Probability - Consistency

Consistence of point estimation

Point estimation is a single value or single vector estimate of model parameter, which is the best estimate or the most possible estimate, in contrast to interval estimation, which gives an interval. We evaluate an estimation by studying its bias, variance and consistence. We know bias analysis and variance analysis well, but what is consistence analysis? An estimate is said to be consistent if the estimate converges to the ground truth as number of observations (or data points) tends to infinity. An unbiased estimator can be consistent, a consistent estimator can be biased.

$E[X_{est,N}]$	=	X_{true}	$\forall N > 0$	definition for unbiased
$\underset{N\to\infty}{p\lim}\, X_{est,N}$	=	X_{true}		definition for consistent

Example, given N points x_n where $n \in [1,N]$, here are two mean estimators :

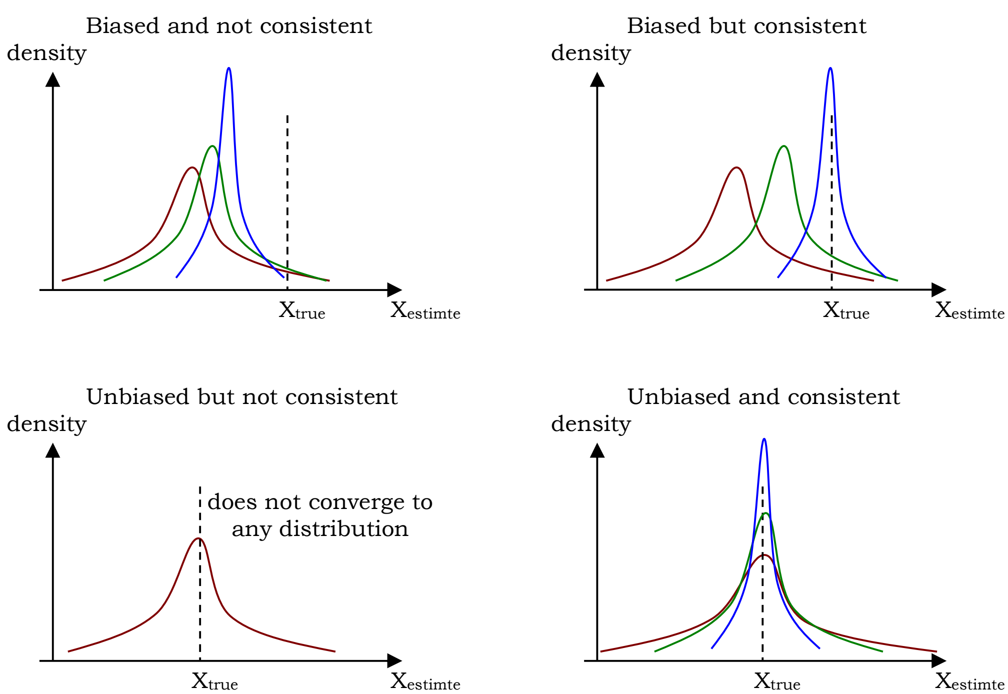
$$X_{est1,N} = x_1$$

$$X_{est2,N} = \left(\textstyle\sum_{n=1}^{N} x_n\right)/N + 1/N$$

The first one is unbiased but not consistent, while the second one is biased but consistent. Lets check.

$E[X_{est,1}]$	=	$E[x_1]$	=	μ		unbiased
$\underset{N\to\infty}{p\lim}\, X_{est1,N}$	=	x_1				non consistent (does not converge)
$E[X_{est,2}]$	=	$E[(\sum_{n=1}^{N} x_n)/N + 1/N]$		=	$\mu + 1/N$	biased
$\underset{N\to\infty}{p\lim}\, X_{est2,N}$	=	$\underset{N\to\infty}{p\lim}(\sum_{n=1}^{N} x_n)/N + \underset{N\to\infty}{p\lim}\, 1/N$		=	μ	non consistent (by law of large number)

Suppose X is our model parameter, X_{true} is the ground truth, we have :



Lets review some concepts that are very important for consistence analysis.

