E7: Maximum Likelihood Estimation (MLE) ← will use the acronym for both.

The maximum likelihood estimator (MLE) has very attractive properties including asymptotic efficiency and a clear procedure for finding it. It is the most popular estimation method in practice.

Ex] DC Level in WGN (modified slightly)

Consider $x[n] = A + w[n]$, $n = 0, 1, \dots, (N-1)$ where A is unknown (with $A > 0$) and $w[n]$ is WGN with unknown variance A. With this model:

$$p(x;A) = \frac{1}{(2\pi A)^{N/2}} e^{-\frac{1}{2A}\left(-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n]-A)^2\right)}$$

Then

$$\frac{\partial \ln p(x;A)}{\partial A} = \frac{-N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n]-A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

$$\overset{?}{=} I(A)(\hat{A}-A) \quad \text{No obvious factorization}$$

We can find the CRLB: $\operatorname{var}(\hat{A}) \geq \dfrac{A^2}{N(A+\frac{1}{2})}$

Trying to use sufficient statistics:

$$p(x;A) = \frac{1}{(2\pi A)^{N/2}} e^{-\frac{1}{2}\left[\frac{1}{A}\sum_{n=0}^{N-1} x^2[n] + NA\right]} e^{N\bar{x}}$$

$$= g\Big(\underbrace{\sum_{n=0}^{N-1} x^2[n]}_{T(x)}, A\Big) \cdot h(x)$$

Can we find a function of $T(x)$ that is unbiased?

$$E\left\{ f\left(\sum_{n=0}^{N-1} x^2[n] \right) \right\} = A \quad \forall A > 0 \quad \text{is needed from } f(\cdot)$$

$$E\left[\sum_{n=0}^{N-1} x^2[n] \right] = N E[x^2[n]] = N\left(var(x[n]) + E^2[x[n]] \right) = N(A+A^2)$$
$$= N A (A+1)$$

A solution for $f(\cdot)$ is not obvious (e.g. scaling does not work). We need to try the second approach.

Let $\hat{A} = x[0]$. $E\left[x[0] \,\middle|\, \sum_{n=0}^{N-1} x^2[n]\right]$ should be MVU.
This is a very challenging conditional expectation to evaluate.

We will propose an estimator that is approximately optimal in the MVU sense. In particular, as $N \to \infty$, $E[\hat{A}] \to A$ and $var(\hat{A}) \to CRLB(\hat{A})$ will emerge as properties, so for large N $\hat{A}$ will be almost optimal. Such an $\hat{A}$ is <u>asymptotically efficient</u>.

Ex] Consider the proposition $\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$.

$$E[\hat{A}] = E\left[-\frac{1}{2} + \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}} \right] \neq A \quad \forall A. \quad \hat{A} \text{ is biased.}$$

However $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] = E[x^2[n]] = A + A^2$, so $\lim_{N\to\infty} \hat{A} = A$

So $\hat{A}$ is asymptotically unbiased. $\left\{ -\frac{1}{2} + \sqrt{A+A^2+\frac{1}{4}} = A \right.$?

Let $g(u) = -\frac{1}{2}\sqrt{u+\frac{1}{4}} \approx g(u_0) + \left.\frac{dg(u)}{du}\right|_{u=u_0} (u-u_0)$. Then

$$\hat{A} \approx A + \frac{(\frac{1}{2})}{A+(1/2)} \left[\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - (A+A^2) \right]$$

For large N, the linearization will be accurate.

Using the linearized approximation, the asymptotic variance is found:

$$\mathrm{var}(\hat{A}) = \left(\frac{1/2}{A+1/2}\right)^2 \mathrm{var}\left[\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right]$$

$$= \frac{1/4}{N(A+1/2)^2}\underbrace{\mathrm{var}(x^2[n])}_{=4A^3+2A^2} = \frac{A^2}{N(A+1/2)} = \mathrm{CRLB}(\hat{A})$$

∴ The proposed estimator is asymptotically unbiased and converges to the CRLB.
⇒ $\hat{A}$ is asymptotically efficient.

Defn: $\hat{\theta}_{ML} = \underset{\bar{\theta}}{\mathrm{argmax}}\ p(x;\bar{\theta}) = \underset{\bar{\theta}}{\mathrm{argmax}}\ \ln p(x;\bar{\theta})$

* It is easy to show that the local optimizers of an ~~function~~ objective and a monotonically increasing function of the objective are identical.

