal: taking $g(x) = 1$ or $g(x) = x$, we obtain mean and variance.

- Maximum Likelihood Estimation: Given N iid observations $x_1, \ldots, x_n$, the MLE θ^* satisfies

$$\nabla_\theta A(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$$

  and the convexity of $A(\theta)$ implies unique global MLE; observe that sufficient statistic $\frac{1}{n}\sum_{i=1}^{n} T(x_i)$ summarizes data.

2.2 LGN

Referenze: [2]

Teorema 2.4 (LGN)**.** *Sia $X_1, X_2, \ldots X_n, \ldots$ una successione di v.a. i.i.d. con media μ. Sia $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Allora con probabilità 1:*

$$\lim_{N\to+\infty} \overline{X}_N = \mu.$$

Esempi di utilizzo.

- Grazie alla convergenza quasi certa dello stimatore $I_M = \frac{1}{M}\sum_{m=1}^{M} f(U_m)$ verso $\mathbb{E}(f(U_1)) = \int_{[0,1]^d} f(x)dx$ possiamo stimare il valore dell'integrale. Notiamo che è sempre possibile effettuare un cambiamento di variabile per riportarci ad avere $[0,1]^d$ come dominio di integrazione.

:) Il vantaggio rispetto alle formule di quadratura: per queste ultime in generale non esiste un modo semplice per passare da N a $N+1$ punti, quindi aumentare la precisione richiede ricominciare i calcoli dall'inizio.

2.3 TCL

Referenze: [2]

Teorema 2.5 (TCL)**.** *Sia (X_n) una successione di v.a. indipendenti, identicamente distribuite, con varianza σ^2 e media μ. Allora la v.a. $Z_n := \frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ converge in legge ad una gaussiana canonica $\mathcal{N}(0,1)$.*

Riformulazione. Possiamo scrivere il TCL come

$$\sqrt{M}(\overline{X}_M - \mathbb{E}(X)) \to \mathcal{N}(0, K)$$

dove K è la matrice di covarianza e la convergenza è in legge. O ancora

$$(\sqrt{K_M})^{-1}\sqrt{M}(\overline{X}_M - \mathbb{E}(X)) \to \mathcal{N}(0, Id)$$

con K_M stimatore non distorto della covarianza.

Esempi di utilizzo.

:) Nell'esempio di calcolo di integrale, il TCL fornisce una stima asintotica sull'errore che **non** dipende dalla dimensione d.

:) La convergenza richiede poche ipotesi.

:) La velocità di convergenza è $\sqrt{M}$ indipendentemente dalla dimensione.

- Errore aleatorio caratterizzato da $\sigma^2 = Var(X)$.

:/ Errore statisticamente più grande secondo la matrice di covarianza.

Teorema 2.6 (Delta method)**.** *Sia X un vettore aleatorio con matrice di covarianza K e sia $f\colon \mathbb{R}^d \to \mathbb{R}$ differenziabile in $\mathbb{E}(X)$. Allora*

$$\sqrt{M}(f(\overline{X}_M) - f(\mathbb{E}(X))) \to \mathcal{N}(0, \nabla f(\mathbb{E}(X))K\nabla f(\mathbb{E}(X))).$$

(La convergenza è in legge per $M \to +\infty$).

Osservazioni sulla *delta method*.

:(La varianza limite in generale non si scrive direttamente come limite di una varianza empirica calcolata insieme alla media empirica.

:(In generale $\mathbb{E}(f(\overline{X}_M)) \neq f(\mathbb{E}(X))$. Anche se $\mathbb{E}(\overline{X}_M) = \mathbb{E}(X)$ (cioè lo stimatore è non distorto), può capitare che lo stimatore $f(\overline{X}_M)$ sia distorto. In particolare:

  - Se f è convessa (*risp. concava*) $\mathbb{E}(f(\overline{X}_M)) \geq f(\mathbb{E}(X))$ (*risp.* $\leq$).
  - Se $f \in C_b^4$ e X ha momenti di ordine $4+\delta$ per $\delta \in (0,1]$:
$$\mathbb{E}(f(\overline{X}_M) - f(\mathbb{E}(X)) = \frac{c_1}{M} + \frac{c_2}{M^2} + o(M^{-2}).$$

  :) Se f è liscia, la distorsione converge a 0 più rapidamente di quanto previsto dal TCL.

- **Ottenere un intervallo di confidenza.** Ci sono vari modi:
  1. Stimare K su un campione $(X_m)_m$ e poi calcolare la derivata di f in $\overline{X}_M \approx \mathbb{E}(X)$.
  2. **Sectioning method.** Applicato generalmente ad algoritmi probabilistici con output centrati in μ e aventi una distribuzione asintotica gaussiana con varianza sconosciuta σ^2. Consiste in suddividere le M simulazioni in n sottoinsiemi di taglia M/n. L'algoritmo produce allora n output indipendenti per ognuno dei quali calcoliamo media e varianza.

Proposizione 2.7. *Siano* $\overline{X}_{1,M} = \frac{2}{M}\sum_{m=1}^{M/2} X_i$ *e* $\overline{X}_{2,M} = \frac{2}{M}\sum_{m=M/2+1}^{M} X_i$. *Gli stimatori*
$$\overline{f}_M := \max(\overline{X}_M, a) \quad \underline{f}_M := \mathbb{1}_{\overline{X}_{1,M} \geq a}\overline{X}_{2,M} + \mathbb{1}_{\overline{X}_{1,M} < a} a$$
convergono q.c. a $\max(\mathbb{E}(X), a)$ *per* $M \to +\infty$ *e vale*
$$\mathbb{E}(\underline{f}_M) \leq \max(\mathbb{E}(X), a) \leq \mathbb{E}(\overline{f}_M).$$

2.4 La speranza condizionale

Proprietà.

- linearità: $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$
- $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$
- Iterazione: $\mathcal{H} \subset \mathcal{G} \Rightarrow \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$
- Monotonia: $X \leq Y \Rightarrow \mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$
- Convergenza monotona: $X_n \geq 0$ $X_n \to X$ q.c. crescente, allora $\mathbb{E}[X_n|\mathcal{G}] \underset{n\to\infty}{\longrightarrow} \mathbb{E}[X|\mathcal{G}]$.
- Indipendenza: se X è indipendente da $\mathcal{G}$, allora $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$
- Se Z est $\mathcal{G}$-misurabile, $\mathbb{E}[XZ|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}]$
- Se X est $\mathcal{G}$-misurabile, $\mathbb{E}[X|\mathcal{G}] = X$
- Disuguaglianza di Jensen: se ϕ è convessa e $\phi(X)$ integrabile,
$$\mathbb{E}[\phi(X)|\mathcal{G}] \geq \phi(\mathbb{E}[X|\mathcal{G}])$$

3 Come calcolare un integrale

Referenze: [1]

I metodi principali per calcolare (o stimare) un integrale

$$I = \int_{\mathbb{R}^d} \psi(f(x))p(x)dx = \mathbb{E}[\psi(f(X))] = \mathbb{E}[\psi(Y)]$$

sono tre: le formule di quadratura, Monte-Carlo, Quasi Monte-Carlo.

Vediamo di seguito le proprietà fondamentali, vantaggi e svantaggi di ciascuno.

3.1 I metodi di quadratura

Referenze: [1]

Approssimiamo

$$I := \int_{\mathbb{R}^d} \psi(f(x))p(x)dx \approx \sum_{i=1}^{n} w_i f(x_i) =: Q(f).$$

I metodi di quadratura in breve.

:($x \mapsto \psi(f(x))$ dev'essere sufficientemente regolare;

:(la dimensione d non dev'essere troppo grande;

3.2 In dimensione 1

Preliminari.

- Non esistono formule di quadratura a n punti di ordine $\geq 2n$ (prendere $f(x) = \prod_{i=1}^{n}(x - x_i)^2$);
- è possibile arrivare all'ordine $2n - 1$ con n punti;

Interpolazione polinomiale.

l'interpolazione polinomiale per mezzo dei **polinomi di Lagrange**:

$$l_j(x) = \prod_{k \neq j} \frac{x - x_k}{x - x_j}$$

(con $l_j(x_k) = \delta_{jk}$) che formano una base dello spazio vettoriale dei polinomi di grado $\leq n$. Il polinomio interpolante i punti $(x_0, y_0), \ldots, (x_n, y_n)$ è

$$L_n(x) = \sum_{j=0}^{n} y_j l_j(x)$$

l'unico polinomio di grado al più n che verifica $L_n(x_i) = y_i$ per ogni $i = 0, \ldots, n$.

Scelta della griglia.

:(Se scegliamo una griglia di punti regolare (cioè in cui i punti sono equispaziati) incontriamo il fenomeno di Runge: oscillazioni ai bordi;

:/ per ogni griglia di punti, esiste una funzione per la quale la successione di polinomi interpolanti diverge;

:) per ogni funzione continua, possiamo costruire una griglia di punti per cui la successione di polinomi interpolanti converge;

per misurare la qualità di un'interpolazione (scelta di una costruzione esplicita di un elemento di uno spazio vettoriale di dimensione finita che approssima la funzione f) si utilizza la **costante di Lebesgue**, cioè la norma dell'operatore di proiezione sullo spazio scelto; ecco alcuni esempi:

- polinomi di Lagrange su griglia regolare:

$$\sim \frac{2^{n+1}}{en \log n}$$

- polinomi di Lagrange con griglia di Chebyshev:

$$\sim \frac{2}{\pi} \log n$$

La tabella 1 presenta le principali proprietà per le formule di quadratura in dimensione $d = 1$.

3.3 In dimensione d

Referenze: [1]

La tabella 2 presenta le principali proprietà per le formule di quadratura in dimensione $d > 1$.

3.4 Monte-Carlo

Referenze: [1]

Il metodo di Monte-Carlo è basato sulla rappresentazione dell'integrale come un valore atteso e sulla possibilità di generare facilmente delle v.a. i.i.d. di legge data.

Lo stimatore MC

$$I_n = \frac{1}{n} \sum_{k=1}^{n} \psi(f(X^{(k)}))$$

è non distorto, convergente con errore quadratico dell'ordine di $\frac{1}{n}$.

:) Nessuna regolarità richiesta per f o ψ;

	Formule di quadratura	Proprietà
Newton-Cotes	$$Q_n(f) = \sum_{j=0}^{n} w_j f(x_j)$$ $$w_j = \frac{1}{b-a}\int_a^b \prod_{0\le i\le n, i\neq j} \frac{x-x_i}{x_j-x_i}dx$$	• ordine: $\begin{cases} n & n \text{ dispari} \\ n+1 & n \text{ pari} \end{cases}$:(instabile quando aumentiamo l'ordine (fenomeno di Runge con polinomi oscillanti e pesi oscillanti); :) utilizzo composito: dividiamo l'intervallo d'integrazione e applichiamo una quadratura semplice su ogni pezzo, poi sommiamo; • stima dell'errore: $$\left\lvert Q_N(f) - \frac{1}{b-a}\int_a^b f(x)dx \right\rvert \le C \left\lVert f^{(r+1)} \right\rVert_\infty \frac{(b-a)^{r+1}}{N^{r+1}}$$
Gauss	$$w_j^{(n)} = \int_{\mathbb{R}} \prod_{i\neq j} \frac{x-x_i^{(n)}}{x_j^{(n)}-x_i^{(n)}} p(x)dx$$ $$w_j^{(n)} = \frac{a_n}{a_{n-1}} \frac{\int_{\mathbb{R}} P_{n-1}(x)^2 p(x)dx}{P_n'(x_j^{(n)})P_{n-1}(x_j^{(n)})}$$	• p è una densità di probabilità e la famiglia di polinomi (P_n) è ortonormale per la densità p; • ogni polinomio P_n ha esattamente n radici distinte $x_j^{(n)}$; :) l'ordine è $2n-1$; :) i pesi sono positivi, e il metodo è stabile; :(i nodi non sono intrecciati (e non possono quindi essere utilizzati per formule d'ordine superiore); • stima dell'errore: $$\left\lvert Q_n(f) - \int_{\mathbb{R}} f(x)p(x)dx \right\rvert \le \frac{\left\lVert f^{(2n)} \right\rVert_\infty}{(2n)!a_n^2}$$
Clenshaw-Curtis	$$Q_n(f) = \frac{\hat{a}_0}{2} + \sum_{k=1}^{n/2-1} \frac{\hat{a}_{2k}}{1-(2k)^2} + \frac{1}{2}\frac{\hat{a}_n}{1-n^2}$$ $$\hat{a}_k = \frac{1}{n}[(-1)^k f(-1) + f(1) + 2\sum_{j=1}^{n-1} f\left(\cos\left(\frac{j\pi}{n}\right)\right)\cos\left(\frac{kj\pi}{n}\right)]$$	• le formule sono per integrare $\frac{1}{2}\int_{-1}^{1} f(x)dx$; • f dev'essere sviluppabile in serie di Fourier; • errore dell'ordine di $\frac{1}{[n^k k]}$; :) i coefficienti $\hat{a}_k$ si calcolano con una FFT;

Tabella 1: Confronto tra i vari metodi di quadratura unidimensionali

Formule di quadratura	Proprietà
Tensorizzazione: $$Q_n(f) = \sum_{j_1=1}^{n} \cdots \sum_{j_d=1}^{n} w_{j_1} \cdots w_{j_d} f(x_{j_1}, \ldots, x_{j_d})$$	:(utilizza n punti in ogni dimensione, n^d in totale; :(le performance peggiorano all'aumentare della dimensione; :/ $O(2^{ld})$ punti se in ogni dimensione ce ne sono $O(2^l)$; • errore d'integrazione: $O(2^{-lr})$;
Griglie sparse di Smolyak : $$Q_l^{(d)}(f) = \left(\sum_{i=1}^{l} (Q_i^{(1)} - Q_{i-1}^{(1)}) \otimes Q_{l-i+1}^{(d-1)} \right)(f)$$ $$Q_i^{(1)} \otimes Q_{l-i+1}^{(d-1)} = \sum_{j=1}^{n_i^{(1)}} \sum_{j'=1}^{n_{l-i+1}^{(d-1)}} w_{j,i}^{(1)} w_{j',l-i+1}^{(d-1)} f(x_{j,i}^{(1)}, x_{j',l-i+1}^{(d-1)})$$	• $Q_n^{(1)}$ è una famiglia di formule di quadratura unidimensionale con griglia annidata; :) $O(2^l l^{d-1})$ punti se in ogni dimensione ce ne sono $O(2^l)$; • errore d'integrazione: $O(2^{-lr} l^{(d-1)(r+1)})$; :/ la convergenza dipende dalla regolarità della funzione e dalla dimensione;

Tabella 2: Confronto tra i vari metodi di quadratura multidimensionali

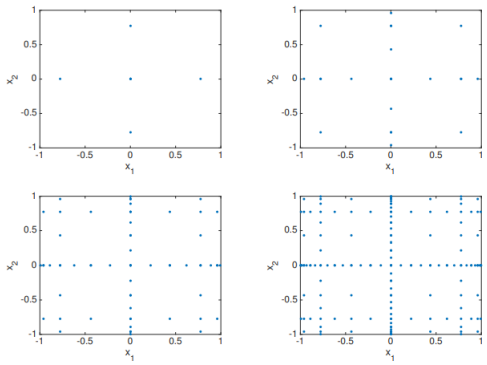

Figura 2: Griglie di Smolyak in dimensione $d = 2$ a partire da griglie di Clenshaw-Curtis ($l = 2, 3, 4, 5$)

:) usando il TCL e il lemma di Slutsky possiamo costruire degli intervalli di fiducia asintotici $\mathbb{P}\left(I \in [I_n - 1.96\frac{\hat{\sigma_n}}{\sqrt{n}}, I_n + 1.96\frac{\hat{\sigma_n}}{\sqrt{n}}]\right) \approx 0.95$.