Convergence of random variable (Asymptotic behaviour)

For a sequence of random variable X_1, X_2, X_3, …, X_N, there are many types of convergence as N tends to infinity : (1) convergence in distribution (or weak convergence), (2) convergence in probability, (3) almost sure convergence, (4) sure convergence and (5) convergence in mean. We will talk about the first three types here. Please note that we are always talking about **convergence in probability** when we study the consistence of an estimator.

	Definition			Notation
convergence in distribution	$\lim_{N\to\infty} F_N(x)$	=	$F(x)$	$X_N \xrightarrow{d} X$
convergence in probability	$\lim_{N\to\infty} \Pr(\vert X_N - X\vert \geq \varepsilon)$	=	0	$X_N \xrightarrow{p} X$ or $\underset{N\to\infty}{p\lim} X_N = X$
almost sure convergence	$\Pr(\lim_{N\to\infty} X_N = X)$	=	1	$X_N \xrightarrow{as} X$

where $F_N(x)$ and $F(x)$ are cumulative density functions of X_N and X respectively. Main difference between convergence in distribution and convergence in probability is that the former claims two random variables having same distribution, they have different probability spaces, while the latter claims two random variables are indeed the same random variable, because they have the same realizations and the same probability space. Besides, convergence in probability implies convergence in distribution, while almost sure convergence implies convergence in probability, but not the other way round.

$$X_N \xrightarrow{p} X \quad\Rightarrow\quad X_N \xrightarrow{d} X$$

and
$$X_N \xrightarrow{as} X \quad\Rightarrow\quad X_N \xrightarrow{p} X$$

Here is the proof of the first part, which makes use of the lemma, for any random variables A and B :

$$\begin{aligned}
\Pr(A \leq a) &= \Pr(A \leq a, B \leq a+\varepsilon) + \Pr(A \leq a, B > a+\varepsilon) \\
&= \Pr(A \leq a, B \leq a+\varepsilon) + \Pr(A - B \leq a - B, B > a+\varepsilon) \\
&= \Pr(A \leq a, B \leq a+\varepsilon) + \Pr(A - B \leq a - B, a - B < -\varepsilon) \\
&= \Pr(A \leq a, B \leq a+\varepsilon) + \Pr(A - B < -\varepsilon) \\
&\leq \Pr(A \leq a, B \leq a+\varepsilon) + \Pr(A - B < -\varepsilon) + \Pr(A - B > \varepsilon) \\
&= \Pr(A \leq a, B \leq a+\varepsilon) + \Pr(|A - B| > \varepsilon) \\
&\leq \Pr(B \leq a+\varepsilon) + \Pr(|A - B| > \varepsilon)
\end{aligned}$$

With this lemma, we have :

$$\Pr(X_N \leq x) \leq \Pr(X \leq x+\varepsilon) + \Pr(|X_N - X| > \varepsilon) \qquad \text{putting } A = X_N\,,\; B = X \text{ and } a = x$$

and
$$\Pr(X \leq x-\varepsilon) \leq \Pr(X_N \leq x) + \Pr(|X_N - X| > \varepsilon) \qquad \text{putting } A = X\,,\; B = X_N \text{ and } a = x-\varepsilon$$

Combine both, we have :

$$\begin{aligned}
\Pr(X \leq x-\varepsilon) - \Pr(|X_N - X| > \varepsilon) &\leq \Pr(X_N \leq x) \leq \Pr(X \leq x+\varepsilon) + \Pr(|X_N - X| > \varepsilon) \\
\Pr(X \leq x-\varepsilon) - \lim_{N\to\infty} \Pr(|X_N - X| > \varepsilon) &\leq \lim_{N\to\infty} \Pr(X_N \leq x) \leq \Pr(X \leq x+\varepsilon) + \lim_{N\to\infty} \Pr(|X_N - X| > \varepsilon)
\end{aligned}$$