Let θ_1 locally maximize $p(x;\theta)$. Then $p(x;\theta_1) \geq p(x;\theta_1+\delta)\ \forall\delta$ in an open ball around 0. This implies that $\ln p(x;\theta_1) \geq \ln p(x;\theta_1+\delta)$ $\forall\delta$ in an open ball around 0, so θ_1 locally maximizes $\ln p(x;\theta)$. The converse is true, shown using the function $\ln^{-1}\cdot = e^{\cdot}$.

Ex) $p(x;A) = \dfrac{1}{(2\pi A)^{N/2}} e^{-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n]-A)^2}$ (same example as before)

The maximizer of $p(x;A)$ or equivalently $\ln p(x;A)$ is found by equating the derivative w.r.t. A to zero (and checking if at the solution the second derivative is negative).

$$\frac{\partial \ln p(x;A)}{\partial A} = \frac{-N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x[n]-A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

Equating to zero and solving for $\hat{A}_{ML}$: $\hat{A}_{ML}^2 + \hat{A}_{ML} - \frac{1}{N}\sum_{n=0}^{N-1} x^2[n] = 0$

$$\Rightarrow \hat{A}_{ML} = -\frac{1}{2} \pm \left(\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}\right)^{1/2}$$

We chose the solution with + instead of − to ensure $\hat{A}_{ML} > 0$.

Verify that for this $\hat{A}_{ML}$ value $\left.\frac{\partial^2 \ln p(x;A)}{\partial A^2}\right|_{A=\hat{A}_{ML}} < 0$.

Ex | DC Level in WGN

$x[n] = A + w[n]$, $n = 0, 1, \ldots, (N-1)$ and $\mathrm{var}(w[n]) = \sigma^2$.

$$p(x;A) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2}$$

$$\left.\frac{\partial \ln p(x;A)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) = 0\right|_{A=\hat{A}_{ML}} \Rightarrow \hat{A}_{ML} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

This is the MVU estimator

Fact: If an efficient estimator exists, MLE will produce it. (Problem 7.12)

Properties of the MLE

Thm 7.1 Asymptotic Properties of the MLE

If $p(x;\theta)$ satisfies some regularity conditions, then the MLE of θ is asymptotically distributed according to $\hat{\theta}_{ML} \stackrel{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta))$, where $I(\theta)$ is the Fisher information.

To simplify the proof let's assume $x[n]$ are iid. The following regularity conditions are assumed to hold.

1) The 1^{st} and 2^{nd} order derivatives of $\ln p(x;\theta)$ are well defined.
2) $E\left[\dfrac{\partial \ln p(x[n];\theta)}{\partial\theta}\right]=0$

* Show MLE is consistent: We will use $\int \ln \dfrac{p(x[n];\theta_1)}{p(x[n];\theta_2)}\, p(x[n];\theta_1)\,dx[n] \geq 0$ where equality holds iff $\theta_1=\theta_2$. This is the Kullback-Leibler divergence and this inequality is a direct consequence of Jensen's inequality for convex functions.

With iid samples $\frac{1}{N}\ln p(x;\theta) = \frac{1}{N}\ln \prod_{n=0}^{N-1} p(x[n];\theta)$

$$= \frac{1}{N}\sum_{n=0}^{N-1} \ln p(x[n];\theta)$$

$$\lim_{N\to\infty} \underbrace{\frac{1}{N}\sum_{n=0}^{N-1} \ln p(x[n];\theta)}_{\text{sample average}} = \underbrace{\int \ln p(x[n];\theta)\, p(x[n];\underline{\underline{\theta_0}})\,dx[n]}_{\text{expected value}}$$

where θ_0 is the true value.

From the $KLD \geq 0$ inequality above

$$\int \ln p(x[n];\theta_1)\,p(x[n];\theta_1)\,dx[n] \geq \int \ln p(x[n];\theta_2)\,p(x[n];\theta_1)\,dx[n]$$

Therefore $\int \ln p(x[n];\theta)\,p(x[n];\theta_0)\,dx[n]$ is maximized by choosing $\theta=\theta_0$. (Being slightly vague here), since the $\lim_{N\to\infty}$ on the left side must make $\hat{\theta}_N$ converge to θ_0 continuously if we were to maximize the sample average instead, $\hat{\theta}_N \to \theta_0$ as $N\to\infty$. Thus, the MLE is consistent.

