Superfici di risposta

0.1 Introduzione

In ambito ingegneristico il numero di valutazioni della funzione obiettivo, in un problema di ottimizzazione, è fortemente limitato da tempo e costi. Spesso i metodi esistenti per l'ottimizzazione richiedono più valutazioni della funzione di quelle che possono essere agevolmente calcolate, per questa ragione è necessario ricorrere a dei modelli matematici che approssimano il problema. L'implementazione di superfici di risposta richiede la conoscenza della funzione in relativamente pochi punti dello spazio e permette di estrapolarne il valore in tutti gli altri punti ottenendo così benefici notevoli in termini di tempo.

Il metodo studiato in questa tesi è il **kriging**. In particolare sono state implementate e confrontate due metodologie: l' **OK** (Ordinary Kriging) e il **DACE** (Design and Analysis of Computer Experiments).

Kriging è il termine usato nella geostatistica per indicare il migliore predittore lineare di funzioni dello spazio. Tale metodo è stato sviluppato dall'ingegnere minerario Sud Africano D.G.Krige (1951). È un metodo largamente usato in geologia, idrologia, monitoraggio ambientale e altri campi per l'interpolazione di dati nello spazio.

Ci sono tre caratteristiche comuni spesso osservate nei dati indicizzati attraverso coordinate spaziali:

- lenta variabilità su larga scala dei valori misurati;
- irregolarità su piccola scala;

- affinità delle misure in punti vicini.

È possibile, tenendo conto di queste caratteristiche, modellare il fenomeno attraverso una funzione aleatoria (**SRF** *Spatial Random Field*). Sotto certe ipotesi semplificative questo modello produce degli stimatori lineari che permettono una facile implementazione della previsione. Questi stimatori, tra cui il *kriging*, forniscono sia una previsione che un errore standard della previsione nei punti non campionati. Ciò permette di costruire una mappa dei valori previsti e del livello di incertezza di tali valori.

0.2 Kriging

Consideriamo un fenomeno deterministico $y(\mathbf{x})$. Dato un insieme di quantità osservate vogliamo stimare tale fenomeno in un nuovo punto. Si tratta di un tipico problema di *inferenza statistica*.

Per poter prevedere il valore incognito consideriamo la funzione deterministica come risultato di un processo stocastico $Y(\mathbf{x})$, il cui modello matematico è un modello di regressione lineare:

$$Y(\mathbf{x}) = \sum_{h=1}^{k} \beta_h f_h(\mathbf{x}) + \epsilon(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \beta + \epsilon(\mathbf{x})$$

in cui ogni $f_h(\mathbf{x})$ è una funzione lineare o non lineare di $\mathbf{x}$, i coefficienti β_h sono incogniti e $\epsilon(\mathbf{x})$ è una funzione stocastica con media nulla e covarianza incognita. Essendo, per ipotesi, la varianza dell'errore σ^2 la stessa per tutte le osservazioni e ricordando che la funzione di correlazione è pari a

$$Corr(\mathbf{w}, \mathbf{x}) = \frac{Cov(\mathbf{w}, \mathbf{x})}{\sigma_w \sigma_x} ,$$

la covarianza tra $\epsilon(\mathbf{w})$ e $\epsilon(\mathbf{x})$ risulta essere

$$Cov(\mathbf{w}, \mathbf{x}) = \sigma^2 Corr(\mathbf{w}, \mathbf{x}).$$

Assunto questo modello per il processo stocastico, note n osservazioni $\mathbf{Y} =$

$(Y(\mathbf{x}_1),\ldots,Y(\mathbf{x}_n))^T$, vogliamo prevedere il valore $Y(\mathbf{x}_0)$. Per fare questo è necessario prima determinare i coefficienti β_h.

Per ottenere uno stimatore di β minimizziamo l'errore quadratico medio (**MSE** *Mean Squared Error*) della previsione. Consideriamo un predittore di tipo lineare :

$$\lambda_0 + \lambda^T \mathbf{Y}$$

soggetto alla condizione di *unbiasedness* ovvero di *non distorsione* del parametro:

$$E[\lambda_0 + \lambda^T \mathbf{Y}] = E[Y(\mathbf{x}_0)] \text{ per ogni } \beta.$$

Tale condizione implica che sia:

$$\lambda_0 + \lambda^T \mathbf{F}\beta = \mathbf{f}(\mathbf{x}_0)^T\beta \text{ per ogni } \beta$$

dove indichiamo con $\mathbf{F}$ la matrice hxn $(\mathbf{f}(\mathbf{x}_1)\mathbf{f}(\mathbf{x}_2)...\mathbf{f}(\mathbf{x}_n))^T$. Dalla precedente segue che:

$$\lambda_0 = (-\lambda^T \mathbf{F} + \mathbf{f}(\mathbf{x}_0)^T)\beta \text{ per ogni } \beta,$$

per cui deve essere:

$$\lambda_0 = 0 \qquad \text{e} \qquad \mathbf{F}^T\lambda = \mathbf{f}(\mathbf{x}_0). \tag{1}$$

Il nostro obiettivo è minimizzare $E[Y(\mathbf{x}_0) - \lambda^T \mathbf{Y}]^2$ vincolato dall'equazione 1. Se λ è la soluzione del problema di minimizzazione vincolata allora $\lambda^T \mathbf{Y}$ è il **miglior predittore lineare non distorto** (**BLUP** *Best Linear Unbiased Predictor*) per $Y(\mathbf{x}_0)$.

Per risolvere questo problema osserviamo innanzi tutto che esiste un predittore lineare non distorto (**LUP** *Linear Unbiased Predictor*) se e solo se $\mathbf{f}(\mathbf{x}_0) \in C(\mathbf{F}^T)$ con C spazio delle colonne della matrice $\mathbf{F}^T$. Da questo momento consideriamo valida questa ipotesi.