:) il metodo è parallelizzabile;

:) possiamo aggiungere punti per migliorare la stima;

:/ la velocità di convergenza è indipendente dalla dimensione, ma lenta;

3.5 Quasi Monte-Carlo

Referenze: [1]

Si veda anche la sottosezione 9.7. L'idea è di costruire una famiglia di punti nel cubo $[0, 1]^d$ che saranno utilizzati come nodi di una formula di quadratura. In pratica il metodo ha le seguenti proprietà:

- è a metà strada tra i metodi di quadratura e Monte-Carlo;

:) riduce l'errore fino a $c_d \frac{(\log n)^d}{n}$ se f è un po' regolare;

:/ la costante c_d esplode con la dimensione;

3.6 Altri metodi

Referenze: [1]

Cumul quadratique (méthode sandwich). Permette di calcolare analiticamente e rapidamente la media e la varianza. Conviene per piccole variazioni delle v.a. in ingresso e un codice sufficientemente regolare. Se $\mu = (\mathbb{E}[X_i])_i$ e $C = (Cov(X_i, X_j))_{i,j}$ allora l'approssimazione ottenuta è:

$$\mathbb{E}[Y] \approx f(\mu) \qquad Var(Y) \approx \nabla f(\mu)^T C \nabla f(\mu).$$

- è un'applicazione del metodo più generale che utilizza un metamodello (che in questo caso con una superficie di risposta affine)

:) ci serve conoscere solamente $f(\mu)$ e $\nabla f(\mu)$ per ottenere questa stima;

:(non abbiamo in generale un controllo sull'errore;

3.7 Probabilità di superamento di soglia

Referenze: [1]

L'interesse è di calcolare la probabilità che una certa soglia y_s sia superata:

$$p_s = \mathbb{P}(Y \geq y_s) = \mathbb{P}(X \in F) = \int_{\mathbb{R}^d} \mathbb{1}_F(x) p(x) dx = \int_F p(x) dx$$

con $F = \{x \in \mathbb{R}^d \mid f(x) \geq y_s\}$.

Metodo FORM (First Order Reliability Method). Si tratta di un metodo analitico che:

:) richiede poche chiamate al codice;

:(non è esatto e non fornisce un controllo sull'errore;

Per utilizzare il metodo FORM, :

- supponiamo X_i indipendenti di legge $\mathcal{N}(0,1)$;
- calcoliamo (supponendolo unico) il punto di superamento più vicino all'origine[1]:
$$x_s = \arg\min_{x \in F} \|x\|^2$$
- approssimiamo F con un semispazio $\hat{F}$ che permette di calcolare esplicitamente $\hat{p}_s = \int_{\hat{F}} p(x) dx$.
- $\hat{F}$ è delimitato da un iperpiano che passa per x_s e gli è ortogonale:
$$\hat{p}_s = \Phi(-\|x_s\|)$$
dove Φ è la funzione di ripartizione della gaussiana $\mathcal{N}(0,1)$.

Metodo SORM (Second Order Reliability Method). Come nel caso FORM, ma invece di un iperpiano, utilizziamo una superfice quadratica che passa per x_s e il cui piano tangente in x_s gli è ortogonale (cioè è l'iperpiano del FORM). Troviamo allora

$$\hat{p}_s = \Phi(-\|x_s\|) \prod_{i=1}^{d-1} \frac{1}{\sqrt{1 + \|x_s\| \kappa_i}}$$

dove κ_i sono le curvature della superficie in x_s, determinate a partire dal gradiente e dall'hessiano di f in x_s.

[1] Si tratta di un problema di ottimizzazione con vincoli.

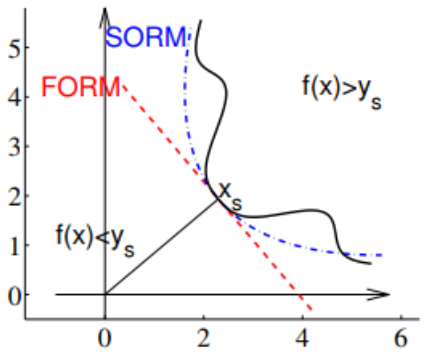

Figura 3: Principio dei metodi FORM-SORM: la linea nera corrisponde alla superficie $\{f(x) = y_s\}$; le linee rosse e blu sono rispettivamente le approssimazioni lineari e quadratiche intorno al punto x_s.

É fondamentale applicare i metodi FORM e SORM nello *spazio standard*, nel quale la densità p è quella di un vettore X di v.a. gaussiane i.i.d. $\mathcal{N}(0,1)$. Per riportarci a questo caso, si possono usare le trasformazioni isoprobabiliste di Rosenblatt:

1. $\phi_i^{(1)}(x) = F_{i|1,\dots,i-1}(x_i|x_1,\dots,x_{i-1})$ funzione di ripartizione della v.a. X_i sapendo $\{X_1 = x_1,\dots,X_{i-1} = x_{i-1}\}$.
2. $\phi^{(1)}(X)$ ha coordinate indipendenti di legge $\mathcal{U}(0,1)$;
3. $\phi^{(2)}(x) = (\Phi^{-1}(z_1),\dots,\Phi^{-1}(z_d))$ è tale che $\phi^{(2)} \circ \phi^{(1)}(X)$ ha coordinate indipendenti di legge $\mathcal{N}(0,1)$.

:) Nel caso in cui il vettore X ha già coordinate indipendenti aventi funzioni di ripartizione F_i, basta porre

$$\phi(X) = ((\Phi^{-1}(F_1(X_1)),\dots,\Phi^{-1}(F_d(X_d))).$$

4 Da ricordare

Referenze: [2]

Metodi e dimensioni. Se la dimensione è bassa, le formule di quadratura e la discretizzazione sono i metodi più efficienti. Spesso è meglio usare i metodi di Quasi Monte-Carlo per medie dimensioni e di Monte-Carlo per grandi dimensioni.

$\boxed{\mathbb{E} \neq \int}$ Possiamo sempre rappresentare un integrale come una speranza interpretando probabilisticamente la misura di Lebesgue, ma non ogni

speranza può essere scritta come un integrale con una funzione di densità esplicita. (In particolare un metodo di Monte-Carlo può essere l'unica soluzione in quest'ultimo caso...)

Vettori gaussiani come trasformazioni affini di v.a. gaussiane standard. Se $X \sim \mathcal{N}(0, I_n)$ e $m \in \mathbb{R}^d$, $L \in \mathbb{R}^{d\times n}$ allora $m + LX \sim \mathcal{N}(m, LL^T)$

Intervalli de fiducia. Fornire un intervallo di confidenza è essenziale e va fatto sistematicamente nella stima di una speranza con metodi di Monte-Carlo.

5 Simulare una variabile aleatoria

Referenze: [2]

Innanzitutto un risultato teorico utile.

Proposizione 5.1. *Sia X una v.a. reale con funzione di ripartizione $F(x) = \mathbb{P}(X \leq x)$. Definiamo il quantile $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. Allora*

$$F^{-1}(U) \sim X$$

e se F è continua, allora $F(X) \sim \mathcal{U}([0,1])$.

Nome	**Densità**	**Funzione di $\mathcal{U}$**
Esponenziale $\mathcal{E}(\lambda)$	$\lambda e^{-\lambda t}\mathbb{1}_{t\geq 0}$	$-\frac{1}{\lambda}\log(U)$
Geometrica $\mathcal{G}(p)$		$1 + \left\lfloor \frac{\log(U)}{\log(1-p)} \right\rfloor$
Cauchy	$\frac{\sigma}{\pi(x^2+\sigma^2)}\mathbb{1}_{x\in\mathbb{R}}$	$\sigma \tan\left(\pi\left(U - \frac{1}{2}\right)\right)$
Rayleigh	$\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}\mathbb{1}_{x\geq 0}$	$\sigma\sqrt{-2\log(U)}$
Pareto	$\frac{ab^a}{x^{a+1}}\mathbb{1}_{x\geq b}$	$\frac{b}{U^{\frac{1}{a}}}$
Weibull	$\frac{a}{b^a}x^{a-1}e^{-\left(\frac{x}{b}\right)^a}\mathbb{1}_{x\geq 0}$	$b(-\log(U))^{\frac{1}{a}}$

Tabella 3: Principali distribuzioni ottenibili a partire da una v.a. uniforme

Proposizione 5.2 (**Simulazione di v.a. discrete**). *Sia $(p_n)_{n\geq 0}$ una successione di numeri reali positivi tale che $\sum_{n\geq 0} p_n = 1$. E sia $(x_n)_{n\geq 0}$ una successione di numeri reali. Allora*

$$X = \sum_{n\geq 0} x_n \mathbb{1}_{p_0+\cdots+p_{n-1}\leq U<p_0+\cdots+p_n}$$

è una v.a. discreta che verifica $\mathbb{P}(X = x_n) = p_n$ per $n \geq 0$.

Il metodo di rigetto.

:) Permette di simulare una v.a. con una distribuzione data a partire dalla simulazione di un'altra v.a. più semplice da generare e da una v.a. uniforme.

:) Per simulare una v.a. Z condizionatamente a un evento A: simulare ripetutamente (Z, A) e rigettare se A non si verifica.

Nome	$(p_n)_n$	$(x_n)_n$
Bernoulli $\mathcal{B}(p)$	$p_0 = 1-p, p_1 = p$	$x_0 = 0, x_1 = 1$
Binomiale $Bin(n,p)$		
Poisson $\mathcal{P}(\theta)$	$p_n = e^{-\theta}\frac{\theta^n}{n!}$	$x_n = n$

Tabella 4: Principali distribuzioni discrete ottenibili col metodo della proposizione 5.2.

L'algoritmo di rigetto ha una durata aleatoria distribuita secondo una legge geometrica $\mathcal{G}(\mathbb{P}(A))$: più A è probabile, più la simulazione è rapida.

Listing 1: Pseudocodice per il metodo di rigetto

```
c = upper bound f/g
while c*u*g(y)<=f(y):
    y = simulation density g
    u = rand
return y;
    /* y ha distribuzione di densità f */
```

Algoritmo per la legge condizionata senza rigetto. Poniamo $S_k = [x_{k-1}, x_k)$. Allora, per ogni k

$$F_k(x) = \mathbb{P}(X \le x | X \in S_k) = \begin{cases} 1 & x > x_k \\ \frac{\mathbb{P}(x_{k-1} \le x < x_k)}{\mathbb{P}(x_{k-1} \le A < x_k)} = \frac{F(x)-F(x_{k-1})}{F(x_k)-F(x_{k-1})} & x_{k-1} \le x \le x_k \\ 0 & x < x_{k-1} \end{cases}$$

Quindi simuliamo $U \sim \mathcal{U}([0,1])$ e calcoliamo $X = F_k^{-1}(U) = F^{-1}(F(x_{k-1}) + U(F(x_k) - F(x_{k-1}))$.

Simulazioni e copula. Per simulare un vettore $(X_1, \dots, X_d)$ con copula C e distribuzioni marginali $F_1, \dots, F_d$:

1. generare un vettore aleatorio $(Y_1, \dots, Y_d)$ con copula C e marginali arbitrarie;
2. calcolare $U_i = F_{Y_i}(Y_i)$;
   - ogni $U_i \sim \mathcal{U}([0,1])$
   - la copula di $U = (U_1, \dots, U_d)$ è ancora C.
3. calcolare $X_i = F_i^{-1}(U_i)$.

5.1 Vettori gaussiani $\mathcal{N}(0, K)$

Referenze: [2], [1]

Se l'obiettivo è di simulare la legge $\pi(x)dx = \mathcal{N}(0, K)$ (con K una matrice di grandi dimensioni) possiamo utilizzare una catena di Markov $(X_n)_n$ che converge a π in legge in un tempo molto grande per cui $X_n, X_{2n}, \dots$ siano approssimativamente indipendenti e di legge π per $n \to +\infty$. In particolare abbiamo il seguente teorema:

Teorema 5.3. *Sia $P(x, dy) = (1-a(x))\delta_x(dy)+p(x, y)dy$ la legge di transizione della catena di Markov $(X_n)_n$ con $p(x, y) > 0$ $\forall x, y \in \mathbb{R}^d$. Supponiamo π invariante per P:*

$$\pi(dy) = \int_{\mathbb{R}^d} \pi(dx)P(x, dy).$$

Allora per ogni punto iniziale x la legge di X_n converge a π per $n \to +\infty$.

Reversibilità e invarianza. La condizione di invarianza è implicata dalla reversibilità:

$$\pi(x)p(x, y) = \pi(y)p(y, x) \quad \forall x \neq y \in \mathbb{R}^d.$$

5.1.1 K^{-1} nota

Nel caso in cui la matrice $V = K^{-1}$ sia nota o facilmente calcolabile $\pi(x) \propto e^{-\frac{1}{2}xVx}$.

Metropolis-Hastings L'algoritmo di Metropolis-Hastings fa parte dei metodi Monte-Carlo basati su Catene di Markov (MCMC). Lo scopo è di generare una catena di Markov ergodia la cui misura invariante è la densità $\pi(x)$.

1. **Proposizione** di una transizione verso y secondo una legge di densità positiva $q(x, y)$;
2. **Accettazione** del punto y proposto con probabilità

$$a(x, y) = \min\left(1, \rho(x, y)\right)^{\,2} \qquad \rho(x, y) := \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

:) Non serve conoscere la costante di normalizzazione di π per il calcolo.

Nel caso in cui a sia scelta come in [2], l'accettazione equivale a una transizione con Dirac in $y = x$ e densità $p(x, y) = \max\left(q(x, y), \frac{\pi(y)q(y,x)}{\pi(x)}\right)$ per $y \neq x$.

[2] In [2] si propone $a(x, y) = \max\left(q(x, y), \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$. In [1] si sottolinea che l'importante è a sia una funzione che verifichi $F(\rho) = \rho F(1/\rho)$; un altro esempio potrebbe essere quindi $F(\rho) = \frac{\rho}{1+\rho}$.

:) $\pi(x)p(x,y)$ è evidentemente simmetrica in x, y e la condizione di reversibilità è quindi soddisfatta.

Gibbs Una transizione $X_n \curvearrowright X_{n+1}$ è decomposta in d transizioni che cambiano una coordinata alla volta.

Passo 1) Ricampionamento di x_1 secondo $\pi(x_1|x_2,\dots,x_d) = \frac{\pi(x_1,x_2,\dots,x_d)}{\int_{\mathbb{R}} \pi(z_1,x_2,\dots,x_d)dz_1}$ per ottenere $y = (y_1, x_2, \dots, x_d)$;

...

Passo i) ricampionamento di x_i secondo $\pi(x_i|y_1,\dots,y_{i-1},x_{i+1},\dots,x_d) = \frac{\pi(y_1,\dots,y_{i-1},x_i,x_{i+1},\dots,x_d)}{\int_{\mathbb{R}} \pi(y_1,\dots,y_{i-1},z_i,x_{i+1},\dots,x_d)dz_i}$ per ottenere $y = (y_1,\dots,y_i,x_{i+1},\dots,x_d)$;

...

Passo d) Ricampionamento di x_d secondo $\pi(x_d|y_1,\dots,y_{d-1}) = \frac{\pi(y_1,\dots,y_{d-1},x_d)}{\int_{\mathbb{R}} \pi(y_1,\dots,y_{d-1},z_d)dz_d}$ per ottenere $y = (y_1, y_2, \dots, y_d)$.

:) La condizione di invarianza è soddisfatta ad ogni passo.

5.1.2 K^{-1} sconosciuta

Nel caso in cui K^{-1} non sia facilmente calcolabile possiamo:

1. ricondurci al caso $Var_\pi(X_i) = K_{ii} = 1$;
2. fare un passo in una direzione aleatoria, cioè porre $X_{n+1} = X_n + (g_n - X_{i_n,n})K_{:i_n}$ dove
   - $K_{:i}$ è la i-esima colonna di K;
   - $X_{i,n}$ la i-esima coordinata di X_n;
   - i_n è scelto a caso uniformemente su $\{1,\dots,d\}$;
   - $g_n \sim \mathcal{N}(0,1)$;
   - tutte queste v.a. sono indipendenti.