* The asymptotic pdf of the MLE: We will use a Taylor series expansion about θ_0 of the first derivative of the log likelihood.

Taylor Thm uses the mean value theorem:

$$\left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta=\theta_0} + \left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\theta=\tilde{\theta}} (\hat{\theta}-\theta_0)$$

By definition of MLE,

$$\left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

There exists such $\tilde{\theta} \in (\theta_0, \hat{\theta})$ which makes exact = possible if $\ln p$ is twice cont. diff.

$$\Rightarrow 0 = \left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta=\theta_0} + \left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\theta=\tilde{\theta}} (\hat{\theta}-\theta_0) \quad \text{for some } \tilde{\theta}$$

From this expression, isolating $(\hat{\theta}-\theta_0)$:

$$\sqrt{N}(\hat{\theta}-\theta_0) = \frac{\frac{1}{\sqrt{N}} \cdot \left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta=\theta_0}}{-\frac{1}{N} \left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\theta=\tilde{\theta}}}$$

The denominator is $-\frac{1}{N} \sum_{n=0}^{N-1} \left.\frac{\partial^2 \ln p(x[n];\theta)}{\partial \theta^2}\right|_{\theta=\tilde{\theta}}$ due to iid x.

Since $\theta_0 < \tilde{\theta} < \hat{\theta}$ and $\hat{\theta}$ is consistent, we must have $\tilde{\theta} \to \theta_0$ as $N \to \infty$. So (continuing to work on the denominator)

$$-\frac{1}{N} \left.\frac{\partial^2 \ln p(x;\theta)}{\partial \theta^2}\right|_{\theta=\tilde{\theta}} \xrightarrow{N\to\infty} -\frac{1}{N} \sum_{n=0}^{N-1} \left.\frac{\partial^2 \ln p(x[n];\theta)}{\partial \theta^2}\right|_{\theta=\theta_0} \rightarrow -E\left[\left.\frac{\partial^2 \ln p(x[n];\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right]$$

$$= i(\theta_0)$$

Law of large numbers

The numerator is $\frac{1}{\sqrt{N}} \underbrace{\sum_{n=0}^{N-1} \frac{\partial \ln p(x[n];\theta)}{\partial \theta}\Big|_{\theta=\theta_0}}_{\text{This is a random variable.}}$ due to iid x.

Using the central limit theorem the numerator has a pdf that converges to a Gaussian with mean

$$E\left[\frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \frac{\partial \ln p(x[n];\theta)}{\partial \theta}\Big|_{\theta=\theta_0} \right] = 0$$ due to the 2nd regularity condition

and variance $E\left[\left(\frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \frac{\partial \ln p(x[n];\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2\right] = \frac{1}{N}\sum_{n=0}^{N-1} E\left[\left(\frac{\partial \ln p(x[n];\theta)}{\partial \theta}\Big|_{\theta=\theta_0}\right)^2\right]$

$= i(\theta_0)$ all $i(\theta_0)$ due to iid x.

<u>Thm</u> (Slutsky) If $x_n \overset{a}{\sim} x$ (x_n asymptotically distributed according to the pdf of x) and $y_n \to c$ (converges to a constant), then $x_n / y_n \overset{a}{\sim} x/c$.

In our case $x \sim \mathcal{N}(0, i(\theta_0))$ and $y_n \to c = i(\theta_0)$.

$$\therefore \sqrt{N}(\hat{\theta} - \theta_0) \overset{a}{\sim} \mathcal{N}(0, i^{-1}(\theta_0))$$ or equivalently

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{N\, i(\theta_0)}\right) = \mathcal{N}(\theta_0, I^{-1}(\theta_0)).$$

Ex | MLE of the Sinusoidal Phase

$$x[n] = A\cos(2\pi f_0 n + \phi) + w[n] \qquad n = 0, 1, \ldots, N-1$$

$w[n] \sim$ WGN with variance σ^2. A, f_0, σ^2 are known.