Se λ soddisfa alla $\mathbf{F}^T\lambda = \mathbf{f}(\mathbf{x}_0)$ allora qualunque predittore lineare non distorto può essere scritto come $(\lambda + \nu)^T\mathbf{Y}$ con $\mathbf{F}^T\nu = 0$. $\lambda^T\mathbf{Y}$ è un *BLUP* se $(\mathbf{K}\lambda - \mathbf{k})^T\nu = 0$ per ogni ν tale che $\mathbf{F}^T\nu = 0$ con $\mathbf{K} = Cov(\mathbf{Y}, \mathbf{Y}^T)$ e $\mathbf{k} = Cov(\mathbf{Y}, Y(\mathbf{x}_0))$ o, equivalentemente, se esiste un vettore μ (moltiplicatore di Lagrange) tale che $\mathbf{K}\lambda - \mathbf{k} = \mathbf{F}\mu$.

Esprimendo in forma matriciale le precedenti equazioni, deve essere quindi:

$$\begin{bmatrix} \mathbf{K} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} \mathbf{k} \\ \mathbf{f}(\mathbf{x}_0) \end{pmatrix} \tag{2}$$

per un certo μ, dove $\mathbf{0}$ è una matrice di zeri. Questo sistema di equazioni lineari ha una soluzione se e solo se $\mathbf{f}(\mathbf{x}_0) \in C(\mathbf{F}^T)$, ipotesi che abbiamo fatto precedentemente. Se $\mathbf{K}$ e $\mathbf{F}$ sono matrici a rango pieno, ovvero matrici non singolari per le quali il determinante è diverso da zero, allora si ottiene:

$$\begin{pmatrix} \lambda \\ \mu \end{pmatrix} = \begin{bmatrix} \mathbf{K} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{k} \\ \mathbf{f}(\mathbf{x}_0) \end{pmatrix} \tag{3}$$

da cui:

$$\lambda = (\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{K}^{-1})\mathbf{k} + \mathbf{K}^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{f}(\mathbf{x}_0). \tag{4}$$

Il **miglior predittore lineare non distorto** è quindi:

$$\lambda^T\mathbf{Y} = (\mathbf{k}^T\mathbf{K}^{-1}(\mathbf{Y} - \mathbf{F}\beta) + \mathbf{f}(\mathbf{x}_0)^T\beta \tag{5}$$

e l'**errore quadratico medio** risulta pari a

$$k_0 - \mathbf{k}^T\mathbf{K}^{-1}\mathbf{k} + \gamma^T(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F})^{-1}\gamma \tag{6}$$

dove $\gamma = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}^T\mathbf{K}^{-1}\mathbf{k}$ e $k_0 = \mathbf{K}(\mathbf{x}_0, \mathbf{x}_0)$.

Il *miglior predittore lineare non distorto* in geostatistica è detto **kriging**. Esistono diversi tipi di Kriging:

- per una generica $\mathbf{f}(\mathbf{x})$ il kriging è detto *universale*;
- se $\mathbf{f}(\mathbf{x}) \equiv 1$, cosicchè la media del processo stocastico è una costante incognita, il kriging è detto *ordinario*;
- se si assume media nulla il kriging è detto *semplice*.

0.3 OK: Ordinary Kriging

Nel **kriging Ordinario** si va a stimare il valore vero della funzione, che è incognito, attraverso una combinazione lineare pesata dei punti campionati. I pesi dipendono da tali punti, quindi variano se usiamo un diverso insieme di punti campionati.

In pratica tale metodo consiste nella risoluzione del sistema:

$$\mathbf{Cw} = \mathbf{C}_0,$$

che in forma estesa diventa:

$$\begin{bmatrix} \tilde{C}_{11} & \tilde{C}_{12} & ... & \tilde{C}_{1n} & 1 \\ \tilde{C}_{21} & \tilde{C}_{22} & ... & \tilde{C}_{2n} & 1 \\ ... & & & & \\ \tilde{C}_{n1} & \tilde{C}_{n2} & ... & \tilde{C}_{nn} & 1 \\ 1 & 1 & ... & 1 & 0 \end{bmatrix} \begin{pmatrix} w_1 \\ w_2 \\ ... \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} \tilde{C}_{10} \\ \tilde{C}_{20} \\ ... \\ \tilde{C}_{n0} \\ 1 \end{pmatrix} \tag{7}$$

con $\tilde{C}$ funzione di covarianza nota in quanto scelta a priori sulla base dei punti campionati. Calcolati così i pesi w_i possiamo determinare la stima della previsione:

$$\hat{y} = \sum_j w_j y_j$$

e la varianza dell'errore:

$$\tilde{\sigma}_R^2 = \tilde{\sigma}^2 - (\sum_j w_j \tilde{C}_{j0} + \mu)$$

dove $\tilde{\sigma}^2$ è la varianza della previsione.

Se $\hat{y}_i$ è la stima dell'i-esimo punto, definiamo l'errore r come la differenza tra il valore stimato e il valore vero in tale punto:

$$r = \hat{y}_i - y_i.$$

L'errore medio su k punti è quindi

$$m_R = \frac{1}{k} \sum_{i=1}^{k} r_i$$

e la varianza dell'errore è data da

$$\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(r_i - m_R)^2 = \frac{1}{k}\sum_{i=1}^{k}[(\hat{y}_i - y_i) - \frac{1}{k}\sum_{i=1}^{k}(\hat{y}_i - y_i)]^2.$$

Se l'errore medio è nullo allora

$$\sigma^2 = \frac{1}{k}\sum_{i=1}^{k}(\hat{y}_i - y_i)^2.$$

La stima di y_i sarà prossima al valore vero se l'errore medio è nullo e la deviazione standard dell'errore è la più piccola possibile. Il valore vero della funzione nel punto i-esimo è però incognito, pertanto non possiamo calcolare l'errore medio nè la varianza dell'errore e siamo costretti a ricavare una soluzione probabilistica del problema.

Consideriamo un processo stocastico $Y(\mathbf{x})$. Nel modello **OK** si fanno due assunzioni sul processo stocastico: la stazionarietà e l'ergodicità.