:) La condizione di invarianza è soddisfatta.

6 Sensibilità del valore atteso

Referenze: [2]

Per misurare la variazione del valore atteso $\mathbb{E}[X^\theta]$ in funzione di un parametro $\theta \in \Theta \subset \mathbb{R}$:

1. **"resimulazione"**:

$\sim$ approssimazione con differenze finite;

$$\partial_\theta \mathbb{E}[f(Y^\theta)] \approx \frac{\mathbb{E}[f(Y^{\theta+\varepsilon})] - \mathbb{E}[f(Y^{\theta-\varepsilon})]}{2\varepsilon}$$

è importante simulare utilizzando gli stessi numeri aleatori nei due termini, facendo variare solo il parametro ε (*Common Random Numbers*);

:(lo stimatore della derivata è distorto;

:(la varianza dello stimatore può tendere a $+\infty$ per $\varepsilon \to 0$ quando la funzione f è poco regolare;

2. **pathwise differentiation** (derivazione):

   se si può, si deriva sotto il segno di integrale:

   $$\partial_\theta \mathbb{E}[f(Y^\theta)] = \mathbb{E}[\nabla f(Y^\theta) \partial_\theta Y^\theta].$$

   :(f deve essere regolare (diciamo C^1 con derivata limitata);

   :/ Y^θ come funzione di θ è C^1 q.o. con derivata limitata da una funzione integrabile;

3. **verosimiglianza**:

   - Nel caso in cui Y^θ abbia una densità $p(\theta, \cdot) > 0$ che sia continuamente differenziabile rispetto a θ e di derivata maggiorata da una funzione integrabile:

     $$\partial_\theta \mathbb{E}[f(Y^\theta)] = \mathbb{E}[f(Y^\theta) \partial_\theta (\log(p(\theta, y)))|_{y=Y^\theta}].$$

   :) non chiediamo la regolarità di f;

   :/ la legge dev'essere una funzione regolare ed esplicita di θ;

   - i pesi $\partial_\theta (\log(p(\theta, y)))|_{y=Y^\theta}$ sono centrati (valore atteso nullo), ma possono avere varianza infinita;

Possiamo misurare contemporaneamente $\mathbb{E}[f(Y^\theta)]$ e la sua sensibilità attraverso metodi di Monte-Carlo.

Metodo	**Applicabile se**
Derivazione	$\theta \mapsto Y^\theta$ derivabile q.o.
Verosimiglianza	$\theta \mapsto$ *legge di* Y^θ regolare, ma legge esplicita

Tabella 5: Recap dei metodi per misurare la sensibilità della speranza rispetto a un parametro.

7 Intervalli di confidenza

Referenze: [2]

1. Stima con la disuguaglianza di Čebyšëv:
   - :) la sola condizione è che la varianza sia finita;
   - :(A ε fissato la decrescenza della probabilità di trovarsi fuori dell'intervallo è lenta;
   - :(poco preciso per ε piccolo.
   - Effetto del TCL : competizione tra ε e $\sqrt{N}$.
2. Teorema Centrale del Limite:
   - :(per aumentare la precisione di un fattore n dobbiamo aumentare il numero di simulazioni di n^2;
   - :/ l'errore è distribuito in maniera gaussiana, ma è impossibile controllarlo esattamente;
   - :/ l'intervallo di confidenza ottenuto è asintotico;
3. Disuguaglianze esponenziali:
   - :) decrescenza rapida della probabilità di essere fuori dall'intervallo, maggiore precisione;
   - :) intervallo di confidenza non asintotico;
   - :/ intervallo di confidenza in generale più largo di quello ottenibile col TCL;
   - :/ più difficile da ottenere;

   In generale sono della forma:

$$\mathbb{P}\left(\left|\frac{1}{M}\sum_{m=1}^{M}(X_m - \mathbb{E}(X_m))\right| > \varepsilon\right) \leq 2e^{-\frac{M\varepsilon^2}{c(M,\varepsilon)}}$$

   con $c(M,\varepsilon) > 0$ una costante che dipende da M, ε e dalla distribuzione delle X_m. Queste **disuguaglianze di concentrazione** descrivono come le v.a. si concentrano attorno alle loro medie. Osserviamo che se $c(M,\varepsilon) = c$ non dipende da M né ε l'intervallo di fiducia che si ottiene assomiglia a quello del TCL (fattore di normalizzazione $\sqrt{M}$).

8 Stimare una densità

Referenze: [1]

Per determinare la legge di una v.a. possiamo basarci su pareri di "esperti" che ci diano la legge, la sua forma, o informazioni qualitative o quantitative; possiamo poi studiare il campione di dati attraverso una delle seguenti tecniche per determinare con maggiore precisione la legge cercata:

- Méthode à noyau;
- Méthode des moments;
- Méthode du maximum de vraisemblance;
- Méthode bayésienne;
- Méthode du maximum d'entropie.

Méthode à noyau. Lo stimatore è

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

K è il nucleo, una funzione reale, positiva, pari e normalizzata:

$$\int_{\mathbb{R}} x^2 K(x)dx = 1 \qquad \int_{\mathbb{R}} K(x)dx = 1 \qquad \int_{\mathbb{R}} xK(x)dx = 0$$

di taglia h, che è il parametro critico di questo metodo, in quanto regola quanto la densità stimata sarà liscia. I risultati asintotici consigliano di prendere $h = \hat{\sigma} n^{-1/5}$. Di seguito qualche caratteristica del metodo:

- non parametrico;
- è la generalizzazione dell'istogramma classico;

:) stima la densità in ogni punto del supporto;

:) nessuna ipotesi sulla forma della densità;

:)  Il n'existe pas d'estimateur non-paramétrique qui converge plus vite que l'estimateur à noyau avec le choix de la fenêtre $h \simeq n^{-1/5}$;

:/  la vitesse de convergence est plus faible que la vitesse typique des méthodes paramétrique ($n^{-4/5}$ vs n^{-1});

:(servono molti dati;

:( la méthode supporte assez mal de monter en dimension (l'errore quadratico medio diventa $n^{-4/(4+d)}$, e la velocità di convergenza molto lenta all'aumentare di d);

Méthode des moments. Aggiustiamo il parametro θ della densità p_θ perché i momenti corrispondono a quelli empirici del campione. Di seguito qualche caratteristica del metodo:

- parametrico: cerchiamo di aggiustare il parametro di una famiglia di densità $\{(p_\theta(x))_{x\in\mathbb{R}}, \theta \in \Theta\}$ per adattarci ai dati;

:) meno dati del metodo a nucleo;

:) l'errore quadratico medio è in $\frac{1}{n}$ **se** la famiglia parametrica contiene la densità cercata;

:(l'ipotesi che la legge appartiene alla famiglia scelta è forte;

Méthode du maximum de vraisemblance. Aggiustiamo il parametro θ della densità p_θ affinché la log-verosimiglianza $\log L_x(\theta) = \sum_{i=1}^n \log p_\theta(x_i)$ sia massimale. Di seguito qualche caratteristica del metodo:

- parametrico: cerchiamo di aggiustare il parametro di una famiglia di densità $\{(p_\theta(x))_{x\in\mathbb{R}}, \theta \in \Theta\}$ per adattarci ai dati;
- $L_x(\theta)$ è sia la verosimiglianza dei dati x condizionati a θ, sia la verosimiglianza del parametro sapendo i dati;

:) meno dati del metodo a nucleo;

:) nei casi standard lo stimatore di massima verosimiglianza converge con rischio quadratico in $\frac{1}{n}$ ed è asintoticamente normale ed asintoticamente efficace;

:/ lo stimatore di massima verosimiglianza può esister o no, essere unico o no;

:(l'ipotesi che la legge appartiene alla famiglia scelta è forte;

Méthode bayésienne. Estensione del metodo del massimo di verosimiglianza che fornisce una distribuzioni a posteriori sul parametro θ. Permette di incorporare un'idea a priori sul parametro per ottenere un modello gerarchico.

 Il faut faire attention avec les tests : la réponse d'un test est "je ne peux pas rejeter l'hypothèse que les données sont compatibles avec la famille de lois" ou "je rejette l'hypothèse que les données sont compatibles avec la famille de lois". Autrement dit, *lorsque le test ne rejette pas l'hypothèse que les données sont compatibles avec la famille de lois, cela peut vouloir dire* que les données sont effectivement issues d'un membre de la famille, ou bien *qu'on n'a pas assez de données pour rejeter l'hypothèse avec suffisamment d'assurance.*

Méthode du maximum d'entropie. L'idea è che se non sappiamo nulla su una legge di probabilità, ma solamente che appartiene a una certa classe, allora scegliamo quella con la più grande entropia tra tutte quelle della classe. Questo per due motivi principali: massimizzare l'entropia significa minimizzare la quantità d'informazione a priori; molti sistemi fisici tendono a evolvere verso una situazione di entropia massimale. L'entropia è definita come:

$$E(p) = -\int_{\mathbb{R}} p(x) \log p(x) dx.$$

In alcuni casi conosciamo la forma della legge che risulterà:

:) la legge che massimizza l'entropia di una v.a. di media e varianza date è gaussiana;

:) la legge che massimizza l'entropia di una v.a. a valori in un intervallo dato è quella uniforme sull'intervallo;

:) la legge che massimizza l'entropia di una v.a. positiva con media data è esponenziale.

8.1 Stimare un quantile

Referenze: [1]

Ricordiamo la definizione di α-quantile:

Definizione L'α-quantile è definito da

$$y_\alpha = inf_{y \in \mathbb{R}} \{F_Y(y) \geq \alpha\}$$

dove F_Y è la funzione di ripartizione di Y.

Se la legge di Y ha densità strettamente positiva, esiste un unico y_α che verifica $F_Y(y_\alpha) = \alpha$ e quindi $\mathbb{P}(Y \leq y_\alpha) = \alpha$.

Definizione Definiamo la statistica d'ordine di un campione $(Y_1, \ldots, Y_n)$ come il campione riordinato $(Y_{(1)}, \ldots, Y_{(n)})$ in modo crescente: $Y_{(1)} \leq \cdots \leq Y_{(n)}$.

Stimatore empirico. Uno stimatore empirico per il quantile di ordine α è $\hat{Y}_{\alpha,n} = Y_{\lceil \alpha n \rceil}$. Corrisponde al quantile empirico, cioè al quantile della funzione di ripartizione empirica:

$$\hat{Y}_{\alpha,n} = inf_{y \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, y]}(Y_i) \geq \alpha \right\}.$$

Abbiamo i seguenti risultati asintotici (per n grande):

- $\mathbb{E}(\hat{Y}_{\alpha,n}) = y_\alpha + O\left(\frac{1}{n}\right)$;

- $Var(\hat{Y}_{\alpha,n}) = \frac{\alpha(1-\alpha)}{np_Y^2(y_\alpha)} + O\left(\frac{1}{n^2}\right)$;
- se la densità $p_Y > 0$ e continua in y_α, allora si ha convergenza in legge:

$$\sqrt{n}(\hat{Y}_{\alpha,n} - y_\alpha) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{p_Y^2(y_\alpha)}\right).$$

  In particolare è difficile stimare un quantile quando $p_Y(y_\alpha)$ diventa molto piccolo perché le fluttuazioni dello stimatore possono diventare enormi.

 Poiché asintoticamente vale $\mathbb{P}(\hat{Y}_{\alpha,n} \leq y_\alpha) \simeq 1/2$, il quantile empirico ha praticamente una possibilità su due di "essere falso", cioè più piccolo del vero quantile.

Stimatore di Wilks. Per rimediare all'osservazione precedente, possiamo utilizzare uno stimatore della forma $\hat{Y}_{\alpha,n} = Y_{(\lceil \alpha n \rceil + r)}$.

:) Se notiamo r il più piccolo intero per cui $\sum_{j=0}^{n-\lceil \alpha n \rceil - r} \binom{n}{j} (1-\alpha)^j \alpha^{n-j} \leq 1-\beta$, allora lo stimatore di Wilks $Y_{(\lceil \alpha n \rceil + r)}$ è sicuro a livello β: $\mathbb{P}(Y_{(\lceil \alpha n \rceil + r)} > y_\alpha) \geq \beta$.

:) Lo stimatore di Wilks vale anche per n piccolo ($n \geq n_c = \left\lceil \frac{\log(1-\beta)}{\log \alpha} \right\rceil + 1$ taglia critica del campione a partire dal quale lo stimatore è ben definito).

- Asintoticamente $r \sim \Phi^{-1}(\beta)\sqrt{\alpha(1-\alpha)n}$ con $\Phi^{-1}(\beta)$ il β-quantile gaussiano standard.

:/ La precisione ha un prezzo: la dispersione dello stimatore di Wilks è superiore a quella dello stimatore empirico.

9 Ridurre la varianza

Referenze: [2]


Ridurre la varianza può aiutare ad accelerare la convergenza di un metodo di Monte-Carlo. Ridurla di un fattore n è asintoticamente equivalente a ridurre dello stesso fattore il numero di simulazioni da effettuare per ottenere la stessa precisione.

9.1 Campionamento antitetico

Referenze: [2]

Lemma 9.1. *Sia Y una v.a. reale. Siano $f, g\colon \mathbb{R} \to \mathbb{R}$ rispettivamente non crescente e non decrescente. Supponiamo $f(Y)$ e $g(Y)$ a quadrato sommabile. Allora la loro covarianza è non positiva:*

$$\mathbb{E}(f(Y)g(Y)) \leq \mathbb{E}(f(Y))\mathbb{E}(g(Y)).$$

Osservazione In particolare, se f è *monotona*, per ogni φ non crescente si ha

$$Cov(f(Y), f(\varphi(Y))) \leq 0$$

il che è interessante se Y e $\varphi(Y)$ hanno la stessa legge, poiché possiamo risparmiare una simulazione di Y sostituendola con $\varphi(Y)$.

Esempi di scelta per φ.

- **Uniforme.** Per $Y \sim \mathcal{U}(0,1)$ possiamo prendere $\varphi(Y) = 1 - Y \sim \mathcal{U}(0,1)$.
- **Gaussiana.** Per simmetria basta prendere $\varphi(Y) = -Y$ per mantenere la stessa distribuzione. Questo vale più in generale per una v.a. con distribuzione **simmetrica**.
- **Cauchy.** Se Y è una v.a. con distribuzione di Cahuchy di parametro σ, possiamo prendere $\varphi(Y) = \frac{\sigma^2}{Y}$ nel caso in cui f è pari o costante su $\mathbb{R}_+$ o $\mathbb{R}_-$.

  Il rapporto di due gaussiane è una v.a. di Cauchy, quindi l'inverso di una v.a. di Cauchy è ancora di Cauchy (rapporto inverso di gaussiane).

Stimatore. Lo stimatore per i metodi di Monte-Carlo diventa $\frac{1}{M}\sum_{m=1}^{M} \frac{f(Y_m)+f(\varphi(Y_m))}{2}$.

:) Per la LGN converge q.c. a $\mathbb{E}(X)$.

:) È in ogni caso preciso almeno quanto lo stimatore classico.

Lemma 9.2. *Se la funzione di ripartizione di Y è continua allora $\varphi(y) = F^{-1}(1 - F(y))$ è non crescente e $\varphi(Y)$ ha la stessa legge di Y (cioè è una variabile antitetica associata a Y).*

Rapporto di performance È definito come il rapporto tra la varianza dello stimatore classico e quella del nuovo stimatore. Più è grande e più la scelta di cambiare è vantaggiosa.

