Maximum Likelihood Estimation - Line Fitting

Andersen Ang

First created: 2014. Last update : 2017

Content

1 The estimation problem

Consider to estimate a straight line signal in the addictive white Gaussian noise environment $x[n] = A+Bn+w[n]$ over N samples $(n = 0, 1, ..., N - 1)$.

- A, B are constant but unknown. They are the parameters to be estimated : $\theta = [A\,B]$. So in this case $p\left(x; \theta\right) = p\left(x; A, B\right)$
- w[n] is a addictive white Gaussian noise $w[n] \sim \mathcal{N}\left(0, \sigma^2\right)$.

2 The estimation for $N = 2$ observations

If there is only 1 observed data $(N = 1)$, we have

$$x[0] = A + w[0]$$

B cannot be estimated. So at least we need to have 2 observations :

$$x[0] = A + w[0]$$
$$x[1] = A + B + w[1]$$

Their PDF will be

$$p\left(x[0]; A, B\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\Big(x[0] - A\Big)^2\right\} \quad p\left(x[1]; A, B\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\Big(x[1] - A - B\Big)^2\right\}$$

The conditional PDF of the model for single obervation is

$$p\left(x; A, B\right) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[\Big(x[0] - A\Big)^2 + \Big(x[1] - A - B\Big)^2\right]\right\}$$

The log-likelihood function will be

$$\ln p\left(x; A, B\right) = -\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left[\Big(x[0] - A\Big)^2 + \Big(x[1] - A - B\Big)^2\right]$$

Take $\frac{\partial}{\partial A}$, $\frac{\partial^2}{\partial A^2}$, $\frac{\partial}{\partial B}$, $\frac{\partial^2}{\partial A \partial B}$ and $\frac{\partial^2}{\partial B^2}$

$$\frac{\partial \ln p(x;A,B)}{\partial A} = \frac{1}{\sigma^2}\left[\left(x[0]-A\right)+\left(x[1]-A-B\right)\right]$$

$$\frac{\partial^2 \ln p(x;A,B)}{\partial A^2} = \frac{-2}{\sigma^2}$$

$$\frac{\partial^2 \ln p(x;A,B)}{\partial A \partial B} = \frac{-1}{\sigma^2}$$

$$\frac{\partial \ln p(x;A,B)}{\partial B} = \frac{1}{\sigma^2}\left[\left(x[1]-A-B\right)\right]$$

$$\frac{\partial^2 \ln p(x;A,B)}{\partial B^2} = \frac{-1}{\sigma^2}$$

The Fisher Information will be

$$I(A) = -\mathbb{E}\begin{bmatrix} \frac{\partial^2 \ln p(x;A,B)}{\partial A^2} & \frac{\partial^2 \ln p(x;A,B)}{\partial A \partial B} \\ \frac{\partial^2 \ln p(x;A,B)}{\partial B \partial A} & \frac{\partial^2 \ln p(x;A,B)}{\partial B^2} \end{bmatrix} = \frac{1}{\sigma^2}\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

Thus the variance of the estimrator will be

$$Var\left(\hat{A}\right) = \frac{1}{[I(\theta)]_{11}} = \frac{\sigma^2}{2} \qquad Var\left(\hat{B}\right) = \frac{1}{[I(\theta)]_{22}} = \sigma^2$$

Which are the variance of the addictive noise. Notice that this is exactly the Cramer-Rao Lower Bound of this estimator, that means

$$\text{The lower bound of the variance of the estimator } \hat{A} = 0.5\sigma^2$$

$$\text{The lower bound of the variance of the estimator } \hat{B} = \sigma^2$$

It's physical meaning is that, if the variance of the addictive noise is small, then the variance of the estimator, (on average) will be as small as the variance of the noise. This is natural, because in the estimation model $x[0] = A + w[0]$, only noise $w[0]$ corrupt our measurement of the A, and therefore a samller noise variance should make the measurement of the DC component easier (having smaller estimator variance).

Note that $\frac{\partial^2}{\partial A \partial B} = \frac{\partial^2}{\partial B \partial A}$, so the Fisher Information Matrix is a symmetric matrix.