Se un processo Y(x) è *stazionario* si ha, indicando con $\mu(x)$ la media, $Corr(x,h)$ la funzione di correlazione e $Cov(x,h)$ la funzione di covarianza:

$$\mu(x) = \mu$$

$$Corr(x,h) = Corr(0,h) = Corr(h)$$

$$Cov(x,h) = R(h) - \mu^2 = Cov(h)$$

dove h è la distanza tra i punti x e $x+h$. Nel nostro caso si richiede che il processo stocastico sia stazionario *in senso largo*, quindi si ha più semplicemente che la speranza matematica è costante e la funzione di correlazione dipende solo dalla distanza tra i punti:

$$E[Y(x)] = \mu$$

$$Corr(x,h) = Corr(0,h) = Corr(h).$$

L'ipotesi di *ergodicità* ci permette di derivare le proprietà statistiche del processo aleatorio. In generale, per fare questo, occorre fare ricorso ad un numero sufficientemente elevato di realizzazioni. Se abbiamo n osservazioni Y_i del generico processo $Y(x)$ la media $\mu(x)$ può essere stimata attraverso la $\hat{\mu}(x) = \frac{\sum_i Y_i}{n}$ solo se n è sufficientemente grande. I *processi ergodici* sono una particolare classe di processi per i quali le proprietà statistiche possono essere dedotte a partire da una singola realizzazione del processo, poichè le medie spaziali calcolate su una singola realizzazione e le medie d'insieme coincidono.

Se con $Y(y_1, y_2, \ldots, y_n)$ indichiamo il processo stocastico dove y_i è la variabile aleatoria relativa al punto $\mathbf{x}_i$, la media d'insieme è data da:

$$\begin{aligned} E[Y(y_1, y_2, \ldots, y_n)] &= \\ &= \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} Y(y_1, y_2, \ldots, y_n) \\ &\quad p_{y_1, y_2, \ldots, y_n}(y_1, y_2, \ldots, y_n)\ dy_1\, dy_2 \ldots dy_n \end{aligned} \tag{8}$$

Per una generica realizzazione del processo ha senso considerare Y come funzione delle distanze $h_1, h_2, \ldots, h_{n-1}$:

$$Y(y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n)) = Y(y(\mathbf{x}_1), y(\mathbf{x}_1 + h_1), \ldots, y(\mathbf{x}_1 + h_{n-1})).$$

La media spaziale è data pertanto da:

$$\lim_{\Delta x \to \infty} \frac{1}{\Delta x} \int_{-\Delta x/2}^{+\Delta x/2} Y(y(\mathbf{x}_1), y(\mathbf{x}_1 + h_1), \ldots, y(\mathbf{x}_1 + h_{n-1}))dh_1 dh_2 \ldots dh_{n-1}.$$

Diremo un *processo ergodico* se la media spaziale di ordine n di una realizzazione della funzione è uguale per tutte le realizzazioni (a meno di un sottoinsieme di probabilità nulla) e coincide con la media d'insieme della stessa funzione.

Inoltre osserviamo che un processo è ergodico se:

- è stazionario in senso stretto
- non contiene sottoinsiemi stazionari in senso stretto con probabilità diversa da 0 o 1.

In sostanza ogni realizzazione di un processo ergodico contiene in sè tutta l'informazione possibile sul processo, in quanto una sorgente ergodica produce, nel corso di una realizzazione, tutte le situazioni ed i casi possibili per il processo con una frequenza pari alla probabilità di detti eventi.

Fatte quindi le ipotesi di stazionarietà ed ergodicità sul processo stocastico $Y(\mathbf{x})$ possiamo scrivere:

$$\hat{Y}(\mathbf{x}_0) = \sum_{i=1}^{n} w_i Y(\mathbf{x}_i)$$

$$R(\mathbf{x}_0) = \hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_0),$$

da cui

$$R(\mathbf{x}_0) = \sum_{i=1}^{n} w_i Y(\mathbf{x}_i) - Y(\mathbf{x}_0).$$

Ci assicuriamo che l'errore in qualunque punto abbia speranza matematica nulla, quindi che sia verificata la *condizione di stimatore non distorto.*

$$\begin{aligned} E[R(\mathbf{x}_0)] &= E[\sum_i w_i Y(\mathbf{x}_i) - Y(\mathbf{x}_0)] \\ &= E[\sum_i w_i Y(\mathbf{x}_i)] - E[Y(\mathbf{x}_0)] \\ &= 0. \end{aligned} \tag{9}$$

Per la stazionarietà è:

$$E[Y(\mathbf{x}_0)] = E[Y].$$

Essendo inoltre:

$$E[\sum_i w_i Y(\mathbf{x}_i)] = \sum_i w_i E[Y(\mathbf{x}_i)] = \sum_i w_i E[Y]$$

possiamo scrivere:

$$\sum_i w_i E[Y] = E[Y(\mathbf{x}_0)],$$

da cui è:

$$\sum_i w_i = 1.$$

A questo punto andiamo a minimizzare la varianza dell'errore $R(\mathbf{x}_0)$. Innanzitutto dobbiamo trovare un'espressione per la varianza. La varianza dell'errore è pari:

$$\begin{aligned} Var[R(\mathbf{x}_0)] &= Cov[\hat{Y}(\mathbf{x}_0)\hat{Y}(\mathbf{x}_0)] - Cov[\hat{Y}(\mathbf{x}_0)Y(\mathbf{x}_0)] - \\ &\quad -Cov[Y(\mathbf{x}_0)\hat{Y}(\mathbf{x}_0)] + Cov[Y(\mathbf{x}_0)Y(\mathbf{x}_0)] \\ &= Cov[\hat{Y}(\mathbf{x}_0)\hat{Y}(\mathbf{x}_0)] - 2Cov[\hat{Y}(\mathbf{x}_0)Y(\mathbf{x}_0)] + \\ &\quad +Cov[Y(\mathbf{x}_0)Y(\mathbf{x}_0)]. \end{aligned} \tag{10}$$

Osserviamo che il primo termine $Cov[\hat{Y}(\mathbf{x}_0)\hat{Y}(\mathbf{x}_0)]$ non è niente altro che la covarianza di $\hat{Y}(\mathbf{x}_0)$ con se stesso ovvero la varianza di $\hat{Y}(\mathbf{x}_0) = \sum_i w_i Y(\mathbf{x}_i)$. Ricordando che la varianza di una combinazione lineare pesata è $Var[\sum_{i=1}^n w_i Y_i] = \sum_{i=1}^n \sum_{j=1}^n w_i w_j Cov[Y_i Y_j]$, tale termine può essere scritto come:

$$Cov[\hat{Y}(\mathbf{x}_0)\hat{Y}(\mathbf{x}_0)] = Var[\sum_i w_i Y_i] = \sum_i \sum_j w_i w_j Cov[Y_i Y_j] = \sum_i \sum_j w_i w_j \tilde{C}_{ij}.$$