9.2 Condizionamento

Referenze: [2]

Nel caso in cui conosciamo esplicitamente $g(z) = \mathbb{E}(X|Z = z)$ abbiamo una riduzione sistematica della varianza:

$$Var(g(Z)) \leq \mathbb{E}(Var(X|Z)) + Var(\mathbb{E}(X|Z)) = Var(X).$$

9.3 Stratificazione

Referenze: [2]

L'idea è di decomporre l'insieme dei valori di Z in k strati $S_1, \ldots, S_k$ per i quali:

- conosciamo $\mathbb{P}(Z \in S_j) = p_j$ con $\sum_{j=1}^{k} p_j = 1$;
- la distribuzione condizionale di X sapendo $Z \in S_j$ può essere simulata in modo efficiente.

Possiamo allora generare M_j simulazioni indipendenti di X_j e considerare lo stimatore

$$I_{M_1,\ldots,M_k} = \sum_{j=1}^{k} p_j \frac{1}{M_j} \sum_{m=1}^{M_j} X_{j,m}$$

che converge q.c. a $\mathbb{E}(X)$ ed ha varianza

$$Var(I_{M_1,\ldots,M_k}) = \sum_{j=1}^{k} p_j^2 \frac{\sigma_j^2}{M_j}$$

dove $\sigma_j^2 = Var(X_j) = Var(X|Z \in S_j)$. Considerando $M = \sum_{j=1}^{k} M_j$ possiamo adottare varie strategie per suddividere il calcolo nei differenti strati:

- **Proporzionale alla probabilità dello strato.** Nel caso $\frac{M_j}{M} = p_j$ abbiamo
$$Var(I_{M_1,\ldots,M_k}) = \frac{\mathbb{E}(Var(X|I)}{M}$$
dove $\{I = j\} = \{Z \in S_j\}$. Questa tecnica riduce sistematicamente la varianza ($\mathbb{E}(Var(X|I)) \leq Var(X)$).

- **Ottimale.** Possiamo minimizzare la varianza e ottenere il numero ottimale di simulazioni per strato $M_j^* = M \frac{p_j \sigma_j}{\sum_{i=1}^k p_i \sigma_i}$ che conduce ad avere

$$Var(I_{M_1,\dots,M_k}) = \frac{1}{M} \left(\sum_{j=1}^k p_j \sigma_j \right)^2 .$$

  Il problema di questa strategia è che richiede di conoscere a priori σ_j.

9.4 Variabili di controllo

Referenze: [2]

Cerchiamo una variabile Z tale che $\mathbb{E}(Z)$ sia una quantità esplicita in termini dei dati del problema e che sia correlata alla v.a. X.

Definizione Una variabile di controllo per il calcolo di $\mathbb{E}(X)$ è una v.a. Z, centrata che può essere simulata contestualmente a X:

$$Z(\beta) = \beta \cdot Z = \sum_{j=1}^d \beta_j Z_j$$

per $\beta \in \mathbb{R}^d$.

Osservazione Aggiungere $Z(\beta)$ alla simulazione di X non cambia la media. Ottimizzando β speriamo di ridurre la varianza. Lo stimatore

$$I_{\beta,M} = \frac{1}{M} \sum_{m=1}^M (X_m - Z_m(\beta))$$

converge q.c. a $\mathbb{E}(X)$ e la varianza asintotica rinormalizzata verifica

$$MVar(I_{\beta,M}) = Var(X - Z(\beta))$$

che speriamo sia migliore di $Var(X)$.

Osservazione Se calcoliamo il β ottimale (che minimizza $\mathbb{E}(X - Z(\beta))$) scopriamo che questo metodo sarà più efficiente se Z e X sono molto correlati.

I polinomi di Hermite Definiti da $H_k(x) = (-1)^k e^{\frac{x^2}{2}} \partial_x^k (e^{-\frac{x^2}{2}})$ sono tali che per G v.a. gaussiana $\mathbb{E}(H_k(G)) = 0$. Una variabile di controllo è data da $Z = (H_1(G), \dots, H_d(G))$.

Altra visione delle variabili di controllo. Seguendo [1]: vogliamo stimare $I = \mathbb{E}(F(X))$ per $F(x) = \psi(f(x))$ funzione determinista. Supponiamo di avere

a disposizione un modello ridotto $f_r(x)$ di $f(x)$. Notiamo $F_r(x) = \psi(f_r(x))$ e $I_r = \mathbb{E}(F_r(X))$. Lo stimatore

$$\hat{I}_n = I_r + \frac{1}{n}\sum_{k=1}^{n} F(X^{(k)}) - F_r(X^{(k)})$$

è non distorto e convergente e di varianza $Var(\hat{I}_n) = \frac{1}{n}(Var(F(X) - F_r(X))$.

Applicazione del metodo [3] On dispose d'un code léger f_r en plus du code lourd f. Le rapport du coût calcul entre un appel à f et un appel à f_r est $q > 1$. On propose l'estimateur

$$\hat{I}_n = \frac{1}{n_r}\sum_{i=1}^{n_r} F_r(X^{(i)}) + \frac{1}{n}\sum_{i=1}^{n} F(X^{(i)}) - F_r(X^{(i)})$$

avec $n_r > n$. La variance totale de l'estimateur est

$$Var(\hat{I}_n) = \frac{1}{n_r}Var(F_r(X)) + \frac{1}{n}Var((F - F_r)(X)).$$

Il s'agit ensuite d'allouer le budget total entre n appels au code lourd f et n_r appels au code léger f_r pour minimiser la variance, sous la contrainte d'un budget total fixé $n_r/q + n(1 + 1/q) = n_{tot}$. Quand q devient très grand (i.e., le code léger est essentiellement gratuit par rapport au code lourd), on retrouve $Var(\hat{I}_n) = \frac{1}{n}(Var(F(X) - F_r(X))$. Cette idée est utilisée en particulier dans le cas où $f(X)$ est la solution d'une équation différentielle (ou aux dérivées partielles) discrétisée finement, avec $f_r(X)$ la solution obtenue avec un schéma de discrétisation grossier (méthode MultiLevel Monte Carlo).

:/ La méthode d'estimation par variable de contrôle n'est en général pas très utile pour l'estimation de probabilité d'événement rare. En effet, le modèle réduit est la plupart du temps conçu pour être raisonnable autour de la tendance centrale, mais pas dans les zones rares.

9.5 Importance sampling

Referenze: [1]

Il metodo nasce dal fatto che la rappresentazione di un integrale come speranza non è unica:

$$I = \mathbb{E}_p(F(X)) = \int_{\mathbb{R}^d} F(x)p(x)dx = \int_{\mathbb{R}^d} \frac{F(x)p(x)}{q(x)}q(x)dx = \mathbb{E}_q\left(\frac{F(X)p(X)}{q(X)}\right).$$

[3] Presente in [1]: Méthodes de Romberg statistiques pour l'estimation de I.

Possiamo allora simulare delle variabili aleatorie secondo una densità "distorta" q che favorisce i valori in una certa zona. Lo stimatore

$$\hat{I}_n = \frac{1}{n}\sum_{k=1}^{n} F(X^{(k)})\frac{p(X^{(k)})}{q(X^{(k)})}$$

è:

:) non distorto: $\mathbb{E}_q(\hat{I}_n) = I$;

:) convergente: $\hat{I}_n \to I$;

:/ con varianza che dipende dalla scelta di q: $Var(\hat{I}_n) = \frac{1}{n}\left(\mathbb{E}_p\left(F(X)^2\frac{p(X)}{q(X)}\right) - \mathbb{E}_p(F(X))^2\right)$.

! In particolare la densità ottimale q_{opt} dipende da I:

$$q_{opt}(x) = \frac{F(x)p(x)}{I} \quad \Rightarrow Var(\hat{I}_n) = 0.$$

Per "trovare" la densità q_{opt} ci sono varie maniere:

- effettuare un'analisi teorica;
- utilizzare un modello ridotto $f_r(x)$ per calcolare la densità ottimale $q_r(x) = \frac{\psi(f_r(x))p(x)}{\int \psi(f_r(x))p(x)dx}$;
- metodi parametrici: scegliere una famiglia di densità e aggiustare i parametri grazie alle simulazioni;
- metodi non parametrici.

Nel caso dei metodi parametrici, il più utilizzato è quello di *cross entropia*. Questo consiste a determinare il parametro θ^* ottimale che minimizza la varianza dello stimatore di *importance sampling* utilizzando una densità $q(x, \theta^*$. Vogliamo allora che questa densità sia "il più vicino possibile" a q_{opt}. Per misurare la distanza tra due densità p e q si usa la distanza di Kullback-Leibler:

$$\mathcal{D}(p, q) = \mathbb{E}_q\left(\log\frac{p(X)}{q(X)}\right) = \int p(x)\log p(x)dx - \int p(x)\log q(x)dx.$$

L'obiettivo è quindi di massimizzare la funzione $\int q_{opt}\log q(x,\theta)dx$ che si esprime come un valore atteso[4]:

$$C(\theta) = \mathbb{E}_p(\psi(f(X))\log q(X,\theta)).$$

Per questo possiamo usare uno stimatore empirico $\hat{C}(\theta) = \frac{1}{n}\sum_{k=1}^{n}\psi(f(X^{(k)}))\log q(X^{(k)},\theta)$.

[4] equivalentemente possiamo massimizzare $\int \psi(f(x))\log q(x,\theta)p(x)dx$.

9.6 Cambi di probabilità

Referenze: [2]

I cambi di probabilità, attraverso le funzioni di verosimiglianza, permettono di codificare la dipendenza dai parametri del modello. Nel caso dei metodi di Monte-Carlo, sono utilizzati negli algoritmi di simulazione di eventi rari.

Definizione Una misura di probabilità $\mathbb{Q}$ sullo spazio $(\Omega, \mathcal{F})$ definisce un cambiamento di probabilità rispetto a $\mathbb{P}$ se esiste una v.a. $L \geq 0$ per cui

$$\mathbb{Q}(A) = \mathbb{E}(L\mathbb{1}_A) \text{ per } A \in \mathcal{F}$$

L è detta densità o verosimiglianza di $\mathbb{Q}$ rispetto a $\mathbb{P}$ e si usano le notazioni:

$$\mathbb{Q} = L\mathbb{P} \quad d\mathbb{Q} = Ld\mathbb{P} \quad \frac{d\mathbb{Q}}{d\mathbb{P}} = L.$$

Le due misure di probabilità si dicono **equivalenti** se $L > 0$ $\mathbb{P}$ q.c. (e in tal caso $\mathbb{P} = L^{-1}\mathbb{Q}$).

Condizione necessaria e condizioni sufficienti. Se $\mathbb{Q}$ è una misura di probabilità, necessariamente $\mathbb{E}_{\mathbb{P}}(L) = 1$. E viceversa se $\mathbb{E}_{\mathbb{P}}(L) = 1$ e $L > 0$ allora $\mathbb{Q}$ è una misura di probabilità.

Log-verosimiglianza. Se le due misure di probabilità sono equivalenti, spesso si considera $L = e^l$ dove l è detta funzione di *log-verosimiglianza.*

Evidente ma importante. Nel caso di equivalenza, un evento è trascurabile per $\mathbb{P}$ se e solo se lo è per $\mathbb{Q}$, quindi ogni uguaglianza q.c. ed ogni limite q.c. valgono per entrambe le probabilità.

Funzioni misurabili. Sia f una funzione misurabile e limitata. Allora

$$\mathbb{E}_{\mathbb{Q}}(f(Y)) = \mathbb{E}_{\mathbb{P}}(Lf(Y)) \quad \mathbb{E}_{\mathbb{P}}(f(Y)) = \mathbb{E}_{\mathbb{Q}}(L^{-1}f(Y)).$$

Densità. Se Y è una v.a. con densità di probabilitàp e q è un'altra densità di probabilità, allora $L = \frac{q}{p}(Y)$ (dove ha senso scriverlo) è una verosimiglianza di una nuova misura di probabilità $\mathbb{Q}$ equivalente a $\mathbb{P}$.

:) L'esempio vale su $\mathbb{R}^d$ se le componenti di Y sono indipendenti sotto $\mathbb{P}$ (cioè $p(y) = p_1(y_1) \dots p_d(y_d)$) e $q(y) = q_1(y_1) \dots q_d(y_d)$:

$$L = \prod_{i=1}^{d} \frac{q_i}{p_i}(Y_i)$$

e le $(Y_i)_i$ sono indipendenti sotto $\mathbb{Q}$, ognuna con distribuzione q_i.

Un cambiamento di probabilità non sempre conserva l'indipendenza! (Per esempio in generale q non si scrive come prodotto delle coordinate...) Ma, se la verosimiglianza dipende da v.a. aggiuntive e indipendenti dalle altre, l'indipendenza è preservata.

A causa della libertà nella scelta di q, la covarianza di Y a priori non è la stessa sotto $\mathbb{P}$ e sotto $\mathbb{Q}$.

Anche se $f(Y)$ è di quadrato sommabile sotto $\mathbb{P}$, può non esserlo sotto $\mathbb{Q}$:

$$Y \sim \mathcal{E}(1) \quad X = e^{\frac{Y}{3}} \quad q(y) = 2e^{-2y} \quad L = \frac{q}{p}(Y)$$

verificano

$$\mathbb{E}(X^2) < +\infty \qquad \mathbb{E}_{\mathbb{Q}}(\left|L^{-1}X\right|^2) = \mathbb{E}_{\mathbb{P}}(L^{-1}\left|X\right|^2) = +\infty$$

Lo stimatore. Se $\mathbb{Q}$ è equivalente a $\mathbb{P}$ e $(L_m, X_m)_m$ è una successione di v.a. distribuite come (L, X) su $\mathbb{Q}$ allora

$$I_{\mathbb{Q},M} = \frac{1}{M}\sum_{m=1}^{M} L_m^{-1}X_m \to \mathbb{E}_{\mathbb{P}}(X) \text{ q.c.}$$

Notiamo che la speranza di una costante in generale non è più esatta se $\frac{1}{M}\sum_{m=1}^{M} L_m^{-1} \neq 1$.

Proposizione 9.3. *Sia X una v.a. reale. $L = \frac{X}{\mathbb{E}(X)}$ definisce una nuova misura di probabilità $\mathbb{Q}$ equivalente a $\mathbb{P}$ per cui $Var(I_{\mathbb{Q},M}) = 0$: basta un solo campione per estimare $\mathbb{E}(X)$.*

Utopia e realtà. In pratica senza conoscere $\mathbb{E}(X)$ è difficile simulare L.

I cambiamenti di probabilità che riducono significativamente la varianza in un calcolo di $\mathbb{E}(X)$ con metodi di Monte-Carlo sono quelli che campionano gli output di X più significativi.

:/ In pratica è difficile mediare tra avere una forma esatta per L, utile per pesare gli output dello stimatore, e una distribuzione esplicita sotto $\mathbb{Q}$, che permetta di generare efficacemente $(X_m)_m$. In generale si preferisce il secondo caso, sebbene niente assicura che L sia vicina a $\frac{X}{\mathbb{E}(X)}$.

Cambiamento di media. Sia Y una v.a. su $\mathbb{R}^d$ con densità $p > 0$.

$$L = \frac{p(Y - \theta_\mu)}{p(Y)}$$

definisce una probabilità $\mathbb{Q}$ equivalente sotto cui Y ha la stessa distribuzione di $Y + \theta_\mu$ sotto $\mathbb{P}$.

Cambiamento di media e varianza. Sia Y una v.a. su $\mathbb{R}^d$ con densità $p > 0$.

$$L = \frac{1}{|\det(\theta_\sigma)|} \frac{p(\theta_\sigma^{-1}(Y - \theta_\mu))}{p(Y)}$$

definisce una probabilità $\mathbb{Q}$ equivalente sotto cui Y ha la stessa distribuzione di $\theta_\sigma Y + \theta_\mu$ sotto $\mathbb{P}$.

