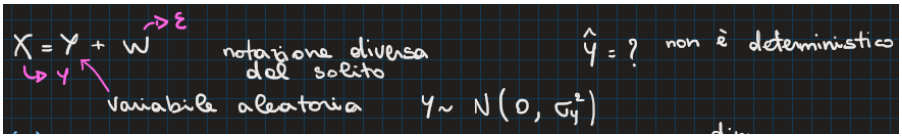
La Stima Bayesiana 1 - 08/10 (Stima parametro aleatorio Y)

Stima Classica e Bayesiana sono due partizioni della statistica.

Stima della variabile aleatoria Y

Riprendiamo l'esercizio di ieri ma lo scriviamo in contesto Bayesiano.

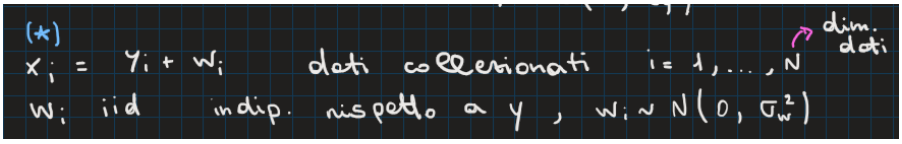


Oggi lo scriviamo X = Y + W ma l'idea è la stessa di ieri, le lettere sono solo lettere (Y = A + ε, Y = θ + W, X = Y + W).

Cosa cambia da Classico a Bayesiano?

La Y in Bayesiano non è più un valore deterministico incognito ma è invece una variabile aleatoria con una sua distribuzione che può variare da un esperimento ad un altro, ha una sua distribuzione a priori (prima di osservare i dati) della quale conosciamo la legge (grazie ad una conoscenza storica) e che vogliamo osservare e potenzialmente modificare sulla base delle nuove osservazioni.

Magari conosciamo il valore medio e la varianza, da una storia di osservazioni, questa è la conoscenza a priori.



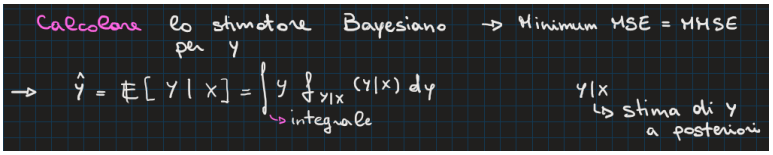
Raccogliamo i dati, $x_i = y_i + w_i$, i va da 1 a N, la data size, la cardinalità dell'insieme delle osservazioni.

Assumiamo che le w_i siano iid e inoltre w_i indipendenti with respect to y.

Assumiamo anche per w_i una distribuzione, sarà normale con media 0 e varianza σ_w^2 con omoschedasticità.

MMSE

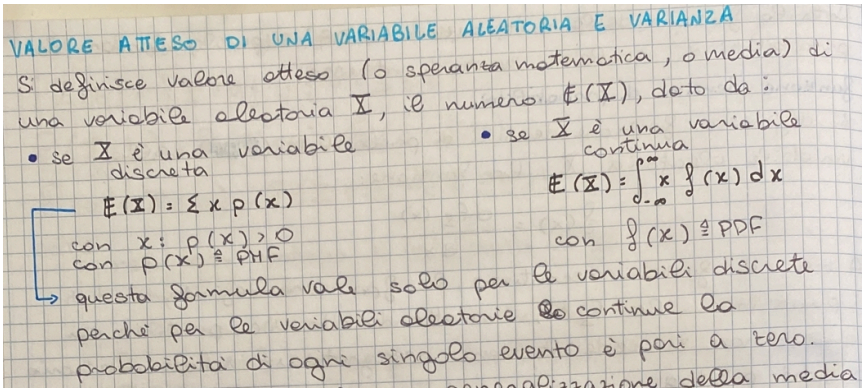
Lo stimatore Bayesiano è detto in ingegneria Minimum Mean Squared Error Estimator, infatti minimizza l'MSE.



Y cappello sarà la media di Y dato X (X sono i nostri dati appena raccolti). Si nota che non c'è differenza dalla funzione di regressione vista in passato, ovviamente dobbiamo particolarizzare per questo problema.