3 The estimation for N observation

Now instead of 1 observation, we have N observations

$$x[n] = A + Bn + w[n] \;\; n = 0, 1, ..., N-1$$

Each data point will have a PDF

$$p(x[k];A,B) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x[k]-A-Bk\right)^2\right\}$$

By the multiplication law of probability, we have

$$p(x[0], x[1], ..., x[N-1]\,;\, A, B) = p(x[0];A)\,p(x[1];A)\,...p(x[N-1];A,B)$$

Denote $p(x[0], x[1], ..., x[N-1]\,;\, A)$ as $p(x;A)$, the PDF of the model will be

$$p(x;A,B)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\Big(x[0]-A\Big)^2\right\}\right)...\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\Big(x[N-1]-A-B(N-1)\Big)^2\right\}\right)$$

Using compact notation

$$p(x;A,B)=\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\prod_{n=0}^{N-1}\exp\left\{-\frac{1}{2\sigma^2}\Big(x[n]-A-Bn\Big)^2\right\}$$

Since $\exp\alpha\exp\beta=\exp(\alpha+\beta)$, the likelihood function is

$$p(x;A,B)=\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\Big(x[n]-A-Bn\Big)^2\right\}$$

The log-likelihood function will be

$$\ln p(x;A,B)=-\frac{N}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\Big(x(n)-A-Bn\Big)^2$$

Take $\frac{\partial}{\partial A}$,$\frac{\partial^2}{\partial A^2}$, $\frac{\partial^2}{\partial A\partial B}$, $\frac{\partial}{\partial B}$ and $\frac{\partial^2}{\partial B^2}$

$$\frac{\partial\ln p(x;A,B)}{\partial A}=\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\Big(x(n)-A-Bn\Big)$$

$$\frac{\partial^2\ln p(x;A,B)}{\partial A^2}=\frac{-N}{\sigma^2}$$

$$\frac{\partial\ln p(x;A,B)}{\partial A\partial B}=\frac{-1}{\sigma^2}\sum_{n=0}^{N-1}n=\frac{-N(N-1)}{\sigma^2}$$

$$\frac{\partial\ln p(x;A,B)}{\partial B}=\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\Big(x(n)-A-Bn\Big)n$$

$$\frac{\partial^2\ln p(x;A,B)}{\partial B^2}=\frac{-1}{\sigma^2}\sum_{n=0}^{N-1}n^2=\frac{-N(N-1)(2N-1)}{6\sigma^2}$$

The Fisher Information Matrix is

$$I(\theta)=-\mathbb{E}\begin{bmatrix}\frac{\partial^2\ln p(x;A,B)}{\partial A^2} & \frac{\partial^2\ln p(x;A,B)}{\partial A\partial B}\\ \frac{\partial^2\ln p(x;A,B)}{\partial B\partial A} & \frac{\partial^2\ln p(x;A,B)}{\partial B^2}\end{bmatrix}=\frac{1}{\sigma^2}\begin{bmatrix}N & \frac{N(N-1)}{2}\\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6}\end{bmatrix}$$

To find the Cramer-Rao Lower Bound , the inverse of the Fisher Information Matrix has to be computed.

$$I^{-1}(\theta)=\sigma^2\begin{bmatrix}N & \frac{N(N-1)}{2}\\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6}\end{bmatrix}^{-1}$$

$$=\frac{\sigma^2}{\frac{N^2(N-1)(2N-1)}{6}-\frac{N^2(N-1)^2}{4}}\begin{bmatrix}\frac{N(N-1)(2N-1)}{6} & -\frac{N(N-1)}{2}\\ -\frac{N(N-1)}{2} & N\end{bmatrix}$$

$$=\sigma^2\begin{bmatrix}2\frac{(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)}\\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)}\end{bmatrix}$$

Therefore, the Cramer-Rao Lower Bounds of the estimrators for having N samples are

$$Var\left(\hat{A}\right) \geq \frac{2(2N-1)\sigma^2}{N(N+1)} \qquad Var\left(\hat{B}\right) \geq \frac{12\sigma^2}{N(N^2-1)}$$

Notice that they have different performances :

$$Var\left(\hat{A}\right) \sim \frac{\sigma^2}{O(N^1)} \qquad Var\left(\hat{B}\right) \sim \frac{\sigma^2}{O(N^3)}$$

Hence B is easier to estimate. Physically it is true since the estimation of B is more sensitive to change of n. Where the estimation of A is insensitive to the change of n. (The change of n have no effect on A).

The comparision of the performance of the estimators can also be observed by computing their ratio

$$\frac{Var(\hat{A})}{Var(\hat{B})} = \frac{\dfrac{2(2N-1)\sigma^2}{N(N+1)}}{\dfrac{12\sigma^2}{N(N^2-1)}} = \frac{(2N-1)(N-1)}{6} > 1 \text{ when } n > 3$$

When $n > 3$, the lower bound of the variance of $\hat{A}$ is larger than that of $\hat{B}$, thus on average B is easier to estimate.