Il caso gaussiano. Nel caso particolare in cui $Y \sim \mathcal{N}(0,1)$ otteniamo:

$$L = e^{\theta_\mu Y - \frac{1}{2}\theta_\mu^2}$$

che fornisce una v.a. $\mathcal{N}(\theta_m, 1)$ e

$$L = \frac{1}{\theta_\sigma} e^{\frac{1}{2}(1-\frac{1}{\theta_\sigma})Y^2 + \frac{Y\theta_\mu}{\theta_\sigma^2} - \frac{\theta_\mu^2}{2\theta_\sigma^2}}$$

per cui Y è distribuita come $\mathcal{N}(\theta_\mu, \theta_\sigma^2)$ sotto $\mathbb{Q}$.

Definizione La funzione generatrice dei momenti di Y v.a. reale è $M(\theta) = \mathbb{E}(e^{\theta Y} > 0$.
Tale funzione è finita su un intervallo aperto contenente lo 0, sulla quale $M \in C^\infty$ con $M^{(k)}(\theta) = \mathbb{E}(Y^k e^{\theta Y})$.
Consideriamo la funzione convessa $\Gamma(\theta) = \log(M(\theta)) = \log(\mathbb{E}(e^{\theta Y}))$.

Lemma 9.4. *$L = e^{\theta Y - \Gamma(\theta)}$ definisce una probabilità $\mathbb{Q}_\theta$ equivalente sotto la quale Y ha funzione generatrice dei momenti data da*

$$\mathbb{E}_{\mathbb{Q}_\theta}(e^{zY}) = e^{\Gamma(\theta+z) - \Gamma(\theta)}$$

per cui

$$\mathbb{E}_{\mathbb{Q}_\theta}(Y) = \Gamma'(\theta) \qquad Var_{\mathbb{Q}_\theta}(Y) = \Gamma''(\theta).$$

Alcuni esempi.

- **Esponenziale $\mathcal{E}(\lambda)$.** Abbiamo

$$\Gamma(\theta) = \log\left(\frac{\lambda}{\lambda - \theta}\right) \mathbb{1}_{\theta<\lambda} + \infty \mathbb{1}_{\theta \geq \lambda} \quad Y \sim_{\mathbb{Q}_\theta} \mathcal{E}(\lambda - \theta)$$

- **Poisson $\mathcal{P}(\lambda)$.** Si ha

$$\Gamma(\theta) = \lambda(e^\theta - 1) \qquad Y \sim_{\mathbb{Q}_\theta} \mathcal{P}(\lambda e^\theta)$$

- **Gaussiana $\mathcal{N}(0,1)$.**

$$\Gamma(\theta) = \frac{\theta^2}{2} \qquad Y \sim_{\mathbb{Q}_\theta} \mathcal{N}(\theta, 1).$$

9.7 Accelerazione con Quasi Monte-Carlo

Questa sezione è legata alla sottosezione 3.5 in cui i metodi di Quasi Monte-Carlo sono presentati per il calcolo di integrali.

Referenze: [2] [1]

Discrepanza di successioni.[5] Misura quantitativa di come una successione $(x_m)_m$ riempie il cubo $[0,1]^d$. Confronta la proporzione di punti nel rettangolo $[0,y_1]\times\cdots\times[0,y_d]$: (punti nel rettangolo - area del rettangolo) e prendiamo il sup sui rettangoli nel cubo:

$$D_M((x_m)_m) = \sup_{y\in[0,1]^d}\left|\frac{1}{M}\sum_{m=1}^{M}\mathbb{1}_{x_m\in[0,y_1]\times\cdots\times[0,y_d]} - y_1\cdot\cdots\cdot y_d\right|.$$

Invece di utilizzare una successione aleatoria per stimare la media, usiamo una successione deterministica a bassa discrepanza $(u_m)_m$:

$$I_M^{QMC} = \frac{1}{M}\sum_{m=1}^{M} f(u_m).$$

L'idea è di riempire il cubo in modo uniforme per ottenere una migliore integrazione numerica.[6] Ci sono varie successioni che servono a questo:

- **Traslazione irrazionale del toro.** $u_m = (Frac(m\sqrt{2}), Frac(m\sqrt{3}))$ (dove si può estendere in dimensione d aggiungendo le parti frazionarie di m per la radice quadrata dei primi d numeri primi).
- **Korobov.** Con a ed n primi tra loro si pone

$$u_i = \frac{i}{n}\begin{pmatrix} 1 \\ a \mod n \\ a^2 \mod n \\ \cdots \\ a^{d-1} \mod n \end{pmatrix} \mod 1$$

Proposizione 9.5. *Per ogni $x\in[0,1]^d$ e $U\sim\mathcal{U}([0,1]^d)$ vale $Frac(U+x)\sim U$.*

Proposizione 9.6. *Se $u_1,\ldots,u_n$ è una successione deterministica, allora lo stimatore $\frac{1}{n}\sum_{i=1}^{n} f(Frac(U+u_i))$ dell'integrale $\int_{[0,1]^d} f(u)du$ è non distorto.*

Teorema 9.7. *Lo stimatore*

$$I_{M,n} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{n}\sum_{i=1}^{n} f(Frac(U_m+u_i))\right)$$

è

[5] Piccola precisione di notazione: [2] chiama discrepanza ciò che [1] chiama *discrepanza stellata*.

[6] In [1], il teorema di Kosma-Hlawka permettono di stimare esplicitamente l'errore tra l'integrale e la sua approssimazione per mezzo della discrepanza (stellata) e della *variazione di Hardy-Krause*.

:) *non distorto:* $\mathbb{E}(I_{M,n}) = \int_{[0,1]^d} f(u)du$*;*

- *ha varianza* $Var(I_{M,n}) = \frac{1}{M} Var\left(\frac{1}{n}\sum_{i=1}^{n} f(Frac(U+u_i))\right)$
- *l'intervallo di fiducia al* 95% *asintotico è della forma*

$$\left[I_{M,n} - 1.96\frac{\sigma_{M,n}}{\sqrt{M}}, I_{M,n} + 1.96\frac{\sigma_{M,n}}{\sqrt{M}}\right]$$

con $\sigma_{M,n}$ *scarto tipo empirico degli stimatori interni alla somma (*$\frac{1}{n}\sum_{i=1}^{n} f(Frac(U+u_i))$*).*

10 Metamodelli

Referenze: [1]

Un métamodèle est une fonction numérique rapide à évaluer capable de mimiquer la sortie d'un code coûteux et complexe ou le résultat d'une expérience. Un métamodèle possède en général des paramètres libres de calibration ou d'ajustement.

Se l'esperimento ha un solo tipo di input (i *punti di funzionamento* x), il metamodello ha in più i parametri di calibrazione β. In generale abbiamo:

- una variabile reale di output $Y_r(x)$;
- una variabile osservata $Y_o(x) = Y_r(x) + \varepsilon_m(x)$ dove $\varepsilon_m(x)$ è un errore di misura modellizzato da v.a. i.i.d.;
- una variabile di output del metamodello $Y_m(x) = f(x, \beta)$.

Nel seguito consideriamo un metamodello lineare (affine in β):

$$f(x, \beta) = f_0(x) + \sum_{i=1}^{p} \beta_i f_i(x)$$

dove $f_i(x)$ sono delle funzioni deterministe date. Gli obiettivi sono:

- **Calibrazione.** Identificare l'insieme di parametri β che permette di aggiustare meglio il metamodello alle osservazioni.
- **Predizione.** Usare il metamodello per predire il valore di Y_r o Y_o in un nuovo punto di funzionamento.
- **Validazione.** Verificare e testare la qualità dei risultati.

Notazioni. Per essere concisi, utilizziamo le seguenti notazioni:

$$y_o = (Y_o(x^{(j)}))_{1 \le j \le n} \qquad y_0 = (f_0(x^{(j)}))_{1 \le j \le n} \qquad H_{ji} = f_i(x^{(j)})$$

10.1 Calibrazione

Cerchiamo il "migliore" β nel senso dei minimi quadrati:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y_o - y_0 - H\beta\|^2 .$$

Nel caso "facile", $n \ge p$ e H di rango massimo p, abbiamo che la matrice $H^T H$ è inversibile ed abbiamo una forma esplicita per $\hat{\beta}$:

$$\hat{\beta} = (H^T H)^{-1} H^T (y_o - y_0).$$

Proposizione 10.1. ***Se*** *esiste β_v tale che $Y_r(x) = f(x, \beta_v)$ e se $\varepsilon_m \sim \mathcal{N}(0, \sigma_m^2)$ (con σ_m^2 noto), allora*

$$\hat{\beta} \sim \mathcal{N}(\beta_v, \sigma_m^2 (H^T H)^{-1}).$$

Osservazione :) La proposizione precedente implica che se $\det(\sigma_m^2 (H^T H)^{-1}) \ll 1$ allora abbiamo identificato con grande precisione β_v per mezzo dello stimatore $\hat{\beta}$.

:) Possiamo costuire un ellissoide di fiducia per β_v:

$$\mathbb{P}\left(\beta_v \in \{\beta \in \mathbb{R}^p (\beta - \hat{\beta})^T H^T H (\beta - \hat{\beta}) \leq \sigma_m^2 \chi_p^2 (1-\alpha)\}\right) = 1 - \alpha$$

con $\chi_p^2(\alpha)$ l'α-quantile della legge χ^2, poiché:

$$\frac{(\beta - \hat{\beta})^T H^T H (\beta - \hat{\beta})}{\sigma_m^2} \sim \chi_p^2.$$

10.2 Caso σ_m^2 noto

10.2.1 Predizione

Posto $h^{(0)} = (f_i(x^{(0)})_{1 \leq i \leq p}$ il vettore di suscettibilità nel nuovo punto $x^{(0)}$, viene naturale utilizzare

$$\hat{Y}(x^{(0)}) = f_0(x^{(0)}) + (h^{(0)})^T \hat{\beta} \sim \mathcal{N}(Y_r(x^{(0)}), \sigma_m^2 q_{pred}^2)$$

dove $q_{pred}^2 = (h^{(0)})^T (H^T H)^{-1} h^{(0)}$.

:) Abbiamo un intervallo di fiducia per $Y_r(x^{(0)})$:

$$\mathbb{P}\left(Y_r(x^{(0)}) \in [\hat{Y}(x^{(0)}) - s_{1-\alpha/2} \sigma_m q_{pred}, \hat{Y}(x^{(0)}) + s_{1-\alpha/2} \sigma_m q_{pred}]\right) = 1 - \alpha$$

con s_α l'α-quantile per la legge $\mathcal{N}(0,1)$.

10.2.2 Validazione

Il test dei residui. Definiamo il vettore dei residui $\hat{\varepsilon} = (Y_o(x^{(j)} - f(x^{(j)}, \hat{\beta}))_{1 \leq j \leq n}$.

- In forma matriciale $\hat{\varepsilon} = (I - H(H^T H)^{-1} H^T)(y_o - y_0) = P(y_o - y_0)$. Osserviamo che P è la matrice di proiezione ortogonale su $\mathrm{Im}(H)^\perp$.
- Se esiste β_v tale che $Y_r(x) = f(x, \beta_v)$ e se $\varepsilon_m \sim \mathcal{N}(0, \sigma_m^2)$, $\hat{\varepsilon} = P\varepsilon_m \sim \mathcal{N}(0, \sigma_m^2 P)$, dunque per il teorema di Cochran 2.3 $\frac{1}{\sigma_m^2} \|\hat{\varepsilon}\|^2 \sim \chi_{n-p}^2$.
- **Se** conosciamo σ_m^2 abbiamo

$$\mathbb{P}\left(\frac{1}{(n-p)\sigma_m^2} \|\hat{\varepsilon}\|^2 \in \left[\frac{\chi_{n-p}^2(\alpha/2)}{n-p}, \frac{\chi_{n-p}^2(1-\alpha/2)}{n-p}\right]\right) = 1 - \alpha.$$

Esempio: Test dei residui per $\alpha = 0.05$.

- Calcoliamo $\frac{1}{(n-p)\sigma_m^2} \|\hat{\varepsilon}\|^2$;
- Se $\frac{1}{(n-p)\sigma_m^2} \|\hat{\varepsilon}\|^2 \in \left[\frac{\chi^2_{n-p}(0.025)}{n-p}, \frac{\chi^2_{n-p}(0.975)}{n-p}\right]$ allora i dati sono compatibili col metamodello;
- altrimenti i dati non sono compatibili.

10.2.3 Validazione LOO

Il principio è di predire il valore di ogni osservazione tenendo conto di tutte le altre: $y_{-i} = (y_o)_{j \neq i}$. In questo modo posiamo comparare la predizione ottenuta senza utilizzare la vera osservazione con il valore osservato. Ciò permette di stimare l'errore commesso per il metodo di predizione su una nuova entrata. Questto procedimento prende il nome di **L**eave **O**ne **O**ut (LOO).

Possiamo calcolare media e varianza della predizione:

$$\begin{aligned} \hat{Y}[y_{-i}](x^{(i)}) =& f_0(x^{(i)}) + (h^{(i)})^T \hat{\beta}[y_{-i}] \\ Var(Y[y_{-i}](x^{(i)})) =& \sigma_m^2 + \sigma_m^2((h^{(i)})^T (H_{-i}^T H_{-i})^{-1} h^{(i)}) \end{aligned}$$

con $\hat{\beta}[y_{-i}] = (H_{-i}^T H_{-i})^{-1} H_{-i}^T (y_{-i} - y_{-i,0})$. Otteniamo una formula per l'errore LOO:

$$\varepsilon_{LOO} = (Y_o(x^{(i)}) - \hat{Y}[y_{-i}](x^{(i)}))_{i=1}^n = \left(\frac{(P(y_o - y_0))_i}{P_{ii}}\right)_{i=1}^n$$

dove $P = I - H(H^T H)^{-1} H^T$. Analogamente il vettore delle varianze di predizione è:

$$V_{LOO} = \left(Var(Y[y_{-i}](x^{(i)}))\right)_{i=1}^n = \left(\frac{\sigma_m^2}{P_{ii}}\right)_{i=1}^n.$$

Il vettore LOO permette di dare una rappresentazione dell'errore del metamodello affidabile e che non necessita di ulteriori chiamate al codice. Inoltre permette di stimare la varianza dell'errore di misura del modello se questa non è nota. Dettagli nella prossima sezione.

10.3 Caso σ_m^2 sconosciuto

10.3.1 Un primo stimatore

Grazie alla sezione sul test dei residui, sappiamo che uno stimatore per σ_m^2 è $\hat{\sigma}^2 = \frac{1}{n-p} \|\hat{\varepsilon}\|^2$. Tale stimatore è:

:) convergente;

:) non distorto;

- con varianza $Var(\hat{\sigma}^2) = \frac{2\sigma_m^4}{n-p}$.

Abbiamo allora:

Proposizione 10.2.

$$\hat{\beta} \sim \mathcal{N}(\beta_v, \sigma_m^2 (H^T H)^{-1}) \qquad e \qquad \hat{\sigma}^2 \sim \frac{\sigma_m^2}{n-p} \chi^2_{n-p}$$

sono indipendenti.