For this problem, we saw that two statistics are jointly sufficient:

$$T_1(x) = \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n), \quad T_2(x) = \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)$$

The MLE is found by maximizing $p(x;\phi)$:

$$p(x;\phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A\cos(2\pi f_0 n+\phi))^2}$$

Equivalently we can minimize w.r.t ϕ

$$J(\phi) = \sum_{n=0}^{N-1}(x[n] - A\cos(2\pi f_0 n+\phi))^2$$

$$\frac{\partial J(\phi)}{\partial \phi} = 2\sum_{n=0}^{N-1}\left(x[n]-A\cos(2\pi f_0 n+\phi)\right) A\sin(2\pi f_0 n+\phi)$$

Equating to zero and solving for $\hat{\phi}_{ML}$

$$\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n+\hat{\phi}_{ML}) = A\sum_{n=0}^{N-1}\sin(2\pi f_0 n+\hat{\phi}_{ML})\cos(2\pi f_0 n+\hat{\phi}_{ML})$$

The right hand side is approximately zero (for large N).

$$\frac{1}{N}\sum_{n=0}^{N-1}\sin(2\pi f_0 n+\hat{\phi}_{ML})\cos(2\pi f_0 n+\hat{\phi}_{ML}) = \frac{1}{2N}\sum_{n=0}^{N-1}\sin(4\pi f_0 n+2\hat{\phi}_{ML}) \approx 0$$

for f_0 not near 0 or $1/2$.

So $\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n+\hat{\phi}_{ML}) \approx 0$

$$\Rightarrow \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)\cos\hat{\phi}_{ML} \approx -\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)\sin\hat{\phi}_{ML}$$

$$\Rightarrow \hat{\phi}_{ML} \approx -\arctan\left(\frac{\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)}{\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)}\right)$$

Note that the (approximate) MLE turned out to be a function of the jointly sufficient statistics $T_1(x)$ and $T_2(x)$.

Recall the Neyman-Fisher factorization:

$$p(x;\phi) = g(T_1(x), T_2(x), \phi)\, h(x) \quad \begin{pmatrix} \text{choosing} \\ h(x)>0 \end{pmatrix}$$

$$\hat{\phi}_{ML} = \arg\max_{\phi} p(x;\phi) = \arg\max_{\phi} g(T_1(x), T_2(x), \phi)$$

$$\hat{\phi}_{ML} \overset{a}{\sim} \mathcal{N}(\phi, I^{-1}(\phi)) \text{ where } I(\phi) = \frac{NA^2}{2\sigma^2} \quad \begin{pmatrix} \text{previously} \\ \text{found} \end{pmatrix}$$

so $\mathrm{var}(\hat{\phi}_{ML}) = \dfrac{1}{N\left(\frac{A^2}{2\sigma^2}\right)} = \dfrac{1}{N \cdot SNR}$.

Ex] DC Level in Nonidependent Non-Gaussian Noise

$x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1$ where $w[n] \sim p_{w[n]}(\bullet)$.

The noise pdf is symmetric around zero ($p(\xi) = p(-\xi)$) and has a maximum at 0 ($p(0) > p(\xi) \;\forall \xi \neq 0$).

In the most extreme case assume that all noise samples are equal ($w[0] = w[1] = \cdots = w[N-1]$). Since all observations are identical, we only need to consider a single observation. Select $x[0]$ without loss of generality (w.l.o.g.).

The pdf of $x[0]$ is a shifted version of p, where the shift amount is A. $p_{x[0]}(x[0];A) = p_{w[0]}(x[0]-A)$.

Therefore, $\hat{A}_{ML} = \arg\max_{A} p_{w[0]}(x[0]-A) = x[0]$.

We have $E\{\hat{A}_{ML}\} = E\{x[0]\} = A$ and

$$\mathrm{var}(\hat{A}_{ML}) = \int_{-\infty}^{\infty} u^2 p_{w[0]}(u)\,du = \mathrm{var}(x[0]) = \mathrm{var}(w[0])$$

The CRLB can be found as

$$\mathrm{var}(\hat{A}) \geq \left[\int_{-\infty}^{\infty} \frac{(dp_{w[0]}(u)/du)^2}{p_{w[0]}(u)}\,du \right]^{-1}$$

and the variance is seen to not attain the CRLB in general for arbitrary $p_w(\cdot)$.

Clearly, dependent samples will cause estimation variance to reduce at a much smaller rate compared to independent data. In this extreme case, replicating (identical copies of data) does not reduce variance at all.

MLE for Transformed Parameters

Thm 7.2 Invariance Property of the MLE

The MLE of the parameter $\alpha = g(\theta)$, where $p(x;\theta)$ is parameterized by θ, is given by $\hat{\alpha}_{ML} = g(\hat{\theta}_{ML})$

The MLE of θ, $\hat{\theta}_{ML}$, is obtained by maximizing $p(x;\theta)$. If g is not a one-to-one function, then $\hat{\alpha}_{ML}$ maximizes the modified likelihood function $\bar{p}_T(x;\alpha)$ given by

$$\bar{p}_T(x;\alpha) = \max_{\{\theta:\,\alpha = g(\theta)\}} p(x;\theta).$$

* Prove as exercise.