4 The estimation when B is known

Consider we have N observations

$$x[n] = A + Bn + w[n] \; n = 0, 1, ..., N-1$$

Now B is known, and we want ot estimate A.
Since now B is known, then we can rearrange the equation as

$$x[n] - Bn = A + w[n] \; n = 0, 1, ..., N-1$$

And re-name $x[n] - Bn$ as $z[n]$, thus the model becomes

$$z[n] = A + w[n] \; n = 0, 1, ..., N-1$$

Then this problem becomes the problem of estimating the DC component of the model $x[n] = A + w[n]$. From the previous study of this model, the Cramer-Rao Lower Bound is

$$Var\left(\hat{A}\right)_{B \text{ known}} \geq \frac{\sigma^2}{N}$$

The Cramer-Rao Lower Bound for unknwon B is

$$Var\left(\hat{A}\right)_{B \text{ unknown}} \geq \frac{2(2N-1)\sigma^2}{N(N+1)}$$

Hence the knowledge of B will improve the Cramer-Rao Lower Bound by the factor of

$$\frac{Var\left(\hat{A}\right)_{B \text{ known}}}{Var\left(\hat{A}\right)_{B \text{ unknown}}} = \frac{\dfrac{\sigma^2}{N}}{\dfrac{2(2N-1)\sigma^2}{N(N+1)}} = \frac{N+1}{2(2N-1)}$$

$$\frac{Var\left(\hat{A}\right)_{B \text{ known}}}{Var\left(\hat{A}\right)_{B \text{ unknown}}} = \frac{N+1}{2(2N-1)} \frac{\frac{1}{N}}{\frac{1}{N}} = \frac{1+\frac{1}{N}}{4-\frac{2}{N}} \to \frac{1}{4} \text{ as } N \to \infty$$

Therefore the knowledge of B reduce the uncertaininty (the variance of $\hat{A}$) by a factor of 4.

5 The estimation when A is known

Consider we have N observations

$$x[n] = A + Bn + w[n] \; n = 0, 1, ..., N-1$$

Now A is known, and we want ot estimate B.
Since now A is known, then we can rearrange the equation as

$$x[n] - A = Bn + w[n] \; n = 0, 1, ..., N-1$$

And re-name $x[n] - Bn$ as $z[n]$, thus the model becomes

$$z[n] = Bn + w[n] \; n = 0, 1, ..., N-1$$

Then this problem becomes the problem of estimating the slope of the model $x[n] = Bn + w[n]$.

The following will be the derivation of the Cramer-Rao Lower Bound for the estimator $\hat{B}$ for the model $x[n] = Bn + w[n]$.

Assume we have N observations

$$x[n] = Bn + w[n] \; n = 0, 1, ..., N-1$$

Their PDF will be

$$p(x; B) = \prod_{k=0}^{N-1} p(x[k]; B) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - Bn \right)^2 \right\}$$

The log-likelihood function will be

$$\ln p(x; B) = -\frac{N}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left(x[n] - Bn \right)^2$$

Take $\frac{\partial}{\partial B}$ and $\frac{\partial^2}{\partial B^2}$

$$\frac{\partial \ln p(x; B)}{\partial B} = \frac{1}{\sigma^2} \left[\sum_{n=0}^{N-1} \left(x[n] - Bn \right) \right] n$$

$$\frac{\partial^2 \ln p(x; B)}{\partial B^2} = \frac{-1}{\sigma^2} \sum_{n=0}^{N-1} n^2 = \frac{-N(N-1)(2N-1)}{\sigma}$$

The Fisher Information will be

$$I(B) = -\mathbb{E}\left[\frac{\partial^2 \ln p(x; B)}{\partial B^2} \right] = \frac{N(N-1)(2N-1)}{\sigma^2}$$

Thus the variance of the estimrator will be

$$Var\left(\hat{B}\right) = \frac{1}{I(B)} = \frac{\sigma^2}{N(N-1)(2N-1)}$$

Therefore, the variance of the estimator for known A is

$$Var\left(\hat{B}\right)_{A \text{ known}} \geq \frac{\sigma^2}{N(N-1)(2N-1)}$$

The Cramer-Rao Lower Bound for unknwon A is

$$Var\left(\hat{B}\right)_{A\text{ unknown}} \geq \frac{12\sigma^2}{N(N^2-1)}$$

Hence the knowledge of B will improve the Cramer-Rao Lower Bound by the factor of

$$\frac{Var\left(\hat{B}\right)_{A\text{ known}}}{Var\left(\hat{B}\right)_{A\text{ unknown}}} = \frac{\frac{\sigma^2}{N(N-1)(2N-1)}}{\frac{12\sigma^2}{N(N^2-1)}}$$

$$= \frac{N+1}{2N-1}\frac{\frac{1}{N}}{\frac{1}{N}} = \frac{1+\frac{1}{N}}{2-\frac{1}{N}} \to \frac{1}{2} \text{ as } N \to \infty$$

Therefore the knowledge of A reduce the uncertaininty (the variance of $\hat{B}$) by a factor of 2.