Osservazione :) Con la proposizione precedente possiamo costruire un ellissoide di fiducia per β_v:

$$\mathbb{P}\left(\beta_v \in \{\beta \in \mathbb{R}^p (\beta - \hat{\beta})^T H^T H (\beta - \hat{\beta}) \leq p\hat{\sigma}^2 f_{p,n-p}(1-\alpha)\}\right) = 1 - \alpha$$

con $f_{p,n-p}(\alpha)$ l'α-quantile della legge di Fisher $\mathcal{F}_{p,n-p}$, poiché:

$$\frac{(\beta - \hat{\beta})^T H^T H (\beta - \hat{\beta})}{p\hat{\sigma}^2} \sim \frac{\frac{\chi^2_p}{p}}{\frac{\chi^2_{n-p}}{n-p}} = \mathcal{F}_{p,n-p}.$$

10.3.2 Predizione

Sempre facendo l'ipotesi (forte) che esiste un β_v tale che $Y_r(x) = f(x, \beta_v)$, possiamo proporre come predizione per $Y_r(x^{(0)}$

$$\hat{Y}(x^{(0)}) = f_0(x^{(0)}) + (h^{(0)})^T \hat{\beta} \sim \mathcal{N}(Y_r(x^{(0)}), \sigma_m^2 q^2_{pred})$$

dove $q^2_{pred} = (h^{(0)})^T (H^T H)^{-1} h^{(0)}$. Questo verifica

$$\frac{\hat{Y}(x^{(0)}) - Y_r(x^{(0)})}{\hat{\sigma}^2 q_{pred}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} = \mathcal{T}_{n-p} \qquad \text{(legge di Student)} .$$

10.3.3 Un altro stimatore

Utilizzando il metodo di validazione incrociata (LOO) possiamo stimare σ_m^2 studiando l'errore di predizione

$$\hat{\mathcal{E}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\varepsilon^2_{LOO,i}}{V_{LOO,i}} = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_o(x^{(i)}) - \hat{Y}[y_{-i}](x^{(i)}))^2}{Var(Y[y_{-i}](x^{(i)}))}$$

L'idea è che $\mathcal{E}$ debba essere vicino a 1 quando il metamodello è corretto, perché in tal caso i rapporti che compaiono nella somma sono delle v.a. gaussiane standard.

(!) Il vettore normalizzato $Y_{LOO} = \left(\frac{\varepsilon_{LOO,i}}{V_{LOO,i}^{1/2}}\right)_{i=1}^{n}$ è gaussiano ma le sue componenti **non** sono indipendenti.

Poiché $\hat{\mathcal{E}}$ dipende da σ_m^2:

$$\hat{\mathcal{E}}(\sigma_m^2) = \frac{1}{n\sigma_m^2}\sum_{i=1}^{n} P_{ii}\varepsilon_{LOO,i}^2$$

uno stimatore per σ_m^2 è

$$\sigma_{CV}^2 = \arg\min_{\sigma^2}\left|\hat{\mathcal{E}} - 1\right| = \frac{1}{n}\sum_{i=1}^{n} P_{ii}\varepsilon_{LOO,i}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_{ii}}\hat{\varepsilon}_i^2.$$

Tale stimatore è:

:) convergente;

:) non distorto;

- con varianza $Var(\hat{\sigma}^2) = \frac{2\sigma_m^4}{n^2}\sum_{i,j=1}^{n}\frac{P_{ij}^2}{P_{ii}P_{jj}}$.

Confronto tra i due stimatori. Per p fisso e n grande, le due varianze sono entrambe equivalenti a $\frac{2\sigma_m^4}{n}$ e i due stimatori sono quindi abbastanza simili.

10.4 Con ipotesi meno forti

L'objectif général est comme dans les sections précédentes de calibrer les paramètres d'un métamodèle et de qualifier ses prédictions en prenant en compte des résultats expérimentaux ou des résultats d'un code complexe et coûteux. La différence est qu'on ne va pas faire l'hypothèse qu'il existe un β_v tel que $Y_r(x) = f(x;\beta_v)$ avec f de la forme $f(x,\beta) = f_0(x) + \sum_{i=1}^{p}\beta_i f_i(x)$.

La variabile di output del metamodello è allora $Y_m(x) = f(x,\beta) + Z_m(x)$, dove $Z_m(x)$ è un errore di modelllo, realizzazione di un processo aleatorio gaussiano di media nulla, stazionario e funzione di covarianza $\mathbb{E}(Z_m(x)Z_m(x')) = C_m(x-x')$. Queste ipotesi sono giustificate nel modo seguente:

- non sappiamo a priori se l'errore del modello è positivo o negativo $\Rightarrow$ prendiamo la media nulla;
- non possiamo dire che l'errore è maggiore in una certo insieme dei punti di funzionamento $\Rightarrow$ prendiamo un processo stazionario;
- abbiamo voglia di calcolare rapidamente $\Rightarrow$ prendiamo un processo gaussiano.

Anche in questo caso ci occupiamo di *calibrazione*, *predizione* e *validazione* del metamodello.

 Ce type de modélisation de l'erreur de modèle par un processus aléatoire gaussien a d'abord été introduite en géostatistique (elle porte alors le nom de krigeage). En effet, bien qu'il n'y ait qu'une seule planète terre, on peut représenter une caractéristique physique du sous-sol (par exemple, la concentration en un minerai) comme une réalisation d'un processus gaussien spatial, qu'on apprend à connaître au fur et à mesure qu'on creuse des puits et des mines. Dans notre cadre, il existe une seule fonction réelle, qu'on représente comme une réalisation d'un processus gaussien et qu'on apprend à connaître au fur et à mesure des simulations ou des expériences menées.

10.4.1 Distribuzione a priori delle osservazioni

Notazioni. Ricordiamo le notazioni per il vettore delle osservazioni, il risultato del metamodello centrale e il tensore di suscettività:

$$y_o = (Y_o(x^{(j)}))_{1\leq j\leq n} \qquad y_0 = (f_0(x^{(j)}))_{1\leq j\leq n} \qquad H_{ji} = f_i(x^{(j)})$$

e aggiungiamo la matrice di covarianza

$$R_{ij} = C_m(x^{(i)} - x^{(j)}) + \sigma_m^2 \delta_{ij}.$$

La funzione C_m. La funzione di covarianza C_m descrive la forma delle fluttuazioni delle variabili di output che non sono incluse e descritte dalla combinazione lineare delle funzioni f_j, cioè l'errore di modello. Tale descrizione è fatta in termine di iperparametri che caratterizzano la statistica dellle fluttuazioni. Tipicamente si hanno fino a tre iperparametri:

- σ^2 lo scarto quadratico medio;
- l_c la lunghezza di correlazione;
- ν il parametro di regolarità.

In base alla scelta di questi tre iperparametri abbiamo scelte diverse per la funzione C_m:

1. modello *nugget*:
$$C_m(x - x') = \sigma^2 \delta(x - x')$$
che rende la matrice R diagonale: $R_{ij} = (\sigma_m^2 + \sigma^2)\delta_{ij}$. In questo caso (l'inverso di) σ^2 misura la fiducia nel modello: se σ^2 è piccolo allora il metamodello riproduce bene la realtà per una scelta opportuna di β.

2. modello gaussiano:
$$C_m(x - x') = \sigma^2 \exp\left(-\frac{\|x - x'\|}{l_c^2}\right).$$
Qui l_c descrive la portata dell'errore di modello: due punti a distanza minore che l_c hanno degli errori di modello correlati (quindi vicini), mentre due

punti a distanza maggiore che l_c hanno errori di modello indipendenti. Il modello *nugget* è un modello gaussiano dove $l_c \to 0$. Possiamo ovviamente generalizzare, scegliendo un parametro $l_{c,i}$ per ogni dimensione dello spazio:

$$C(x - x') = \sigma^2 \exp\left(- \sum_{i=1}^{r} \frac{(x_i - x_i')^2}{l_{c,i}^2} \right).$$

3. modello di Matérn:

$$C(x - x') = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{2\sqrt{\nu} \|x - x'\|}{l_c} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu} \|x - x'\|}{l_c} \right),$$

dove K_ν è la funzione di Bessel modificata[7]. L'iperparametro $\nu \in [1/2, +\infty)$ descrive la regolarità dell'errore: il modello gaussiano corrisponde a $\nu \to +\infty$, nel quale $Z_m(x)$ è q.o. di classe C^∞; quando $\nu = 1/2$ troviamo il modello esponenziale continuo ma non derivabile q.o.: $C_m(x - x') \sim \exp\left(-\sqrt{2} \|x - x'\| / l_c\right)$. Osserviamo che quando $\nu = n + 1/2$ per $n \in \mathbb{N}$, la funzione di covarianza è prodotto di un polinomio di grado n e di una funzione esponenziale. Il modello di Matérn è utilizzato perché la densità di potenza spettrale (cioè la trasformata di Fourier di $C_m(x)$) ha una forma semplice (se tralasciamo la costante davanti):

$$\hat{C}_m(k) = \frac{2^d \pi^{d/2} \Gamma(\nu + d/2)(2\nu)^\nu}{\Gamma(\nu) l_c^{2\nu}} \left(\frac{2\nu}{l_c^2} + \|k\|^2 \right)^{-\nu - d/2} \approx \left(\frac{2\nu}{l_c^2} + \|k\|^2 \right)^{-\nu - d/2}.$$

10.4.2 Calibrazione

Sappiamo che la legge di y_o condizionalmente a β è gaussiana di media $y_0 + H\beta$ e matrice di covarianza R.Resta a determinare la legge condizionale di β dato y_o.

Distribuzione a priori dei parametri gaussiana. Abbiamo una distribuzione a priori di β che è $\mathcal{N}(\beta_{prior}, Q_{prior})$. Allora la distribuzione a posteriori di β è gaussiana $\mathcal{N}(\beta_{post}, Q_{post})$ con

$$\begin{aligned} \beta_{post} &= \beta_{prior} + (Q_{prior}^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} (y_o - y_0 - H\beta_{prior}) \\ Q_{post} &= (Q_{prior}^{-1} + H^T R^{-1} H)^{-1} \end{aligned}$$

- Se $\|R\| \to 0$ (cioè le osservazioni sono quasi perfette) allora possiamo considerare che la distribuzione a posteriori non dipende più da quella a priori:

$$\begin{aligned} \beta_{post} &\simeq (H^T R^{-1} H)^{-1} H^T R^{-1} (y_o - y_0) \\ Q_{post} &\simeq (H^T R^{-1} H)^{-1} \end{aligned}$$

[7] Fonction de Bessel modifiée sur Wikipedia.

- Se $\|R\| \to \infty$ (cioè le osservazioni sono cattive) allora semplicemente

$$\beta_{post} \simeq \beta_{prior} \qquad Q_{post} \simeq Q_{prior}.$$

Dai punti precedenti vediamo che R è una misura della fiducia che mettiamo nelle osservazioni.

- La varianza generalizzata a posteriori è sempre più piccola della varianza generalizzata a priori:

$$\det(Q_{post}) \leq \det(Q_{prior})$$

e vale anche $Q_{post} \leq Q_{prior}$ nel senso di matrici positive.

La riduzione di varianza ci dice che abbiamo (quasi) certamente trovato i parametri migliori, cioè il miglior metamodello possibile tra quelli della classe dei metamodelli affini, ma questo può non essere un buon metamodello in generale.

- La covarianza a posteriori per β non dipende dai valori osservati. Possiamo allora considerare la varianza generalizzata in funzione dei punti di funzionamento:

$$\mathcal{E}(x^{(1)}, \ldots, x^{(n)}) = \det\left(Q_{prior}^{-1} + H^T R^{-1} H\right)^{-1}$$

e minimizzarla per ridurre il più possibile l'incertezza sui parametri. Questo procedimento è detto *D-ottimale*.

- I calcoli di inversione delle matrici sono facili quando ci sono più osservabili che parametri ($n \geq p$). Può succedere che no sia il caso o che R non sia invertibile, o mal condizionata.

Senza distribuzione a priori dei parametri. Non avere informazioni a priori su β è come avere Q_{prior} molto grande, quindi possiamo ritrovare la distribuzione a posteriori di β sapendo y_o passando al limite $\left\|Q_{prior}^{-1}\right\| \to 0$ nelle formule del caso precedente:

$$\begin{aligned} \beta_{post} &= (H^T R^{-1} H)^{-1} H^T R^{-1} (y_o - y_0) \\ Q_{post} &= (H^T R^{-1} H)^{-1} \end{aligned}$$

- La matrice $(H^T R^{-1} H)^{-1}$ è da interpretare come la pseudo-inversa di $H^T R^{-1} H$.

10.4.3 Predizione

Distribuzione a priori dei parametri gaussiana.

Notazioni. Poniamo

$$h_i = f_i(x)\ i = 1, \ldots, p \qquad r_j = C_m(x - x^{(j)})\ j = 1, \ldots, n.$$

La distribuzione di $Y_o(x)$ condizionata alle osservazioni y_o è gaussiana di media

$$\hat{Y}_{post}(x) = f_0(x) + h^T \beta_{post} + r^T R^{-1}(y_o - y_0 - H\beta_{post})$$

e varianza

$$Var_{post}(Y(x)) = C_m(0) + \sigma_m^2 - \begin{pmatrix} h \\ r \end{pmatrix}^T \begin{pmatrix} -Q_{prior}^{-1} & H^T \\ H & R \end{pmatrix}^{-1} \begin{pmatrix} h \\ r \end{pmatrix}.$$

La varianza si può riscrivere come:

$$Var_{post}(Y(x)) = C_m(0) + \sigma_m^2 - r^T R^{-1} r + (h - H^T R^{-1} r)^T (H^T R^{-1} H + Q_{prior}^{-1})^{-1} (h - H^T R^{-1} r),$$

o anche

$$Var_{post}(Y(x)) = C_m(0) + \sigma_m^2 - h^T Q_{prior} h - (r + HQ_{prior} h)^T (HQ_{prior} H^T + R)^{-1} (r + HQ_{prior} h).$$

- La media a posteriori contiene due termini: la predizione della combinazione lineare delle funzioni f_j col parametro β aggiustato, e un termine correttivo che tiene conto dell'errore di modello nel nuovo punto.

:) Possiamo calcolare un intervallo di fiducia per $Y_o(x)$ condizionato a y_o:

$$\mathbb{P}\left(Y_o(x) \in [\hat{Y}_{post}(x) - Var_{post}(Y(x))^{1/2} z_\alpha, \hat{Y}_{post}(x) + Var_{post}(Y(x))^{1/2} z_\alpha]\right) = 1 - \alpha$$

con z_α l'$1 - \alpha/2$-quantile della distribuzione $\mathcal{N}(0, 1)$.

:) Poiché la varianza a priori di $Y_o(x)$ è $C_m(0) + \sigma_m^2 + h^T Q_{prior} h$, constatiamo che facendo delle osservazioni riduciamo la varianza.

- La varianza a posteriori per $Y_o(x)$ non dipende dai valori delle osservazioni y_o. Allora prendendo una densità di probabilità $w(x)$ centrata su una zona di funzionamento che ci interessa, possiamo considerare la funzione

$$\mathcal{E}_x(x^{(1)}, \ldots, x^{(n)}) = \begin{pmatrix} h(x) \\ r(x) \end{pmatrix}^T \begin{pmatrix} -Q_{prior}^{-1} & H^T \\ H & R \end{pmatrix}^{-1} \begin{pmatrix} h(x) \\ r(x) \end{pmatrix}$$

che misura la riduzione dell'errore di predizione locale in x. Con questa, definiamo la funzione che misura la riduzione dell'errore globale:

$$\mathcal{E}_w(x^{(1)}, \ldots, x^{(n)}) = \int \mathcal{E}_x(x^{(1)}, \ldots, x^{(n)}) w(x) dx$$
$$= tr\left(\begin{pmatrix} -Q_{prior}^{-1} & H^T \\ H & R \end{pmatrix}^{-1} \int \begin{pmatrix} h(x)h^T(x) & h(x)r^T(x) \\ r(x)h^T(x) & r(x)r^T(x) \end{pmatrix} w(x) dx\right).$$

Possiamo quindi massimizzarla per trovar i punti di funzionamento dove fare le esperienze per ridurre il più possibile l'incertezza nella predizione. Questo procedimento è detto *I-ottimale*.

- Se cerchiamo la legge a posteriori di $Y_r(x)$ senza errore di misura, basta eliminare σ_m^2 nella formula di $Var_{post}(Y(x))$.