Per il secondo termine si ha:

$$\begin{aligned} 2Cov[\hat{Y}(\mathbf{x}_0)Y(\mathbf{x}_0)] &= 2Cov[(\sum_i w_i Y_i)Y_0] \\ &= 2E[(\sum_i w_i Y_i)Y_0] - 2E[\sum_i w_i Y_i]E[Y_0] \\ &= 2\sum_i w_i E[Y_i Y_0] - 2\sum_i w_i E[Y_i]E[Y_0] \\ &= 2\sum_i w_i Cov[Y_i Y_0] \\ &= 2\sum_i w_i \tilde{C}_{i0}. \end{aligned} \tag{11}$$

L'ultimo termine della varianza dell'errore, $Cov[Y(\mathbf{x}_0)Y(\mathbf{x}_0)]$, è la varianza di $Y(\mathbf{x}_0)$. Avendo assunto che tutte le variabili aleatorie abbiano la stessa varianza $\tilde{\sigma}^2$, si può scrivere:

$$Cov[Y(\mathbf{x}_0)Y(\mathbf{x}_0)] = \tilde{\sigma}^2.$$

Combinando i termini precedenti otteniamo la seguente espressione per la varianza dell'errore:

$$\tilde{\sigma}_R^2 = \tilde{\sigma}^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \tilde{C}_{ij} - 2\sum_{i=1}^{n} w_i \tilde{C}_{i0}. \tag{12}$$

La precedente espressione è un'equazione in n variabili che dobbiamo minimizzare tenendo conto del vincolo di non distorsione dello stimatore. Questo problema può essere facilmente risolto attraverso il *metodo di Lagrange*. Il vincolo $\sum_i w_i = 1$ può essere equivalentemente scritto come $2(\sum_i w_i - 1) = 0$. La funzione lagrangiana è quindi:

$$\tilde{\sigma}_R^2 = \tilde{\sigma}^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \tilde{C}_{ij} - 2\sum_{i=1}^{n} w_i \tilde{C}_{i0} + 2\mu(\sum_{i=1}^{n} w_i - 1) \tag{13}$$

con μ **moltiplicatore di Lagrange**. La condizione che deve verificarsi è:

$$\begin{cases} \frac{\partial \sigma_R^2}{\partial w_1} = 2\sum_{j=1}^{n} w_j \tilde{C}_{1j} - 2\tilde{C}_{10} + 2\mu = 0 \\ \\ \frac{\partial \sigma_R^2}{\partial w_2} = 2\sum_{j=1}^{n} w_j \tilde{C}_{2j} - 2\tilde{C}_{20} + 2\mu = 0 \\ \dots \\ \frac{\partial \sigma_R^2}{\partial w_i} = 2\sum_{j=1}^{n} w_j \tilde{C}_{ij} - 2\tilde{C}_{i0} + 2\mu = 0 \\ \dots \\ \frac{\partial \sigma_R^2}{\partial w_n} = 2\sum_{j=1}^{n} w_j \tilde{C}_{nj} - 2\tilde{C}_{n0} + 2\mu = 0 \\ \\ \frac{\partial \sigma_R^2}{\partial \mu} = 2\sum_{i=1}^{n} w_i - 2 = 0 \end{cases} \tag{14}$$

che è equivalente a

$$\begin{cases} \sum_{j=1}^{n} w_j \tilde{C}_{ij} + \mu = \tilde{C}_{i0} \text{ per ogni } i = 1, \ldots, n \\ \sum_{i=1}^{n} w_i = 1 \end{cases} \tag{15}$$

e in notazione matriciale:

$$\mathbf{Cw} = \mathbf{C}_0.$$

Determinati i pesi w_i la varianza dell'errore minimizzata è:

$$\tilde{\sigma}_R^2 = \sigma^2 - (\sum_{i=1}^{n} w_i \tilde{C}_{i0} + \mu)$$

e la previsione nel punto x_0:

$$\hat{y}(x_0) = \sum_{i=1}^{n} w_i y_i.$$

In genere quando si estrapola una funzione con il kriging ordinario non si considera tutto il set di allenamento ma solamento i k punti più vicini al punto di estrapolazione. Quindi avremo:

$$\hat{y}(x_0) = \sum_{i=1}^{k-nearest} w_i y_i$$

e

$$\tilde{\sigma}_R^2 = \sigma^2 - (\sum_{i=1}^{k-nearest} w_i \tilde{C}_{i0} + \mu).$$

Quando ricaviamo l'espressione della varianza dell'errore assumiamo che le variabili aleatorie della funzione stocastica abbiano tutte la stessa media e varianza. Queste assunzioni ci permettono di ricavare la seguente relazione tra

la funzione variogramma e la covarianza:

$$\begin{aligned}\gamma_{ij} &= \frac{1}{2}E[(Y_i - Y_j)^2] \\ &= \frac{1}{2}E[Y_i^2] + \frac{1}{2}E[Y_j^2] - E[Y_iY_j] \\ &= E[Y^2] - E[Y_iY_j] \\ &= E[Y^2] - \tilde{m}^2 - E[Y_iY_j] + \tilde{m}^2 \\ &= \tilde{\sigma}^2 - \tilde{C}_{ij}. \end{aligned} \tag{16}$$

La validità di queste relazioni ci consente di esprimere le equazioni del kriging ordinario in termini di funzione variogramma:

$$\begin{cases} \sum_{j=1}^{n} w_j \tilde{\gamma}_{ij} - \mu = \tilde{\gamma}_{i0} & \text{per ogni } i = 1, \ldots, n \\ \sum_{i=1}^{n} w_i = 1. \end{cases} \tag{17}$$

Per completezza ricordiamo che la definizione di funzione variogramma è:

$$\tilde{\gamma}(h) = \frac{1}{2}E[(Y(x) - Y(x+h))^2] = \tilde{C}(0) - \tilde{C}(h).$$

La funzione variogramma ha la stessa forma della funzione di covarianza tranne che è invertita: mentre la covarianza parte da un massimo $\tilde{\sigma}^2$ per $h = 0$ e tende a zero per h crescente, la funzione variogramma parte da zero per $h = 0$ e tende a $\tilde{\sigma}^2$ per h crescente (fig.1).