Ex] Transformed DC Level in WGN

Consider the data $x[n] = A + w[n]$ $n = 0, 1, \ldots, (N-1)$

where $w[n] \sim$ WGN with var σ^2. $\alpha = e^A$. Find $\hat{\alpha}_{ML}$.

$$p(x; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2}, \quad -\infty < A < \infty$$

Since there is a 1-1 transformation between α & A, we can equivalently parameterize the pdf using α:

$$p_T(x; \alpha) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\ln\alpha)^2}, \quad \alpha > 0$$

Our signal model is now $x[n] = \ln\alpha + w[n]$.

Setting $\frac{\partial p_T(x;\alpha)}{\partial\alpha} = 0$ yields $\sum_{n=0}^{N-1}(x[n] - \ln\hat{\alpha}_{ML})\frac{1}{\hat{\alpha}_{ML}} = 0$

or $\hat{\alpha}_{ML} = e^{\bar{x}}$. Notice that $\bar{x}$ is the MLE of A; $\hat{A}_{ML} = \bar{x}$.

So we have $\hat{\alpha}_{ML} = e^{\hat{A}_{ML}}$.

Ex] Transformed DC Level in WGN (non 1-1)

Now consider $\alpha = A^2$. Since $A = \mp\sqrt{\alpha}$, the transformation is not 1-1. We require two sets of pdfs:

$$p_{T_1}(x; \alpha) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\sqrt{\alpha})^2}, \quad \alpha \geq 0$$

$$p_{T_2}(x; \alpha) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]+\sqrt{\alpha})^2}, \quad \alpha > 0$$

We need to find $\hat{\alpha}_{ML}$ which maximizes either one.

$$\hat{\alpha}_{ML} = \arg\max_{\alpha}\left\{\max\left(p_{T_1}(x;\alpha), p_{T_2}(x;\alpha)\right)\right\}.$$

$$\hat{\alpha}_{ML} = \arg\max_{\alpha \geq 0} \max\{ p(x;\sqrt{\alpha}), p(x;-\sqrt{\alpha}) \}$$

$$= \left[\arg\max_{\sqrt{\alpha} \geq 0} \max\{ p(x;\sqrt{\alpha}), p(x;-\sqrt{\alpha}) \} \right]^2$$

$$= \left[\arg\max_{-\infty < A < \infty} p(x;A) \right]^2 = \hat{A}_{ML}^2 = \bar{x}^2 .$$

Once again $\hat{\alpha}_{ML} = \hat{A}_{ML}^2$ indicates that the invariance property holds. In this case, however, given $\hat{\alpha}_{ML}$, we cannot identify $\hat{A}_{ML}$ uniquely.

Ex| Power of WGN in dB

Given N samples of WGN with variance σ^2, estimate the power of this process in dB. $P(\text{dB}) = 10 \log_{10} \sigma^2$.

$$p(x;\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n]}$$

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

$$= \left. -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} x^2[n] \right|_{\sigma^2 = \hat{\sigma}^2_{ML}} = 0$$

$$\Rightarrow \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] .$$

Using the invariance property, $\hat{P}_{ML} = 10 \log_{10} \left[\frac{1}{N} \sum_{n=0}^{N-1} x^2[n] \right]$

Ex | Exponential in WGN

$x[n] = r^n + w[n] \quad n = 0, 1, \ldots, (N-1)$, $w[n] \sim$ WGN with var σ^2

Estimate $r > 0$.

$$p(x;r) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-r^n)^2}$$

Maximizing $p(x;r)$ wrt r is the same as minimizing

$$J(r) = \sum_{n=0}^{N-1} (x[n] - r^n)^2$$

$$\frac{\partial J(r)}{\partial r} = 2\sum_{n=0}^{N-1} (x[n] - r^n) n r^{n-1} \Big|_{r = \hat{r}_{ML}} = 0$$

This is a nonlinear root finding problem for which we cannot determine the solution analytically. Numerical methods need to be employed.

For root finding, Newton-Raphson or ~~the~~ secant methods could be employed. Alternatively, we can minimize $J(r)$ numerically using gradient descent, Newton method or quasi Newton methods.