Se consideriamo che Y è Gaussiana allora questa media sarà l'integrale della pdf condizionata, cioè la nuova y dati i dati raccolti x (y assoluto già lo abbiamo ipotizzato all'inizio, sono le info che già si hanno per l'esperienza, le informazioni a priori, mentre ora con i dati raccolti avremo le informazioni a posteriori).

▼ Media e integrali



Supponiamo di avere due variabili casuali X e Y. La **media condizionata** di Y dato $X = x$ è denotata come $\mathbb{E}(Y|X = x)$, ed è calcolata utilizzando la distribuzione condizionata di Y dato $X = x$.

Nel caso di variabili continue, la media condizionata $\mathbb{E}(Y|X = x)$ può essere espressa come:

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy$$

dove $f_{Y|X}(y|x)$ è la **densità di probabilità condizionata** di Y dato $X = x$.

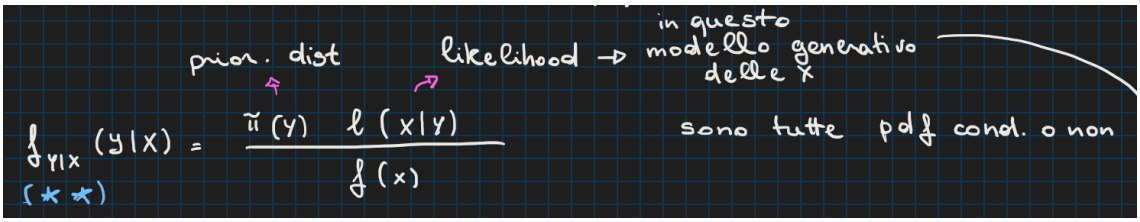
Il calcolo di $f_{Y|X}(y|x)$

Il problema diventa ora il calcolo di $f_{Y|X}(y|x)$.

Partiamo ricordando la Regola di Bayes:

La Regola di Bayes ci torna utile perché siamo in possesso della f(Y), che è l'informazione a priori, e ci serve l'informazione a posteriori cioè la f(Y|X).

Definimmo la Regola di Bayes in insiemistica e ora la riscriviamo per le pdf (anche se usiamo simboli variegati come π stiamo sempre parlando di pdf).



La pdf di Y prende il nome di distribuzione a priori, o prior distribution.

La probabilità condizionata $l(x|y)$ è la probabilità di come sono generati i dati che raccolgo dato un fissato livello di Y. Avendo Y possiamo avere le X. Questa è la funzione che genera le X a partire dalle Y, si chiamano modelli generativi, questo è il modello generativo delle X.

Avendo il modello di W, se abbiamo Y possiamo ricavare X.

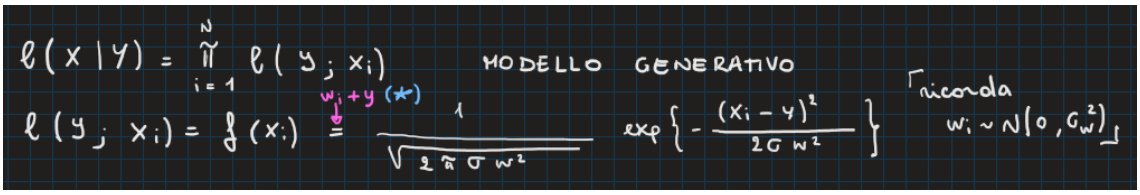
Questa l è la congiunta delle X dato Y, la congiunta delle X dati i parametri è la verosimiglianza (vedere definizione formale) ed è il modello generativo delle X conoscendo le Y, ovviamente questo si può fare solo conoscendo la relazione tra X ed Y. La barra nella verosimiglianza non è proprio da puristi, i puristi vorrebbero l(parametro (nel nostro caso y):dato (nel nostro caso x)).

Non tutti i modelli generativi sono funzioni di verosimiglianza.

La pdf di X è la condizionante a primo membro.