Osseviamo che i pesi del kriging ordinario e la risultante varianza dell'errore dipendono direttamente dalla scelta della funzione di correlazione e quindi dalla scelta della matrice di covarianza. La scelta del modello di covarianza è una prerogativa del kriging ordinario. Il primo passo è costruire la funzione variogramma attraverso l'interpolazione dei punti campionati. Il variogramma sperimentale è ottenuto attraverso il calcolo della varianza di ogni punto dell'insieme rispetto agli altri e rappresentando graficamente le varianze in funzione della distanza tra i punti. Diverse formule possono essere usate per calcolare la

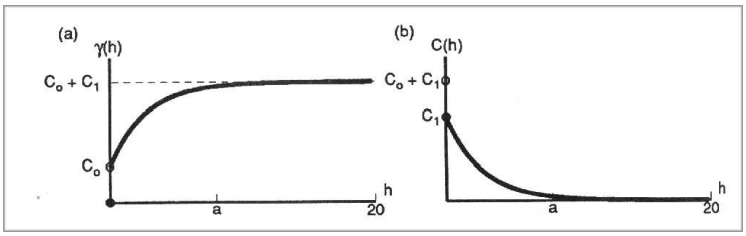

Figura 1: Esempio di una funzione variogramma esponenziale (a) e di una funzione di covarianza esponenziale (b).

varianza, ma normalmente è calcolata come $\frac{1}{2}(y_i - y_j)^2$. Una volta calcolato il variogramma sperimentale il passo successivo è definire una semplice funzione matematica che modella l'andamento del variogramma sperimentale (fig.2). Per avere una e una sola soluzione del sistema 15 dobbiamo assicurarci che la

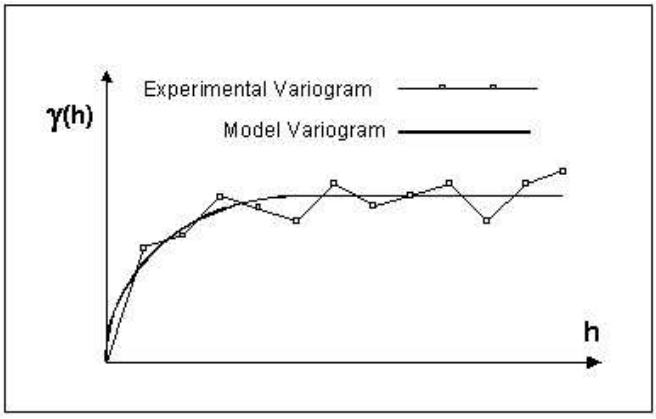

Figura 2: Variogramma sperimentale e modello.

matrice di covarianza sia definita positiva. Garantiamo l'esistenza e l'unicità della soluzione interpolando il variogramma campionato con funzioni che sono

definite positive.

Sebbene l'interpolazione del variogramma campionato sia l'approccio più comune per scegliere il modello di continuità spaziale per il processo aleatorio, non è l'unico e non è necessariamente il migliore. Ci sono molti casi in cui è meglio basare la scelta di tale modello su un'interpretazione più qualitativa. L'esperienza con simili insiemi di dati può spesso essere una guida migliore che non un modello di continuità spaziale ottenuto attraverso pochi campioni disponibili.

In genere quando estrapoliamo una funzione con il kriging ordinario non si considera tutto il set di allenamento ma solamento gli nk punti più vicini al punto di estrapolazione.

0.4 DACE

Supponiamo di aver valutato in n punti un fenomeno deterministico, funzione di k variabili. Indichiamo i punti discreti con $\mathbf{x}(i) = (x_1(i), \ldots, x_k(i))$ e gli associati valori della funzione con $y(i) = y(\mathbf{x}(i))$ con $i = 1, \ldots, n$.

Il modo più semplice di adattare una superficie di risposta a tali dati è la regressione lineare, quindi il modello matematico adottato è lo stesso del kriging ordinario. Le osservazioni sono trattate come se fossero generate da:

$$y(\mathbf{x}(i)) = \mu + \epsilon(\mathbf{x}(i)) \qquad \text{per i=1,...,n} \tag{18}$$

dove μ è la media del fenomeno stocastico e l'errore $\epsilon(\mathbf{x}(i))$ ha distribuzione normale $(0, \sigma^2)$.

Nel modello DACE, come nel kriging, i termini $\epsilon(\mathbf{x}(i))$ non sono indipendenti ma "correlati" tra loro. È ragionevole pensare che la correlazione sia alta quando $\mathbf{x}$(i) e $\mathbf{x}$(j) sono vicini e bassa quando i due punti sono lontani. Assumiamo che la correlazione tra gli errori sia legata alla distanza tra i corrispondenti punti. Non usiamo la distanza euclidea siccome questa distanza pesa ugualmente tutte le variabili, piuttosto usiamo la seguente formula di

distanza pesata:

$$d(\mathbf{x}(i), \mathbf{x}(j)) = \sum_{h=1}^{k} \theta_h |x_h(i) - x_h(j)|^{p_h} \qquad \text{con } \theta_h > 0 \qquad \text{e} \qquad p_h \in [1, 2]. \tag{19}$$

Usando questa funzione di distanza la correlazione tra gli errori in $\mathbf{x}(i)$ e $\mathbf{x}(j)$ è

$$Corr[\epsilon(\mathbf{x}(i)), \epsilon(\mathbf{x}(j))] = exp[-d(\mathbf{x}(i), \mathbf{x}(j))]. \tag{20}$$

La funzione di correlazione definita in 19 e 20 è tale che quando la distanza tra $\mathbf{x}$(i) e $\mathbf{x}$(j) è piccola la correlazione è circa uno, quando la distanza tra i punti è alta la correlazione è prossima a zero.