Generalizzazione multidimensionale. Possiamo generalizzare il risultato sulla distribuzione a posteriori di $Y(x)$ nel caso di più output. La distribuzione congiunta degli output è gaussiana e caratterizzata dalla media e matrice di covarianza a posteriori degli singoli output. Nel caso di due punti x_a e x_b, posti $h_i^a = f_i(x_a)$ e $r_j^a = C_m(x_a - x^{(j)})$ (e analogamente per x_b), la matrice di covarianza a priori per l'output (y_a, y_b) è

$$Cov_{prior} = \begin{pmatrix} C_m(0) & C_m(x_a - x_b) \\ C_m(x_a - x_b) & C_m(0) \end{pmatrix} + \sigma_m^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Allora la media a posteriori è il vettore:

$$\begin{pmatrix} \hat{Y}_{post}(x_a) \\ \hat{Y}_{post}(x_b) \end{pmatrix} = \begin{pmatrix} f_0(x_a) \\ f_0(x_b) \end{pmatrix} + \begin{pmatrix} (h^a)^T \\ (h^b)^T \end{pmatrix} \beta_{post} + \begin{pmatrix} (r^a)^T \\ (r^b)^T \end{pmatrix} R^{-1}(y_o - y_0 - H\beta_{post})$$

e la matrice di covarianza a posteriori:

$$Cov_{post} = Cov_{prior} - \begin{pmatrix} h^a & h^b \\ r^a & r^b \end{pmatrix}^T \begin{pmatrix} -Q_{prior}^{-1} & H^T \\ H & R \end{pmatrix}^{-1} \begin{pmatrix} h^a & h^b \\ r^a & r^b \end{pmatrix}.$$

Per simulare una realizzazione di $Y_r(x)$ o $Y_o(x)$ condizionata alle osservazioni y_o nei punti $(x^{(j)})_{1 \le j \le n}$ ci sono essenzialmente due metodi.

Choleski.

a) prendiamo la griglia dei punti di funzionamento $\{x, x \in \mathcal{D}\}$;

b) calcoliamo la media $m = \{\hat{Y}_{post}(x) | x \in \mathcal{D}\}$ e la matrice di covarianza a posteriori $\{Cov_{post}(x, x') | x, x' \in \mathcal{D}\}$;

c) calcoliamo S, radice quadrata della matrice di covarianza (tramite l'algoritmo di Choleski);

d) preso $z \sim \mathcal{N}(0, Id)$ poniamo $y = m + Sz$.

e) y ha la stessa legge di $(Y_r(x))_{x \in \mathcal{D}}$ condizionato alle osservazioni y_o.

Simulazione condizionata. L'idea di questo metodo è che se $\tilde{y}(x)$ segue la legge a priori, allora $y(x) = \hat{Y}_{post}(x) - \hat{\tilde{Y}}_{post}(x) + \tilde{y}(x)$ segue la legge a posteriori sapendo y_o.

a) prendiamo la griglia dei punti di funzionamento $\{x, x \in \mathcal{D}\}$;

b) prendiamo $\{\tilde{y}(x) | x \in \mathcal{D}\}$, cioè:

i) preso $\beta \sim \mathcal{N}(\beta_{prior}, Q_{prior})$;

ii) simuliamo $\{z_m(x)|x \in \mathcal{D}\}$ secondo la legge $\mathcal{N}(0, (C_m(x-x'))_{x,x' \in \mathcal{D}})$;

iii) simuliamo $\{\varepsilon_m(x)|x \in \mathcal{D}\}$ secondo la legge $\mathcal{N}(0, \sigma_m^2 Id)$;

iv) poniamo $\tilde{y}(x) = f_0(x) + \sum_{i=1}^p f_i(x)\beta_i + z_m(x) + \varepsilon_m(x)$ per $x \in \mathcal{D}$.

c) calcoliamo $\beta_{post} = \beta_{prior} + (Q_{prior}^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}(y_o - y_0 - H\beta_{prior})$;

d) calcoliamo $\tilde{\beta}_{post} = (Q_{prior}^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1}(\tilde{y}_o - y_0 - H\beta_{prior})$ con le osservazioni $\tilde{y}_o$;

e) poniamo

$$\begin{aligned} y(x) &= \tilde{y}(x) + \hat{Y}_{post}(x) - \hat{\tilde{Y}}_{post}(x) \\ &= \tilde{y}(x) + h(x)^T(\beta_{post} - \tilde{\beta}_{post}) + r(x)^T R^{-1}(y_o - \tilde{y}_o - H(\beta_{post} - \tilde{\beta}_{post})). \end{aligned}$$

Questo vettore ha la stessa legge di $(Y_r(x))_{x \in \mathcal{D}}$ condizionato alle osservazioni y_o.

Senza distribuzione a priori dei parametri. Possiamo ritrovare la distribuzione a posteriori di $Y_o(x)$ condizionata alle osservazioni y_o ponendo $Q_{prior}^{-1} = 0$ nelle formule del caso precedente:

$$\beta_{post} = (H^T R^{-1} H)^{-1} H^T R^{-1}(y_o - y_0)$$

La distribuzione di $Y_o(x)$ condizionata alle osservazioni y_o è gaussiana di media

$$\hat{Y}_{post}(x) = f_0(x) + h^T \beta_{post} + r^T R^{-1}(y_o - y_0 - H\beta_{post})$$

e varianza

$$Var_{post}(Y(x)) = C_m(0) + \sigma_m^2 - \begin{pmatrix} h \\ r \end{pmatrix}^T \begin{pmatrix} 0 & H^T \\ H & R \end{pmatrix}^{-1} \begin{pmatrix} h \\ r \end{pmatrix}.$$

La varianza si può riscrivere come:

$$Var_{post}(Y(x)) = C_m(0) + \sigma_m^2 - r^T R^{-1} r + (h - H^T R^{-1} r)^T (H^T R^{-1} H)^{-1} (h - H^T R^{-1} r),$$

10.4.4 Validazione

L'ordre de grandeur de l'erreur de modèle est caractérisée par la variance $C_m(0)$, la portée de l'erreur de modèle est caractérisée par la largeur de la fonction de corrélation $x \mapsto C_m(x)/C_m(0)$, et la régularité spatiale de l'erreur de modèle est caractérisée par la régularité de la fonction de covariance en 0. On peut vérifier si le modèle est acceptable par un procédé de validation croisée. L'idée est de se servir de toutes les observations, sauf une, pour calibrer le modèle, puis d'utiliser les formules de prédiction pour voir ce que prédit le modèle au point manquant, et enfin de comparer avec la valeur observée qu'on avait mise de côté.

Leave One Out

1. on calcule la prédiction de la valeur de la sortie en $x^{(i)}$ en termes de moyenne $\hat{Y}_{post}[y_{-i}](x^{(i)})$ et de variance $Var_{post}(Y[y_{-i}](x^{(i)}))$ en utilisant seulement les observations formées du vecteur y_o duquel on a retiré le i-eme résultat;

2. on regarde l'erreur de prédiction globale normalisée de type Leace One Out (LOO):
$$\mathcal{E} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{Y}_{post}[y_{-i}](x^{(i)}))^2}{Var_{post}(Y[y_{-i}](x^{(i)}))}.$$

3. $\mathcal{E}$ doit être proche de 1 pour que le modèle puisse être déclaré bon. Sinon, on peut faire essentiellement deux choses :

   (a) essayer de changer le métamodèle, dans le sens où on cherche une fonction de covariance C_m mieux adaptée.

   (b) essayer de changer le problème et considérer la fonction $1/Y(x)$ ou la fonction $\log Y(x)$, qui peut être mieux adaptée à une approche par processus gaussiens que $Y(x)$.

:) le calcul des erreurs de prédiction globale de type LOO peut être effectué très rapidement:
$$y_i - \hat{Y}_{post}[y_{-i}](x^{(i)}) = \frac{1}{(\tilde{R}^-)_{ii}}(\tilde{R}^-\tilde{y})_i$$
$$Var_{post}(Y[y_{-i}](x^{(i)})) = \frac{1}{(\tilde{R}^-)_{ii}}$$
dove

   i) $\tilde{R}^- = (HQ_{prior}H^T + R)^{-1}$ e $\tilde{y} = y_o - y_0 - H\beta_{prior}$ se la distribuzione a priori dei parametri è gaussiana.

   ii) $\tilde{R}^- = R^{-1} - R^{-1}H(H^TRH)^{-1}H^TR^{-1}$ e $\tilde{y} = y_o - y_0$ se non c'è una distribuzione a priori dei parametri.

10.4.5 Selezionare un modello

Possiamo provare a stimare C_m sulle stesse osservazioni, attraverso un procedimento parametrico, cioè ricercando la funzione in una classe parametrata $\{C_m(x) = \sigma^2 C_0(x,\theta) \mid \sigma^2 \in \mathbb{R}^+, \theta \in \Theta\}$. Ad esempio con $\theta = l_c$ abbiamo il modello gaussiano e $\theta = (l_c, \nu)$ quello di Matérn.

Distribuzione a priori dei parametri gaussiana. Se β segue una distribuzione a priori $\mathcal{N}(\beta_{prior}, Q_{prior})$, supponendo $R = \sigma^2 R_0$, lo stimatore di massima verosimiglianza è
$$\hat{\sigma^2} = \arg\min_{\sigma^2 \in \mathbb{R}^+} \{n\log(\sigma^2) + (y_o - y_0 - H\beta_{prior})^T V(\sigma^2)(y_o - y_0 - H\beta_{prior})\}$$

dove $V(\sigma^2) = (HQ_{prior}H^T + \sigma^2 R_0)^{-1} = \sigma^{-2}R_0^{-1} - \sigma^{-2}R_0^{-1}H(H^T R_0^{-1}H + \sigma^2 Q_{prior}^{-1})^{-1}H^T R_0^{-1}$. Supponendo invece $R = \sigma^2 R_0(\theta)$, lo stimatore di massima verosimiglianza è

$$(\hat{\sigma^2}, \hat{\theta}) = \arg \min_{\sigma^2 \in \mathbb{R}^+, \theta \in \Theta} \left\{ n \log(\sigma^2) + \log(\det R_0(\theta)) + (y_o - y_0 - H\beta_{prior})^T V(\sigma^2, \theta)(y_o - y_0 - H\beta_{prior}) \right\}$$

con $V(\sigma^2, \theta) = (HQ_{prior}H^T + \sigma^2 R_0(\theta))^{-1} = \sigma^{-2}R_0^{-1}(\theta) - \sigma^{-2}R_0^{-1}(\theta)H(H^T R_0^{-1}(\theta)H + \sigma^2 Q_{prior}^{-1})^{-1}H^T R_0^{-1}(\theta)$.

Senza distribuzione a priori dei parametri. Supponendo $R = \sigma^2 R_0$, lo stimatore di massima verosimiglianza è

$$\hat{\sigma^2} = \frac{1}{n}(y_o - y_0)^T W_{yy}(y_o - y_0)$$

dove $W_{yy} = R_0^{-1} - R_0^{-1}H(H^T R_0^{-1}H)^{-1}H^T R_0^{-1}$. Supponendo $R = \sigma^2 R_0(\theta)$, lo stimatore di massima verosimiglianza è

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \{ n \log((y_o - y_0)^T W_{yy}(\theta)(y_o - y_0)) + \log \det R_0(\theta) \}$$
$$\hat{\sigma^2} = \frac{1}{n}(y_o - y_0)^T W_{yy}(\hat{\theta})(y_o - y_0)$$

dove $W_{yy}(\theta) = R_0^{-1}(\theta) - R_0^{-1}(\theta)H(H^T R_0^{-1}(\theta)H)^{-1}H^T R_0^{-1}(\theta)$. Osserviamo che nel caso $R_0 = Id$, $\hat{\sigma^2} = \frac{1}{n}\|P(y_o - y_0)\|^2$ con $P = I - H(H^T H)^{-1}H^T$, che è l'estimatore dei residui (a meno di una costante moltiplicativa $n/(n-p)$).

Selezione per validazione incrociata. Un metodo un po' più generale, per una classe di modelli di matrici di covarianza $\{R(\theta) \mid \theta \in \Theta\}$: 

1. on calcule la prédiction de la valeur de la sortie en $x^{(i)}$ en termes de moyenne $\hat{Y}_{post}[y_{-i}, \theta](x^{(i)})$ et de variance $Var_{post}(Y[y_{-i}, \theta](x^{(i)}))$ en utilisant seulement les observations formées du vecteur y_o duquel on a retiré le i-eme résultat et la matrice de covariance $R(\theta)$ de laquelle on a retiré la i-ème ligne et la i-ème colonne;

2. On cherche le minimum en θ de l'erreur globale de prédiction de type LOO :

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{Y}_{post}[y_{-i}, \theta](x^{(i)}))^2 \right\}$$

ou de la version normalisée :

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{Y}_{post}[y_{-i}, \theta](x^{(i)}))^2}{Var_{post}(Y[y_{-i}, \theta](x^{(i)}))} - 1 \right| \right\}.$$

11 Roba che per ora non so dove mettere

Referenze: [2]

Copule. Misura intrinseca della dipendenza senza utilizzare le distribuzioni marginali. Invarianti sotto trasformazioni crescenti del vettore aleatorio iniziale.

Esempi di copule.

- **Indipendente**: $C(u_1, \dots, u_d) = u_1 \dots u_d$.
- **Co-monotona**: $C^+(u_1, \dots, u_d) = \min(u_1, \dots, u_d)$.
- **Gaussiana**: Se $X \sim \mathcal{N}(0, K)$ con $K_i i = Var(X_i) = 1$, allora il vettore $U = (F_{\mathcal{N}(0,1)}(X_1), \dots, F_{\mathcal{N}(0,1)}(X_d))$ ha copula gaussiana di covarianza K.

Stimatore di polinomi in $\mathbb{E}(X)$ Se X è limitata, ad un campione i.i.d $(X_m)_m$ associamo una replica indipendente $(X'_m)_m$. Allora

- $I_1 = \left(\frac{1}{M}\sum_{m=1}^M X_m\right)^2$ è uno stimatore distorto di $\mathbb{E}(X)^2$;
- $I_2 = \frac{1}{M}\sum_{m=1}^M X_m X'_m$ è uno stimatore non distorto di $\mathbb{E}(X)^2$.

X e Y v.a. reali di quadrato integrabile $\Rightarrow XY$ integrabile X e Y v.a. reali integrabili e indipendenti $\Rightarrow XY$ integrabile

$MVar(I_1) \approx 4\mathbb{E}(X)^2 Var(X)$ per $M \to +\infty$ $MVar(I_2) = Var(XX') = (\mathbb{E}(X^2) + \mathbb{E}(X)^2)Var(X)$ dunque il secondo estimatore ha una varianza più piccola se $\mathbb{E}(X^2) + \mathbb{E}(X)^2 \leq 4\mathbb{E}(X)^2 \iff Var(X) \leq 2\mathbb{E}(X)^2$.

Generalizzando: se $f(x) = \sum_{k=1}^n a_k x^k$ e cerchiamo uno stimatore non distorto per $f(\mathbb{E}(X))$ possiamo creare n copie inidipendenti del campione i.i.d $(X_i)_{1\leq i\leq M}$ e considerare

$$I_2^{(M)} = \frac{1}{M}\sum_{m=1}^M \left(\sum_{k=1}^n a_k \prod_{i=1}^k X_m^{(i)}\right)$$

che è uno stimatore integrabile, convergente q.c. e non distorto. In alternativa

- possiamo sorteggiare v.a. indipendenti per ogni grado del polinomio;
- randomizzare i gradi con probabilità $\mathbb{P}(K = k) = \frac{|a_k|}{\sum_{i=1}^n |a_i|}$ per evitare di sorteggiare n repliche per ogni simulazione m.