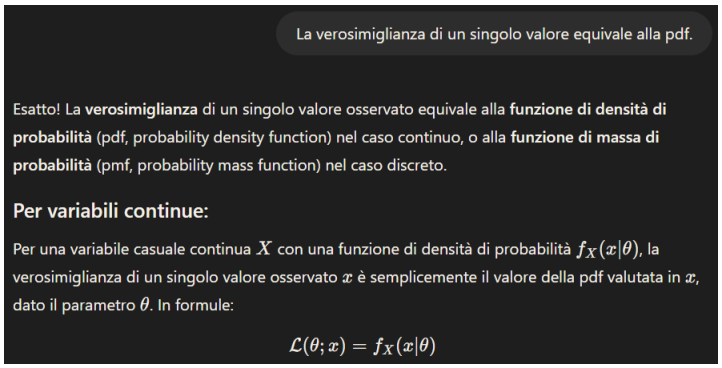
Ricordiamo che i nostri dati sono tutti iid.

f(y|x) non c'è l'abbiamo e quindi ribaltiamo rispetto al condizionamento e il problema diventa l(x|y) che conoscendo il modello possiamo trovarlo perché non sono altro che le x se conosco il modello di legame tra x e y.

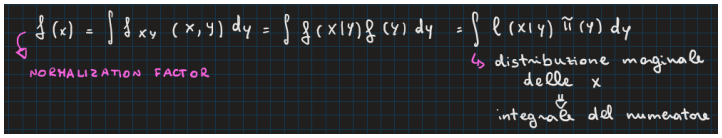


La verosimiglianza di un singolo valore equivale alla pdf.

Ricordiamo che w sono Gaussiane e la y è un numero (essendo un'osservazione) e quindi contribuisce spostando la media di w.



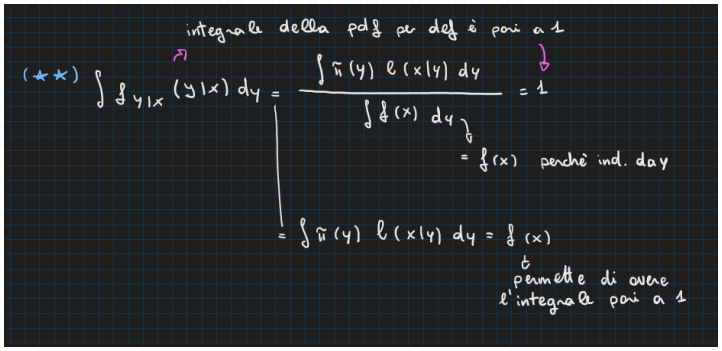
A questo punto ci manca solo da definire $f(x)$.



La pdf della x lo possiamo scrivere come integrale della congiunta rispetto a y. (è una regola che si può ricavare la marginale integrando la congiunta per le variabili che si vogliono togliere, in questo caso y)

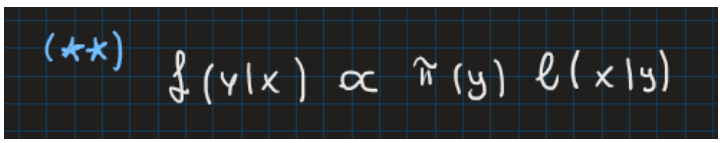
Poi riscriviamo la congiunta con la probabilità condizionata.

Così abbiamo rivelato che la distribuzione marginale delle x non è altro che l'integrale del numeratore, quindi è un termine di normalizzazione.



Lo statistico vorrebbe avere termini e oggetti di facile manipolazione, quindi la f(x) a denominatore ha come scopo SOLO rendere il numeratore pari ad 1, quindi per la mia inferenza non mi serve, lo ignoro e poi il risultato del numeratore si divide per una costante per renderlo 1.

Una cosa è fare un generatore di cose mai viste e una cosa è fare un generatore di gaussiane, noi facciamo sempre manipolazioni per portarci a casi semplici perché si lavora meglio in questo modo.