All of these numerical methods will find a local minimizer of $J(r)$. Global minimizers are harder to find.

Extension to a Vector Parameter

Let $p(x;\theta)$ be parameterized by a $p\times 1$ dim θ.

$\hat{\theta}_{ML} = \arg\max_{\theta} p(x;\theta)$ is now a multidimensional optimization problem. We must have $\left.\frac{\partial \ln p(x;\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{ML}} = 0$ if $p(x;\theta)$ is twice cont. differentiable. The Hessian of $p(x;\theta)$ or $\ln p(x;\theta)$ at $\hat{\theta}_{ML}$ will also need to be negative (semi) definite.

Ex DC Level in WGN

$x[n] = A + w[n] \quad n=0,\dots,(N-1) \quad w[n] \sim$ WGN with var σ^2.

$$\theta = \begin{bmatrix} A \\ \sigma^2 \end{bmatrix} \Rightarrow \frac{\partial \ln p(x;\theta)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)\Big|_{\hat{\sigma}^2_{ML},\hat{A}_{ML}} = 0$$

$$\frac{\partial \ln p(x;\theta)}{\partial \sigma^2} = \frac{-N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-A)^2\Big|_{\hat{\sigma}^2_{ML},\hat{A}_{ML}} = 0$$

From the first equation: $\hat{A}_{ML} = \bar{x}$.

From the second equation: $\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\bar{x})^2$

$$\therefore \hat{\theta}_{ML} = \begin{bmatrix} \bar{x} \\ \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\bar{x})^2 \end{bmatrix}$$

Thm 7.3 Asymptotic Properties of the MLE (vector param.)

If $p(x;\theta)$ is the pdf of data x, with parameters θ and it satisfies some regularity conditions, then the MLE of θ is asymptotically Gaussian:

$$\hat{\theta}_{ML} \overset{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta))$$

where $I(\theta)$ is the Fisher information matrix evaluated at θ_{true}.

Thm 7.4 Invariance Property of MLE (Vector parameters)

The MLE of $\alpha = g(\theta)$, where α is $r\times 1$, θ is $p\times 1$ and the data pdf is $p(x;\theta)$, is given by $\hat{\alpha}_{ML} = g(\hat{\theta}_{ML})$.

If g is not an invertible function, then $\hat{\alpha}_{ML}$ maximizes the modified likelihood function $\bar{p}_T(x;\alpha)$ given by

$$\bar{p}_T(x;\alpha) = \max_{\{\theta : \alpha = g(\theta)\}} p(x;\theta)$$

Thm 7.5 Optimality of the MLE for the Linear Model

If $x = H\theta + w$ where H is a known $N\times p$ matrix $(N>p)$ of rank p, θ is $p\times 1$, and w is $N\times 1$ with pdf $\mathcal{N}(0, C)$, then the MLE of θ is $\hat{\theta}_{ML} = (H^TC^{-1}H)^{-1}H^TC^{-1}x = \hat{\theta}_{MVU}$. The pdf is

$$\hat{\theta}_{ML} \sim \mathcal{N}(\theta, (H^TC^{-1}H)^{-1}).$$

We can employ any numerical optimization technique to solve for $\hat{\theta}_{ML}$. A popular method that does not stem from numerical optimization theory is the Expectation-Maximization (EM) Algorithm.

EM assumes that a hypothetical dataset y for which the determination of MLE is easier exists. The original data is called the incomplete data. Suppose a transformation exists such that $x = g(y)$. Here g could be a many-to-one transformation. We would like to find $\hat{\theta}_{ML} = \arg\max_{\theta} \ln p_x(x;\theta)$, but this is difficult. Instead, if we had y, we could solve $\hat{\theta}_{ML} = \arg\max_{\theta} \ln p_y(y;\theta)$. Since y is unavailable, consider instead

$$E_{y|x}[\ln p_y(y;\theta)] = \int \ln p_y(y;\theta)\, p(y|x;\theta)\, dy .$$

One iteration of EM has the following two steps

<u>E step</u>: Determine $U(\theta,\theta_k) = \int \ln p_y(y;\theta)\, p(y|x;\theta_k)\, dy$

<u>M step:</u> Let $\theta_{k+1} = \arg\max_{\theta} U(\theta,\theta_k)$

Ex] $x[n] = \sum_{i=1}^{p} \cos(2\pi f_i n) + w[n]$ $\quad n = 0, \ldots, (N-1)$

$w[n] \sim WGN$, var σ^2

Estimate $f = [f_1 \cdots f_p]^T$. Due to AWGN, MLE reduces to

$$\hat{f}_{ML} = \arg\min_f J(f) \quad \text{where } J(f) = \sum_{n=0}^{N-1} \left(x[n] - \sum_{i=1}^{p} \cos(2\pi f_i n) \right)^2$$

If we had access to $y_i[n] = \cos 2\pi f_i n + w_i[n]$ $\quad i=1,\ldots,p;\ n=0,\ldots,N-1$

where $w_i[n]$ is WGN with var σ_i^2, then the problem would be decoupled and each f_i could be estimated individually.