According to the definition of convergence in probability, we have :

$$\Pr(X \leq x-\varepsilon) \leq \lim_{N\to\infty} \Pr(X_N \leq x) \leq \Pr(X \leq x+\varepsilon) \qquad \text{since } \lim_{N\to\infty} \Pr(|X_N - X| > \varepsilon) = 0$$

$$\Rightarrow \qquad \lim_{N\to\infty} \Pr(X_N \leq x) = \Pr(X \leq x) \qquad \text{by sandwich theorem}$$

The calculation of plim is not straight forward, it can be made easier with the help of central limit theorem, law of large number, Slutsky theorem and Levy's continuity theorem.

Central limit theorem and Law of large number
Suppose X_1, X_2, X_3, …, X_N is a sequence of independent and identically distributed random variables, they share the same mean and variance, i.e.

$$
\begin{array}{lll}
E[X_n] & = & \mu \qquad\qquad \forall n \in [1,N] \\
Var[X_n] & = & \sigma^2 \qquad\qquad \forall n \in [1,N]
\end{array}
$$

then : $\lim_{N\to\infty} \sum_{n=1}^N X_n / N \xrightarrow{d} normal(\mu, \sigma/\sqrt{N})$ by central limit theorm

$\lim_{N\to\infty} \sum_{n=1}^N X_n / N \xrightarrow{p} \mu$ by weak law of large number

$\lim_{N\to\infty} \sum_{n=1}^N X_n / N \xrightarrow{as} \mu$ by strong law of large number

However, the above central limit theorem is incorrect. We should not put N on both sides of the formula (as N tends to infinity). Thus central limit theorem should be restated as :

$$
\begin{aligned}
& \lim_{N\to\infty} \sqrt{N}\left(\sum_{n=1}^N X_n / N - \mu\right) \\
= & \lim_{N\to\infty} \frac{\sum_{n=1}^N X_n / N - \mu}{1/\sqrt{N}} \xrightarrow{d} normal(0,\sigma) \\
\text{or} \quad & \lim_{N\to\infty} \frac{\sum_{n=1}^N X_n / N - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} normal(0,1) \\
\text{or} \quad & \lim_{N\to\infty} \frac{\sum_{n=1}^N X_n / N - \mu}{\sqrt{\sigma^2/N}} \xrightarrow{d} normal(0,1)
\end{aligned}
$$

Thus central limit theorem shows the relation between sample mean, real mean and real variance.

Slutsky theorem
Suppose X_1, X_2, X_3, …, X_N is a sequence of random variables (or random matrices) that converge to a random variable (or random matrices) X, while Y_1, Y_2, Y_3, …, Y_N is a sequence of random variables (or random matrices) that converge to a constant value (or constant matrix) C, then we have :

$$
\begin{array}{llll}
\text{If} & X_n & \xrightarrow{d} & X \\
 & Y_n & \xrightarrow{d} & C \\
\text{then} & X_n + Y_n & \xrightarrow{d} & X + C \\
 & Y_n X_n & \xrightarrow{d} & CX \\
 & Y_n^{-1} X_n & \xrightarrow{d} & C^{-1} X
\end{array}
$$

The above statements are not valid if C is random. Furthermore, the above statements are still valid for convergence in distribution, convergence in probability and almost sure convergence.

Levy continuity theorem
Levy continuity theorem states the relation between convergence in distribution and pointwise convergence in characteristic function.

$$
\begin{array}{llll}
\text{If} & \varphi_n(t) & = & E[\exp(itX_n)] \qquad \forall n \in [1,N] \\
\text{and} & \varphi(t) & = & E[\exp(itX)] \\
\text{and} & \varphi_n(t) & \longrightarrow & \varphi(t) \\
\text{then} & X_n & \xrightarrow{d} & X
\end{array}
$$