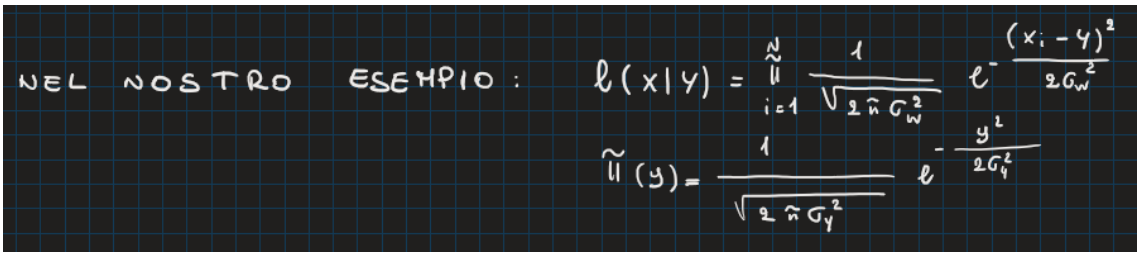


Quindi diciamo che la nostra f(y|x) sarà proporzionale rispetto ad una costante (il denominatore). Quindi consideriamo solo il numeratore.

> 💡 Questa versione alleggerita della regola di Bayes è quello che si usa sempre in statistica Bayesiana.

Esercizio - ricavare la conoscenza a posteriori

Data la premessa teorica spiegata fino ad ora, partiamo con un esercizio.



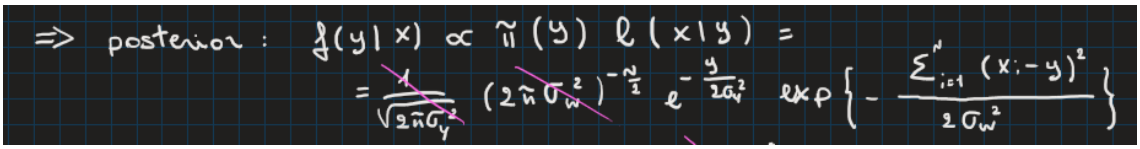
Per adoperare la versione stringata della regola di Bayes che abbiamo trovato ci occorrono la likelihood, anche detta modello generativo, e la distribuzione delle y (prior distribution).

La y tendenzialmente si ha a causa di una serie storica, uno statistico chiamato da una fabbrica di resistori può aspettarsi che la fabbrica abbia già i dati della distribuzione delle resistenza che producono.

Supponiamo di misurare poi i dati e scopriamo che la media risulta lontana da quella della conoscenza a priori, cioè la mia a priori è molto lontana dalla mia likelihood, questo può capitare ma potrebbe portare a stime non buone, spesso c'è un errore da parte dello statistico che ha raccolto male i dati o altro, generalmente la prior distribution ha una certa affidabilità.

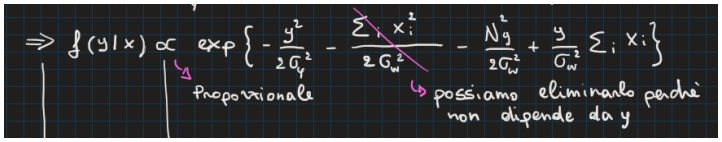
Comunque il nostro obiettivo è mettere insieme prior e likelihood per ottenere la conoscenza a posteriori. Sappiamo dalla formula precedentemente ricavata che per farlo basta il prodotto.

Calcolo della $f_{Y|X}(y|x)$



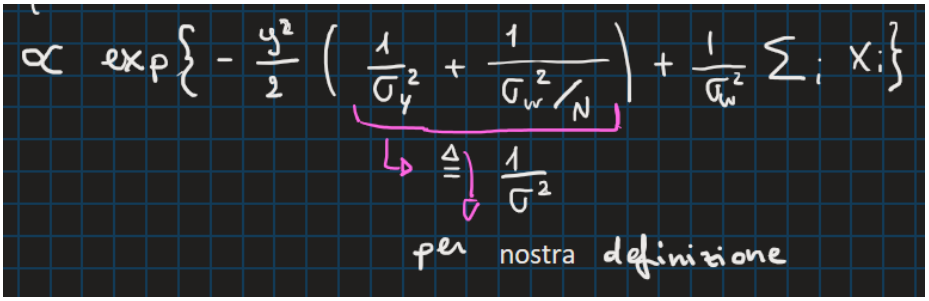
Nel primo esponenziale la y doveva essere al quadrato

Come abbiamo eliminato f(X) perché era equivalente ad una costante ora possiamo rimuovere anche tutte le altre costanti che troviamo.