Il parametro θ_h nella formula della distanza 19 può essere interpretato come misura dell'importanza o dell'*attività* della variabile x_h. Dire la variabile x_h è attiva significa che anche piccoli valori di $|x_h(i) - x_h(j)|$ comportano una correlazione bassa tra gli errori $\epsilon(\mathbf{x}(i))$ e $\epsilon(\mathbf{x}(j))$. Osservando le equazioni 19 e 20, si vede che effettivamente se θ_h è molto elevato allora piccoli valori di $|x_h(i) - x_h(j)|$ si traducono in distanze elevate e quindi basse correlazioni.

Il modello descritto dalle equazioni 18, 19 e 20 è chiamato **stochastic process model** poichè il termine errore $\epsilon(\mathbf{x})$ è un processo stocastico, cioè è un insieme di variabili indicizzate da uno spazio (in questo caso dallo spazio k-dimensionale di $\mathbf{x}$) correlate casualmente. È diventato comune chiamare questo modello **DACE stochastic process model**, dove DACE è l'acronimo di *Design and Analysis of Computer Experiments*.

Nel DACE la stima dei parametri μ e σ^2 ha un'interpretazione affatto immediata, poichè devono essere combinati con la stima dei parametri di correlazione θ_h e p_h. Il modello DACE ha in effetti $2k + 2$ parametri : μ, σ^2, $\theta_1, \ldots, \theta_k$, $p_1, \ldots, p_k$. Per stimare questi parametri si va a massimizzare la probabilità del campione.

Ricordiamo la definizione di *distribuzione normale multivariabile*.
Sia $\mathbf{x}$ un vettore di variabili aleatorie di lunghezza q avente distribuzione normale multivariabile con $E[\mathbf{x}] = \mu$ e matrice di covarianza $Cov(\mathbf{x}, \mathbf{x}^T) = \Sigma$. Allora $\mathbf{a}^T\mathbf{x}$ è $N(\mathbf{a}^T\mu, \mathbf{a}^T\Sigma\mathbf{a})$ per ogni fissato vettore $\mathbf{a} \in \Re^q$.

Più in generale se A è una matrice con q colonne, $\mathbf{Ax}$ è $N(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)$.
Se la matrice $\mathbf{\Sigma}$ è definita positiva, allora $\mathbf{x}$ ha densità di probabilità:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{q/2}(det\mathbf{\Sigma}^{1/2})} exp[-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)].$$

Nel nostro caso la densità di probabilità è funzione della variabile aleatoria Y e dei parametri θ e $\mathbf{p}$, pertanto varia al variare di tali parametri. Un modo ragionevole di costruire uno stimatore di y($\mathbf{x}$) è assegnare alla variabile aleatoria i valori osservati e considerare i valori dei parametri per cui la densità di probabilità è massima. Tale funzione prende il nome di **funzione di verosimiglianza** e lo **stimatore** è detto **di massima verosimiglianza**.
Indichiamo con $\mathbf{y} = (y_1, \ldots, y_n)'$ il vettore di dimensione n dei valori osservati della funzione, con $\mathbf{R}$ la matrice nxn il cui elemento (i, j) è uguale a $Corr[\epsilon(\mathbf{x}(i)), \epsilon(\mathbf{x}(j))]$ e con $\mathbf{1}$ il vettore di dimensione n $(1, 1, \ldots, 1)'$. Allora la funzione di verosimiglianza è:

$$\frac{1}{(2\pi)^{n/2}\sigma^{2^{n/2}}(det\mathbf{R})^{1/2}} exp\frac{(\mathbf{y}-\mathbf{1}\mu)'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\mu)}{2\sigma^2} \tag{21}$$

Notiamo che la dipendenza dai parametri θ_h e p_h con $h = 1, \ldots, k$ si ha attraverso la matrice di correlazione $\mathbf{R}$.

Noti i parametri di correlazione θ_h e p_h con $h = 1, \ldots, k$ possiamo determinare i valori di μ e σ^2:

$$\hat{\mu} = \frac{\mathbf{1}'\mathbf{R}^{-1}\mathbf{y}}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}} \tag{22}$$

e

$$\hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{1}\hat{\mu})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\hat{\mu})}{n}. \tag{23}$$

Sostituendo le equazioni 22 e 23 nella funzione di verosimiglianza otteniamo la cosiddetta *concentrated likelihood function*, che dipende solo dai parametri θ_h e p_h con $h = 1, \ldots, k$. Questa è la funzione che in pratica massimizziamo per ottenere gli stimatori θ_h e p_h e quindi determinare la matrice di correlazione $\mathbf{R}$. Con le formule 22 e 23 poi otteniamo le stime di $\hat{\mu}$ e $\hat{\sigma}^2$.

Si osservi che il modello DACE descritto dalle equazioni 18, 19 e 20 è essenzialmente un modello ai minimi quadrati generalizzato (**GLS** *General Least Squares*) con un insieme semplice di regressori, ovvero il termine costante, e una speciale matrice di correlazione tra i punti campionati.

Allo scopo di avere una comprensione intuitiva su come gli errori correlati agiscano sulla previsione, consideriamo l'illustrazione della figura 3 dove c'è un'unica variabile x in ingresso. Il punto x^* in cui stiamo facendo delle previsioni è vicino al secondo punto campionato x_2. Il valore $y(x_2)$ è in modo significativo al di sopra del valore medio stimato μ. Il fatto che $y(x_2)$ stia sopra la linea di regressione implica che i termini omessi in x, che costituiscono l'errore, abbiano nel punto x_2 un valore positivo elevato. Siccome x^* è vicino a x_2 ha senso supporre che anche in x^* i termini omessi siano positivi, sebbene non uguali. La nostra previsione in x^* non deve quindi essere calcolata semplicemente inserendo la x^* nell'equazione di regressione, che ci darebbe solo il valore medio μ, ma deve essere uguale al valore dell'equazione di regressione μ aumentato per tener conto della correlazione con l'errore nel vicino punto x_2.