$\{y_1[n], \ldots, y_p[n]\}$ is the complete data and $x[n] = \sum_{i=1}^{p} y_i[n]$, $w[n] = \sum_{i=1}^{p} w_i[n]$. For this to hold, we need $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

We have $\ln p_y(y;\theta) = \sum_{i=1}^{p} \ln p(y_i;\theta_i)$

$$= \sum_{i=1}^{p} \ln \left[\frac{1}{(2\pi\sigma_i^2)^{N/2}} e^{-\frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (y_i[n] - \cos(2\pi f_i n))^2} \right]$$

c is a constant

$g(y)$ does not depend on f.

$$= c - \sum_{i=1}^{p} \frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} (y_i[n] - \cos(2\pi f_i n))^2$$

$$= g(y) + \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} \left(y_i[n] \cos(2\pi f_i n) - \frac{1}{2}\cos^2(2\pi f_i n) \right)$$

Using $\sum_{n=0}^{N-1} \cos^2(2\pi f_i n) \approx \frac{N}{2}$, we have

$$\ln p_y(y;\theta) \approx h(y) + \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} y_i[n] \cos(2\pi f_i n)$$

and letting $c_i = [1 \ \ \cos(2\pi f_i) \ \cdots \ \cos(2\pi f_i (N-1))]^T$

$$\ln p_y(y;\theta) = h(y) + \sum_{i=1}^{p} \frac{1}{\sigma_i^2} c_i^T y_i$$

vectorized data $y_i[n]$.

Let $c = \left[\frac{c_1^T}{\sigma_1^2} \cdots \frac{c_p^T}{\sigma_p^2} \right]^T$ and $y = [y_1^T \cdots y_p^T]^T$.

$$\ln p_y(y;\theta) = h(y) + c^T y.$$

$$U(\theta,\theta_k) = E[\ln p_y(y;\theta) \mid x;\theta_k] = E[h(y) \mid x;\theta_k] + c^T E[y \mid x;\theta_k]$$

Since $E[h(y) \mid x;\theta_k]$ does not depend on θ and in the M step we will maximize $U(\theta,\theta_k)$ wrt θ, we don't need to evaluate the first term.

Since $x = \sum_{i=1}^{p} y_i = [\underbrace{I \cdots I}_{N\times N \text{ identity matrices}}]\, y$, x & y are jointly Gaussian. Then

$$E[y \mid x;\theta_k] = E[y] + C_{yx} C_{xx}^{-1} (x - E[x])$$
$$= \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} + C_{yx} C_{xx}^{-1} \left(x - \sum_{i=1}^{p} c_i\right)$$

We have $C_{xx} = \sigma^2 I$ and $C_{yx} = E\left[\begin{pmatrix} w_1 \\ \vdots \\ w_p \end{pmatrix} w^T\right] = \cdots = \begin{bmatrix} \sigma_1^2 I \\ \vdots \\ \sigma_p^2 I \end{bmatrix}$

Then $E[y \mid x;\theta_k] = \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} + \frac{1}{\sigma^2} \begin{bmatrix} \sigma_1^2 I \\ \vdots \\ \sigma_p^2 I \end{bmatrix} \left(x - \sum_{i=1}^{p} c_i\right)$ $w = \sum_{i=1}^{p} w_i$

or $E[y_i \mid x;\theta_k] = c_i + \frac{\sigma_i^2}{\sigma^2} \left(x - \sum_{i=1}^{p} c_i\right)$ $i = 1, \ldots, p$.

Here c_i is computed using θ_k (current frequency estimates).

Define $\hat{y}_i \triangleq E[y_i \mid x;\theta_k]$:

$$\hat{y}_i[n] = \cos(2\pi f_{i_k} n) + \frac{\sigma_i^2}{\sigma^2}\left(x[n] - \sum_{i=1}^{p} \cos(2\pi f_{i_k} n)\right)$$

iteration index.