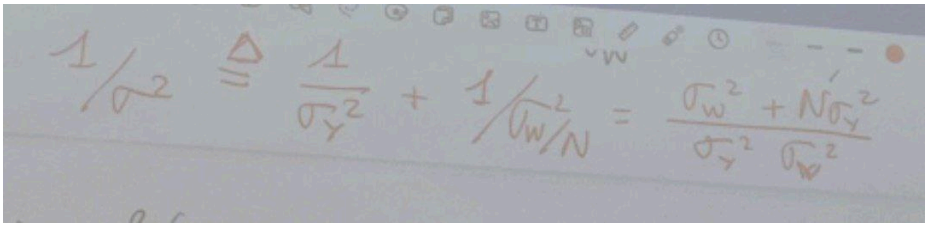


il terzo termine nell'esponenziale è sommatoria su i delle y al quadrato, le y sono tutte uguali quindi è n volte y al quadrato

e poi:



Manca la y a numeratore dell'ultimo pezzo nell'esponenziale



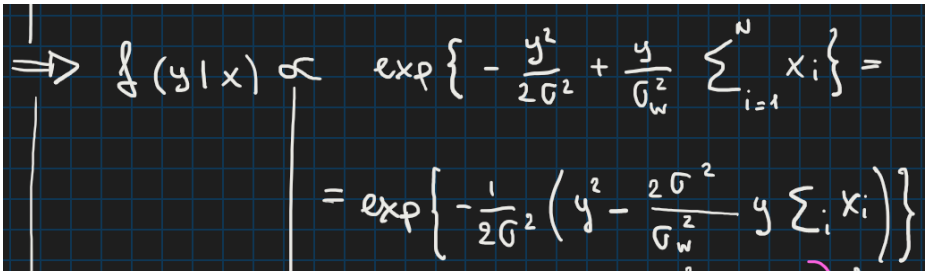
Questa assomiglia quasi al reciproco di una media armonica.

Lo chiamiamo $1/\sigma^2$ perché ha le dimensioni del σ^2.

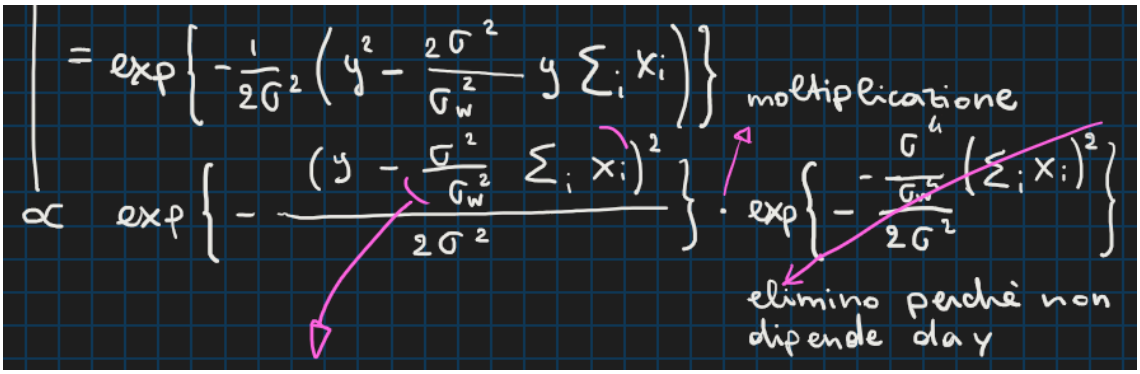
Calcoli per ottenere una Gaussiana

Si può trasformare l'esponenziale che abbiamo nella parte esponenziale della pdf di una Gaussiana (tanto le costanti dividendo le possiamo eliminare fintanto che teniamo a mente che quello che otteniamo in integrale deve fare 1 alla fine).

Quindi possiamo dire che la pdf condizionata che stiamo cercando sarà un qualcosa di simile ad una Gaussiana.