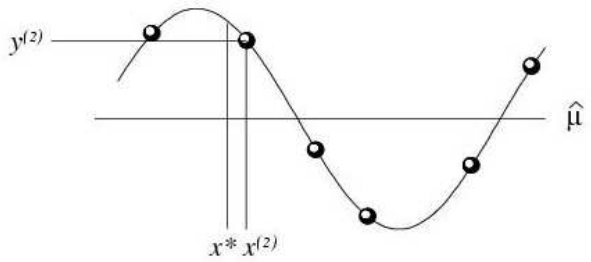

Figura 3: Effetto della correlazione sulla previsione.

Formalmente, sia **r** il vettore di dimensione n delle correlazioni tra i termini errore in $\mathbf{x}^*$ e i termini errore nei punti campionati. Ovvero l'elemento i-esimo di **r** sia $\mathbf{r}_i(\mathbf{x}^*) \equiv Corr[\epsilon(\mathbf{x}^*), \epsilon(\mathbf{x}_i)]$ calcolato usando la formula per la correlazione 19 e 20. Risulta che la **miglior previsione lineare non distorta**

di $y(\mathbf{x}^*)$ è

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \mathbf{r}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}). \tag{24}$$

Al secondo membro dell'equazione 24 il primo termine $\hat{\mu}$ è il risultato ottenuto semplicemente inserendo $\mathbf{x}^*$ nell'equazione di regressione e il secondo termine rappresenta la correzione di questa previsione basata sulla correlazione tra $\epsilon(\mathbf{x}^*)$ e i termini errore nei punti campionati. Notiamo che se la correlazione è nulla ($\mathbf{r} = 0$), allora prevediamo proprio $\hat{y}(\mathbf{x}^*) = \hat{\mu}$.

Per verificare che la previsione DACE va effettivamente a interpolare i dati supponiamo di fare una previsione nell'i-esimo punto campionato e quindi sia $\mathbf{x}^* = \mathbf{x}_i$. In questo caso $\mathbf{r}$ sarà equivalente alla i-esima colonna di $\mathbf{R}$, che possiamo indicare con $\mathbf{R}_i$. Quindi è

$$\mathbf{r}'\mathbf{R}^{-1} = (\mathbf{R} - \mathbf{1}\mathbf{r})' = (\mathbf{R} - \mathbf{1}\mathbf{R}_i)' = e_i' \tag{25}$$

dove e_i è l'i-esimo vettore unitario. L'equazione 24 si riduce così a

$$\hat{y}(\mathbf{x}_i) = \hat{\mu} + e_i'(\mathbf{y} - \mathbf{1}\hat{\mu}) = \hat{\mu} + (y_i - \hat{\mu}) = y_i. \tag{26}$$

Quindi la previsione in $\mathbf{x}_i$ è effettivamente y_i e si dimostra che la previsione DACE interpola i dati.

La correlazione degli errori va a influenzare anche l'accurateza della nostra stima. Ritornando alla figura 3 è facile intuire che siccome x^* è molto vicino al punto x_2 siamo più sicuri della nostra previsione su $y(x^*)$ di quel che saremmo se x^* fosse lontano dai punti campionati.

Tale fatto è descritto dalla formula generale per l'**errore quadratico medio** della previsione:

$$s^2(\mathbf{x}^*) = \sigma^2[1 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}\mathbf{R}^{-1}\mathbf{r})^2}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}}]. \tag{27}$$

Nell'espressione di $s^2(\mathbf{x}^*)$ il termine $-\mathbf{r}'\mathbf{R}^{-1}\mathbf{r}$ rappresenta la riduzione nell'errore previsto dovuta al fatto che $\mathbf{x}^*$ è correlata con i punti campionati. Notiamo che se la correlazione fosse nulla, cioè $\mathbf{r} = 0$, tale correzione sarebbe nulla. Il termine $(1 - \mathbf{1}\mathbf{R}^{-1}\mathbf{r})^2/\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}$ riflette l'incertezza che deriva dal fatto di non conoscere esattamente μ, ma di averlo stimato a partire dai dati campionati.

Supponiamo nuovamente di fare una previsione nell'i-esimo punto campionato, cosicchè $\mathbf{x}^* = \mathbf{x}_i$. Come mostrato in 25 avremmo $\mathbf{R}^{-1}\mathbf{r} = e_i$, quindi

$$\mathbf{r}'\mathbf{R}^{-1}\mathbf{r} = \mathbf{r}'e_i = r_i(\mathbf{x}^*) \equiv Corr(\mathbf{x}^*, \mathbf{x}_i) = Corr(\mathbf{x}_i, \mathbf{x}_i) = 1 \tag{28}$$

e

$$\mathbf{1}'\mathbf{R}^{-1}\mathbf{r} = \mathbf{1}e_i = 1. \tag{29}$$

Sostituendo la 28 e la 29 nell'equazione 27 segue che $s^2(\mathbf{x}^*) = 0$. Infatti con una funzione deterministica una volta che abbiamo campionato un punto conosciamo il suo valore, quindi la nostra incertezza in tale punto deve essere nulla. Spesso è conveniente lavorare con la radice quadrata dell'errore quadratico medio, $s = \sqrt{s^2(\mathbf{x})}$. Ciò fornisce il **RMSE** (*Root Mean Squared Error*) che dà una misura dell'incertezza delle previsioni.

Riassumendo: l'RMSE in un punto campionato è nullo e in un punto molto distante dal campione (dove $\mathbf{r} \sim \mathbf{0}$) è circa σ. Tra questi due estremi l'RMSE è σ diminuito di una quantità che dipende da quanto vicino, e perciò quanto correlato, è il punto in questione ai punti campionati. Affermazioni simili possono essere fatte anche per la previsione DACE $\hat{y}(\mathbf{x})$. In un punto campionato $\hat{y}(\mathbf{x})$ concorda con i dati. In un punto molto lontano dai dati (dove $\mathbf{r} \sim 0$) $\hat{y}(\mathbf{x})$ è circa $\hat{\mu}$. Tra quest estremi $\hat{y}(\mathbf{x})$ è ottenuto attraverso un'interpolazione "smussata" dei dati secondo l'equazione 24.