Simulare $\mathbb{E}(f(O, \mathbb{E}(R|O)))$**.** Monte-Carlo di Monte-Carlo: simulazione in due passaggi:

- speranza esterna: M simulazioni indipendenti O_i

- speranza condizionale interna: per ogni O_m N simulazioni indipendenti di $R|O_m$ per avere
$$\mathbb{E}(R|O_m) \approx \frac{1}{N}\sum_{k=1}^{N} R_{m,k}$$
- approssimare con
$$\mathbb{E}(f(O, \mathbb{E}(R|O))) \approx \frac{1}{M}\sum_{m=1}^{M} f\left(O_m, \frac{1}{N}\sum_{k=1}^{N} R_{m,k}\right).$$

Teorema 11.1. *Sotto ipotesi che* $\sup_x \mathbb{E}(|R - \mathbb{E}(R|O = x)|^4) < +\infty$ *se:*

- $f\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ *è limitata e lipschitziana nella seconda variabile allora*
$$\mathbb{E}(|I_{M,N} - I|)^2)^{1/2} \leq O\left(\frac{1}{\sqrt{N}} + \frac{1}{\sqrt{M}}\right)$$
- *se* $f \in C^1$ *nella seconda variabile con derivata limitata e lipschitziana, allora*
$$\mathbb{E}(|I_{M,N} - I|)^2)^{1/2} \leq O\left(\frac{1}{N} + \frac{1}{\sqrt{M}}\right)$$

Complessità Il costo di calcolo è $O(NM)$. Se ε è la tolleranza cercata, abbiamo:

- $M = N = \varepsilon^{-2}$ per un costo $O(\varepsilon^{-4})$
- $M = N^2 = \varepsilon^{-2}$ per un costo $O(\varepsilon^{-3}$.

Il caso polinomiale Se $f(x, y) = \sum_{k=1}^{n} a_k(x)y^k$ sorteggiamo un campione di taglia M per O, poi per ogni O_m generiamo n copie i.i.d. di $R|O_m$. A quel punto lo stimatore
$$I^{(M)} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{n} a_k(O_m)\prod_{i=1}^{k} R_{m,i}$$
è integrabile, convergente q.c. e non distorto per $\mathbb{E}(f(O, \mathbb{E}(R|O)))$. Il costo di calcolo è $O(M)$ quindi per una tolleranza ε possiamo prendere $M = \varepsilon^{-2}$ ed avere una velocità di convergenza migliore del caso generale.

11.1 Regressione non parametrica

Poniamo $\mathcal{M}(x) = \mathbb{E}(R|O = x)$ la funzione di regressione e sia $\mathcal{M}_M$ la sua approssimazione su uno spazio finito dimensionale.

:(La legge di (R, O) non è sempre esplicita ma facile a simulare.

dobbiamo conoscere $\mathcal{M}$ ovunque

- Non possiamo supporre che le funzioni di base siano ortonormali rispetto al prodotto scalare di $L_2(\mathbb{P} \circ O^{-1})$ perché numericamente, essendo la distribuzione di O sconosciuta o difficile da trattare, non si può utilizzare Gram-Schmidt.
- L'approssimazione migliora all'aumentare della dimensione K dello spazio. Per evitare l'*over-fitting* supponiamo $M \geq K$.
- $\mathcal{M}(O) = \arg\min_{\mathcal{M}_O} \mathbb{E}(R - \mathcal{M}_O)^2$ (dove $\mathcal{M}_O$ è una v.a. $\sigma(O)$-misurabile). In particolare scegliamo
$$\mathcal{M}_M = \alpha^M \cdot \phi$$
che minimizzi il criterio quadratico $\frac{1}{M}\sum_{m=1}^{M}(R^{(m)} - \varphi(O^{(m)}))^2$. Notiamo che i coefficienti sono aleatori (dipendono dalla simulazione) empirici:
$$\alpha^M := \arg\min_{\alpha \in \mathbb{R}^K} \frac{1}{M}\sum_{m=1}^{M}(R^{(m)} - \alpha \cdot \phi(O^{(m)}))^2 \quad \mathcal{M}_M(O) = \alpha^M \cdot \phi$$
Ciò conduce infine ad avere lo stimatore
$$\mathbb{E}(f(O, \mathbb{E}(R|O))) \approx \frac{1}{M}\sum_{m=1}^{M} f(O^{(m)}, \mathcal{M}_M(O^m))$$
approssimando dunque una funzione determinista con una aleatoria.
- A causa di possibili collinearità tra le funzioni di base

11.1.1 Caso $M = +\infty$

I coefficienti ottimali sono quindi
$$\alpha^* := \arg\min_{\alpha \in \mathbb{R}^K} \mathbb{E}(R - \sum_{k=1}^{K} \alpha_k \phi_k(O))^2$$
soluzione delle equazioni normali
$$A\alpha = \mathbb{E}(R\phi) \qquad \text{con } A_{ij} = \mathbb{E}(\phi_i(O), \phi_j(O)) \quad (\mathbb{E}(R\phi))_i = \mathbb{E}(R\phi_i(O)).$$

:/ Se scegliamo le ϕ_i ortonormali in $L_2(O)$ la matrice A diventa l'identità ma quando $M < +\infty$ i coefficienti non sono più quelli giusti: $\alpha^M \neq \widehat{\mathbb{E}(R\phi)}^M$.

:/ In molti casi la legge di O non è esplicita.

:(Anche se A non è degenere, $\hat{A}^M$ può esserlo.

Come scegliere lo spazio di approssimazione:

1. **Polinomi globali**.

:) Adatta a funzioni molto regolari

:/ Convergenza lenta in caso di singolarità

:/ Stima dell'errore difficile

:(I polinomi non ??? sono densi in $L_2(O)$.

2. **Polinomi locali.**

   - L'idea è di dividere il dominio in ipercubi, su ognuno dei quali scegliamo un polinomio a d variabili di grado al più k (in pratica risolviamo tanti problemi più piccoli con meno coefficienti)
   - In totale ci sono $2H/\Delta)^d$ ipercubi se il dominio è $[-H, H[^d$ e scegliamo un passo Δ per il lato di un ipercubo.
   - La dimensione dello spazio di approssimazione è $K = (2H/\Delta)^d(k+1)^d$.

   :) Il costo numerico è basso poiché approssimiamo su ipercubi disgiunti con polinomi di grado piccolo.

   :) Per $H \to +\infty$ e $\Delta \to 0$ abbiamo convergenza per funzioni continue (a tratti)

   :) La velocità di convergenza è data da

   $$\inf_{\varphi} |\varphi - \mathcal{M}|_{L_2(\mu)} \leq c\Delta^k$$

   se $\mathcal{M} \in C^k$ e O ha legge μ e momenti esponenziali, con $H \sim c\log(1/\Delta)$.

11.1.2 Caso $M < +\infty$

I coefficienti ottimali sono calcolati con una decomposizione SVD.

Entropia.

Definizione Sia μ una misura di probabilità su $\mathbb{R}^d$. Definiamo l'entropia di una funzione misurabile $f\colon \mathbb{R}^d \to \mathbb{R}_+$:

$$\mathrm{Ent}_\mu(f) := \mathbb{E}_\mu(f\log(f)) - \mathbb{E}_\mu(f)\log(\mathbb{E}_\mu(f)).$$

dove $\mathbb{E}_\mu(f) = \int_{\mathbb{R}}^d f(y)\mu(dy)$.

Proprietà dell'entropia. È facile verificare le seguenti proprietà dell'entropia:

- **positività:** $\mathrm{Ent}_\mu(f) \geq 0$;
- **omogeneità:** per $\lambda > 0$ vale $\mathrm{Ent}_\mu(\lambda f) = \lambda\mathrm{Ent}_\mu(f)$;
- **caratterizzazione:** $\mathrm{Ent}_\mu(f) = \sup\{\mathbb{E}_\mu(fg) : \mathbb{E}_\mu(e^g) = 1\}$.

Diciamo che la misura μ verifica *la disuguaglianza di Sobolev logaritmica* con costante $C_\mu \in (0, +\infty)$ per la classe di funzioni $\mathcal{A}$ se

$$\mathrm{Ent}_\mu(f^2) \leq C_\mu \mathbb{E}_\mu(|\nabla f|^2) \quad \forall f \in \mathcal{A}.$$

Teorema 11.2. *Sia μ una misura che soddisfi una disuguaglianza di Sobolev logaritmica con costante C_μ. Allora per ogni $f \in C_c^2(\mathbb{R}^d, \mathbb{R})$ con $|\nabla f| \leq 1$ vale:*

$$\mathbb{E}(e^{\lambda f(Y)}) := \mathbb{E}_\mu(e^{\lambda f}) \leq e^{\lambda \mathbb{E}_\mu(f) + \frac{C_\mu \lambda^2}{4}} =: e^{\lambda \mathbb{E}(f(Y)) + \frac{C_\mu \lambda^2}{4}} \quad \forall \lambda \in \mathbb{R}.$$

Corollario 11.3 (Disuguaglianza di concentrazione)**.** *Sia $Y \in \mathbb{R}^d$ un vettore aleatorio di legge μ che verifichi una disuguaglianza di Sobolev logaritmica di costante $C_\mu > 0$. Allora per ogni funzione $f \colon \mathbb{R}^d \to \mathbb{R}$ lipschitziana vale*

$$\mathbb{P}(|f(Y) - \mathbb{E}(f(Y))| > \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{C_\mu |f|_{Lip}^2}} \quad \forall \varepsilon \geq 0$$

Lemma 11.4. *Sia $n \geq 2$. Se ogni μ_i $1 \leq i \leq n$ soddisfa la disuguaglianza di Sobolev logaritmica con costante C_{μ_i}, allora la misura prodotto $\mu^{\otimes n}$ soddisfa ancora la disuguaglianza di Sobolev logaritmica con costante $C_{\mu^{\otimes n}} = \max_{1 \leq i \leq n} C_{\mu_i}$.*

Corollario 11.5. *Se (Y_i) i.i.d. di legge μ che soddisfa una disuguaglianza di Sobolev logaritmica di costante C_μ allora la legge di $(Y_1, \ldots, Y_n)$ soddisfa la disuguaglianza di Sobolev logaritmica con la stessa costante C_μ.*

Teorema 11.6. *La misura gaussiana in dimensione 1 soddisfa la disuguaglianza di Sobolev logaritmica con costante 2. Per il lemma di tensorizzazione la costante è la stessa per la misura gaussiana in dimensione d qualunque.*

Teorema 11.7 (Concentrazione gaussiana)**.** *Siano $(Y_1, \ldots, Y_M)$ vettori aleatori indipendenti di legge gaussiana d dimensionale (centrata e ridotta). Allora per ogni $f \colon \mathbb{R}^d \to \mathbb{R}$ lipschitziana:*

$$\mathbb{P}\left(\left|\frac{1}{M}\sum_{m=1}^{M} f(Y_m) - \mathbb{E}(f(Y))\right| > \varepsilon\right) \leq 2e^{-\frac{M\varepsilon^2}{2|f|_{Lip}^2}}.$$

Teorema 11.8 (Disuguaglianza di Borell)**.** *Sia $(Y_1, \ldots, Y_d)$ un vettore gaussiano centrato d dimensionale. Allora*

$$\mathbb{P}\left(\left|\max_{1 \leq i \leq d} Y_i - \mathbb{E}(\max_{1 \leq i \leq d} Y_i)\right| > \varepsilon\right) \leq 2e^{-\frac{\varepsilon^2}{2\sigma^2}} \quad \forall \varepsilon \geq 0$$

dove $\sigma^2 = \max_{1 \leq i \leq d} \mathbb{E}(Y_i^2)$.

12 Moto browniano

Referenze: [2]

12.1 Integrale stocastico e formule di Ito

Referenze: [2]

 The concept of an adapted process is essential, for instance, in the definition of the Itō integral, which only makes sense if the integrand is an adapted process.

An **adapted process** (also referred to as a non-anticipating or non-anticipative process) is one that cannot "see into the future".

Interpretazione X is adapted if and only if, for every realisation and every n, X_n is known at time n.

Definizione Un processo X è detto *adatto ? alla filtrazione* $(\mathcal{F}_t)_{t\geq 0}$ se per ogni $t \geq 0$, X_t è $\mathcal{F}_t$-misurabile.

Esempi 

- Take the natural filtration $\mathcal{F}^X$, where $\mathcal{F}_t^X$ is the σ-algebra generated by the pre-images $X_s^{-1}(B)$ for Borel subsets $B \subset \mathbb{R}$ and times $0 \leq s \leq t$. Then X is automatically $\mathcal{F}^X$-adapted. Intuitively, the natural filtration contains "total information" about the behaviour of X up to time t.
- This offers a simple example of a non-adapted process $X\colon [0,2] \times \Omega \to \mathbb{R}$: set $\mathcal{F}_t$ to be the trivial σ-algebra $\{, \Omega\}$ for times $0 \leq t < 1$, and $\mathcal{F}_t = \mathcal{F}_t^X$ for times $1 \leq t \leq 2$. Since the only way that a function can be measurable with respect to the trivial σ-algebra is to be constant, any process X that is non-constant on $[0,1]$ will fail to be $\mathcal{F}$-adapted. The non-constant nature of such a process "uses information" from the more refined "future" σ-algebras $\mathcal{F}_t$, $1 \leq t \leq 2$.

Un **tempo di arresto**, conosciuto anche come tempo di Markov, è uno specifico tipo di "tempo casuale", il cui valore dipende solo dagli eventi successi prima o nell'istante stesso. Ad esso può essere associato una regola di arresto, ovvero una regola per definire il tempo d'arresto.

Rispetto a una sequenza di variabili aleatorie $X_1, X_2, \ldots$ un tempo di arresto T è una variabile aleatoria con la proprietà che per ogni t l'evento $\{T = t\}$ dipende solo dalle variabili $X_1, X_2, \ldots, X_t$.

Definizione τ è un tempo d'arresto per la filtrazione $(\mathcal{F}_t)_{t\geq 0}$ se per ogni $t \geq 0$, l'evento $\{\tau \leq t\} \in \mathcal{F}_t$.

Esempi

- Il primo tempo per cui un moto Browniano tocca lo 0 ($\inf\{t \geq 0 \mid W_t = 0\}$)è un tempo d'arresto (perché dipende solamente dagli eventi passati).
- L'ultima volta che il Browniano tocca lo 0 prima del tempo $T = 1$ non è un tempo d'arresto (perché dipende da un evento futuro).

13 Markov Processes

Referenze: Bayesian Estimation & Stochastic Opt - Vikram Krishnamurthy 2019 [3]

Definizione A discrete time process x_k is Markov on $\mathcal{X}$ if for any $S \subseteq \mathcal{X}$

$$\mathbb{P}(x_{n+1} \in S \mid x_1, \dots, x_n) = \mathbb{P}(x_{n+1} \in S \mid x_n).$$

We are given an initial distribution $\int_S \pi_0(x)dx = \mathbb{P}(x_0 \in S)$.

The Strong Markov property is:

$$\mathbb{P}(x_\tau \mid x_{k-1}, x_{k-2}, \dots) = \mathbb{P}(x_\tau \mid x_{k-1})$$

for any stopping time $\tau \geq k$.

We know that an iid process has no memory:

$$\mathbb{P}(x_k \mid x_{k-1}, x_{k-2}, \dots) = \mathbb{P}(x_k \mid x_{k-1})$$

There is a martingale representation of Markov chains:

$$x_{k+1} = P^T x_k + M_{k+1} \qquad \mathbb{E}(M_{k+1} \mid X_1, \dots, X_k) = 0$$

where M_k is a martingale increment, i.e., $Z_k = \sum_{n=1}^{k} M_n$ is a martingale process:

$$\mathbb{E}(Z_{k+1} \mid Z_1, \dots, Z_k) = Z_k.$$

Martingales are more general than iid.

Riferimenti bibliografici

[1] Josselin Garnier. *Gestion des incertitudes et analyse de risque MAP568.* École Polytechnique, 2017.

[2] Emmanuel Gobet. *Monte-Carlo methods and stochastic processes: from linear to non-linear.* Chapman and Hall/CRC, 2016.

[3] Vikram Krishnamurthy. *Partially observed Markov decision processes.* Cambridge University Press, 2016.

Elenco delle figure

Elenco delle tabelle

Elenco dei Teoremi

Elenco delle Proposizioni

Elenco dei Corollari

Elenco dei Lemmi

Elenco delle Definizioni