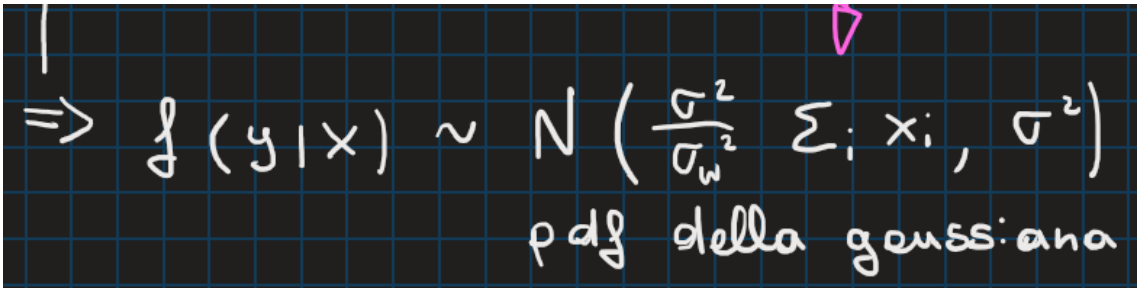


Notiamo che abbiamo un doppio prodotto, quindi mi riscrivo questo pezzo come quadrato, per farlo bisognava sottrarre un termine opportuno, quindi dobbiamo anche sommarlo.



Per le proprietà degli esponenziali me lo porto in un altro esponenziale il termine sommato.

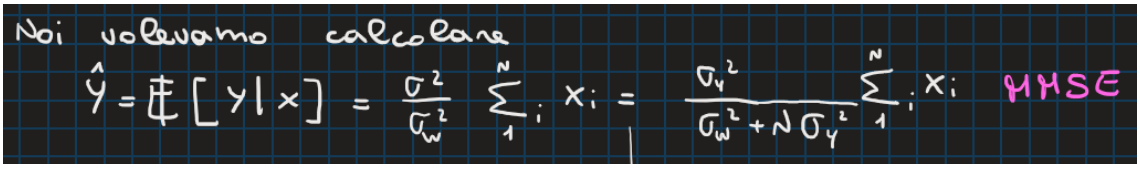
Moltiplicando per un altro esponenziale per togliere l'esponenziale aggiunto solo per permettere di fare il quadrato otteniamo finalmente solo quello che ci serve:



In questo caso mettendo bene i pezzi siamo riusciti ad ottenere una gaussiana, non sempre è possibile e se non è possibile si usano altri algoritmi.

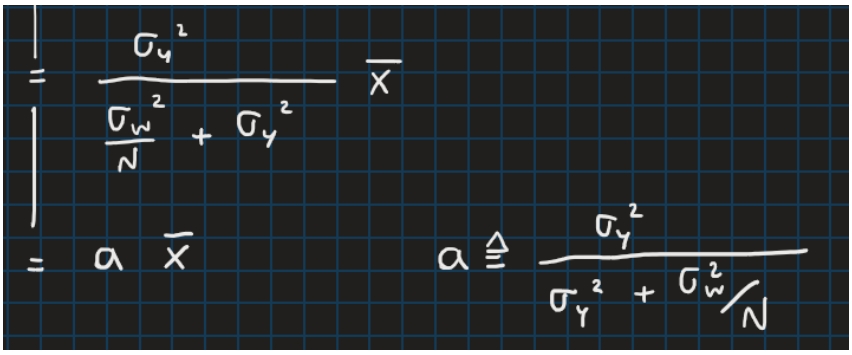
Ritorniamo al problema originario, tutto questo serve a stimare Y

Noi volevamo trovare $\hat{Y}$, lo stimatore Bayesiano, avevamo detto che era il valore atteso di Y dato X che avevamo detto essere uguale all'integrale della distribuzione condizionata Y|X MA noi abbiamo già trovato la media di questa distribuzione che abbiamo scoperto essere Gaussiana e quindi non c'è nessun bisogno di calcolare integrali.



Abbiamo trovato il nostro MMSE.

Mettendo N in evidenza e collegandolo alla sommatoria otteniamo il prodotto di una costante e della media campionaria.