I parametri del modello, μ, σ^2, $\theta_1,\ldots,$ θ_k, $p_1,\ldots,$ p_k, descrivono il comportamento caratteristico della funzione obiettivo. In particolare i parametri $\theta_1,\ldots,$ θ_k descrivono quanto è sensibile la funzione rispetto alle variabili in ingresso e i parametri $p_1,\ldots,$ p_k indicano quanto la funzione varia rapidamente rispetto ad ogni variabile. Quando stimiamo questi parametri attraverso la massima verosimiglianza, stiamo essenzialmente cercando i valori dei parametri che meglio descivono il comportamento della funzione evidenziato dal nostro campione. Quando prevediamo il valore della funzione in qualche nuovo punto $\mathbf{x}^*$ stiamo in pratica calcolando il valore della funzione che è più consistente con questo comportamento caratteristico. Facciamo semplicemente l'ipotesi

che il valore della funzione in $\mathbf{x}^*$ è un certo numero y^* e quindi aggiungiamo questa pseudo-osservazione relativa al punto n-esimo+1. Potremmo calcolare la probabilità del campione aumentato. Questa probabilità misurerebbe quanto bene la pseudo-osservazione si adatta ai dati originali, cioè la probabilità che essi siano generati dallo stesso modello. È evidente che potremmo ipotizzare diversi valori per y^* e che per ogni ipotesi otteremmo un diverso valore della probabilità. Risulta che il valore di y^* che massimizza questa probabilità aumentata, e quindi il più consistente con il comportamento caratteristico della funzione, è esattamente il predittore dato dall'equazione 24.

Un limite del metodo DACE è che il valore previsto e il suo errore quadratico medio sono ricavati sotto l'assunzione che i parametri σ^2, $\theta_1,\ldots,\theta_k$, $p_1,\ldots,p_k$ siano noti. In realtà i veri valori di questi parametri sono incogniti e i parametri da noi utilizzati nell'estrapolazione sono semplicemente dei valori stimati. Questo *gioco di prestigio* sembra non avere serie conseguenze, sebbene sia probabile che nei campioni di piccole dimensioni conduca a una lieve sottostima dell'errore previsto.

Osserviamo infine che il modello DACE potrebbe assomigliare al modello di regressione lineare standard. In effetti i due modelli condividono una comune cornice matematica consistente in regressori ed errori, ma l'enfasi su tali termini è abbastanza diversa. La regressione lineare si focalizza interamente sui regressori e la stima dei loro coefficienti e fa assunzioni semplicistiche sugli errori, cioè li considera indipendenti gli uni dagli altri. Viceversa il DACE fa assunzioni semplicistiche sui regressori, cioè introduce una semplice costante, e si focalizza interamente sulla struttura di correlazione degli errori. Quindi è probabilmente meglio pensare la regressione e il DACE come modelli diametralmente opposti. La regressione va a stimare i coefficienti di regressione che assieme alla forma di funzione assunta descrivono completamente il comportamento della funzione. IL DACE va a stimare i parametri di correlazione che descrivono tale comportamento. Il DACE fa previsioni attraverso l'interpolazione e l'estrapolazione dai dati campionati in modo consistente con il comportamento caratteristico stimato.

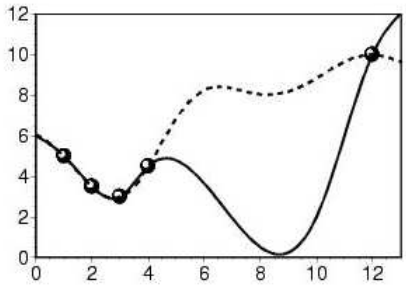

Figura 4: La linea continua rappresenta una funzione campionata nei cinque punti indicati. La linea tratteggiata rappresenta la funzione estrapolata a partire dai punti di allenamento.

0.5 Superfici di risposta adattative

Il *kriging*, oltre a fornire in un punto il valore estrapolato della funzione, associa a questo anche la deviazione standard (vedi fig.4 e 5). Tanto più la deviazione standard è elevata tanto più alta è l'incertezza dell'estrapolazione in tale punto. La conoscenza della deviazione standard ci offre l'opportunità di implementare una superficie di risposta adattativa.

Questo metodo consiste, dato un insieme di allenamento iniziale, nell'andare a determinare nel dominio della funzione il punto di massima deviazione standard, ovvero di massima incertezza, e quindi campionare la funzione in tale punto. Successivamente l'informazione viene aggiunta all'insieme di allenamento. Procedendo iterativamente riusciamo a migliorare la conoscenza sul comportamento globale della funzione.

In alcune applicazioni potremmo essere interessati ad avere una conoscenza della funzione solo localmente più accurata, ad esempio in una zona di minimo o di massimo. In questi casi non adatteremo la superficie di risposta sulla deviazione standard ma su un indice di errore che tenga conto sia della deviazione standard che del valore estrapolato della funzione (vedi fig.6).

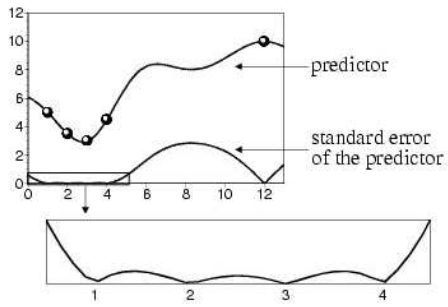

Figura 5: Funzione estrapolata e deviazione standard.

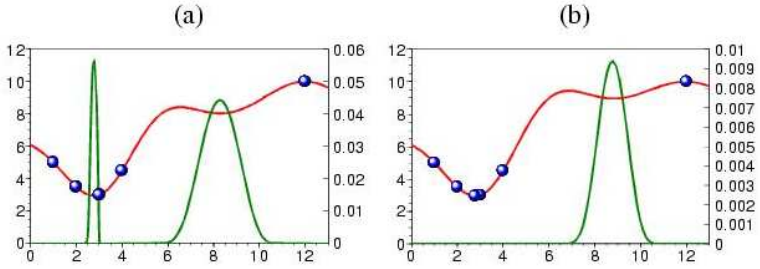

Figura 6: (a) La linea rossa rappresenta la funzione estrapolata. La linea verde rappresente l'indice di errore sulla deviazione standard e sul minimo della funzione. (b) Valori aggiornati dopo aver aggiunto il punto di campionamento in cui precedentemente l'indice di errore era massimo.