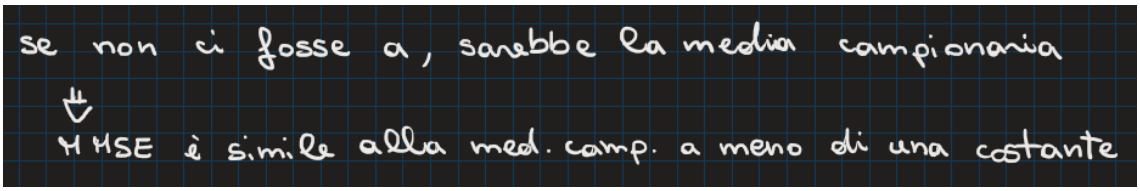


Abbiamo tutti i dati per fare questi calcoli.

Notiamo che togliendo la a lo stimatore della media è proprio la **media campionaria**.

Lo stimatore a massima verosimiglianza della media della combinazione Y + w sarebbe stato la media campionaria.

L'MMSE non è proprio la media campionaria ma è simile a meno di una costante che dipende dalle varianze in gioco.



Il valore atteso della media campionaria delle x sarebbe proprio la x, già sappiamo che la media campionaria è uno stimatore unbiased ma qui il problema è diverso visto che Y è variabile aleatoria.

Infatti nel Bayesiano **parlare di unbiased non ha senso** perché abbiamo l'effetto di tutte le varianze, in classico possiamo avere il concetto di unbiased perché c'è un valore vero da trovare, qui invece c'è una distribuzione.

Ma poi anche a livello di quello che è il nostro obiettivo in Bayesiano noi abbiamo la nostra conoscenza a priori, ci mettiamo i dati e l'obiettivo è proprio modificare, spostare, il prior, in un certo senso noi vogliamo polarizzare la nostra conoscenza a priori usando i dati, quindi il concetto di unbiased perde di significato.

La correttezza non ci serve perché **noi stiamo modificando la nostra conoscenza** spostando magari anche la media.

I dati spostano il prior e se non avessi i dati resterebbe solo il prior, infatti con il Bayesiano si può fare inferenza anche senza dati, usando solo la conoscenza a priori, è un caso limite ma comunque è si può fare.

Considerazioni finali

$a \neq 1$

Se a è diverso da 1 non ho proprio la y valore attuale ma sto modificando l'informazione, del resto l'approccio di bayes si usa per arricchire l'informazione.

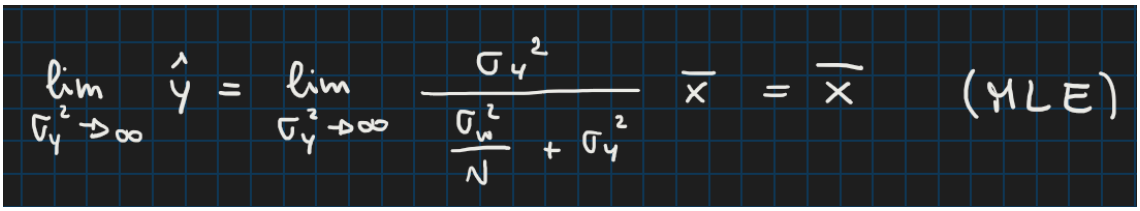
$\sigma_y^2 \to \infty$

In questo caso la prior è non informativa.

L'informazione a priori ha una varianza infinita, può essere ovunque, e quindi non ho informazioni a priori, in genere nel bayesiano se questo è vero si mettono oggetti con varianze molto grandi.

La condizionata deve essere una pdf però, così posso usare le pdf improprie, si aprono i filoni non informativi.

Facciamo il limite.



Eliminando l'informazione a priori dell'MMSE resta solo la media campionaria, cioè lo stimatore con i suoi dati cioè la massima verosimiglianza.

$\sigma_y^2 \to 0$

Se la varianza tende a 0 la distribuzione di Y diventa un valore deterministico, la media, che sappiamo essere 0.

Del resto il limite per sigma quadro di y che tende a 0 fa 0.
La mia distribuzione degenera alla media della distribuzione a priori che è 0.

$N \to \infty$

N è importante perché mi dice quante informazioni sto raccogliendo, se N tende all'infinito, con infiniti dati prevalgono i dati sulla prior e prevale la stima a massima verosimiglianza.