Let $U'(\theta, \theta_k) = \sum_{i=1}^{p} c_i^T y_i$ (the part that depends on θ).

Then in the M step $f_{i_{k+1}} = \arg\max_{f_i} c_i^T \hat{y}_i$. Since σ_i^2 are not unique, they can be chosen arbitrarily as long as $\sum_{i=1}^{p} \sigma_i^2 = \sigma^2$, or $\sum_{i=1}^{p} \beta_i = \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma^2} = 1$.

In summary, we get

E step: $\hat{y}_i[n] = \cos(2\pi f_{i_k} n) + \beta_i \left(x[n] - \sum_{i=1}^{p} \cos(2\pi f_{i_k} n) \right)$

M step: For $i = 1, \dots, p$ $f_{i_{k+1}} = \arg\max_{f_i} \sum_{n=0}^{N-1} \hat{y}_i[n] \cos(2\pi f_i n)$

where β_i's are selected arbitrarily but with $\sum_{i=1}^{p} \beta_i = 1$.

Disadvantages: The choice of complete data definitions is arbitrary. Determining the conditional expectation in the E step may be difficult. Slow to converge.

My interpretation: Think of it like increasing the dimensionality of the parameters and optimizing the slack variables (extra parameters, the complete data by minimizing the least squares objective and then optimizing ~~then~~ the original parameters by maximizing ~~the~~ log-likelihood. If the slack variables are selected appropriately, the alternating optimization algorithms work nicely together.

Asymptotic MLE

For a high dimensional data pdf that is Gaussian, the computation of likelihood requires inverting a large covariance matrix. Consider $x \sim \mathcal{N}(0, C(\theta))$; $p(x;\theta) = \frac{1}{(2\pi)^{N/2} |C(\theta)|^{1/2}} e^{-\frac{1}{2} x^T C^{-1}(\theta) x}$.

If $C^{-1}(\theta)$ cannot be evaluated in a symbolic fashion, then for each θ, the inverse will have to be computed from scratch numerically.

When x is data from a 0-mean WSS random process, the covariance matrix is Toeplitz and has the autocorrelation sequence values in its diagonals. (Review in Appendix 1.) In that case, the asymptotic log-likelihood function is given by (as shown in Chapter 3), as N gets large:

$$\ln p(x;\theta) = \frac{-N}{2} \ln(2\pi) - \frac{N}{2} \int_{-1/2}^{1/2} \left[\ln P_{xx}(f) + \frac{I(f)}{P_{xx}(f)} \right] df$$

where $I(f) = \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-j2\pi fn} \right|^2$ is the periodogram of the data and $P_{xx}(f)$ is the PSD of the random process $x[n]$. As we have seen earlier in (E3), the PSD is parameterized by θ, which is not shown explicitly here.

At the (asymptotic) MLE solution, the gradient $\nabla_\theta^T \ln p(x;\theta) = 0$. Specifically, $\frac{\partial \ln p(x;\theta)}{\partial \theta_i} = \frac{-N}{2} \int_{-1/2}^{1/2} \left[\frac{1}{P_{xx}(f)} - \frac{I(f)}{P_{xx}^2(f)} \right] \frac{\partial P_{xx}(f)}{\partial \theta_i} df = 0 \; \forall i$.

The Hessian $\nabla_\theta^T \nabla_\theta \ln p(x;\theta)$ is also given in (E3 ; App 3D in the book). These could be used as asymptotic approximations during the numerical optimization of the maximum likelihood problem w.r.t. θ.

Ex Gaussian Moving Average Process

Suppose $x[n]$ is generated by WGN passing through an FIR filter with impulse response $b[n]$. Then the autocorr of $x[n]$ is

$$r_{xx}[k] = \begin{cases} 1 + b^2[1] + b^2[2] &, k=0 \\ b[1] + b[1]b[2] &, k=1 \\ b[2] &, k=2 \\ 0 &, k \geq 0 \end{cases}$$

for some particular $b[n]$.

The (auto) PSD of x is the Fourier transform of $r_{xx}[k]$:

$$P_{xx}(f) = \left| 1 + b[1] e^{-j2\pi f} + b[2] e^{-j4\pi f} \right|^2$$

Here the FIR filter ~~sys~~ transfer function is $\mathcal{B}(z) = 1 + b[1]z^{-1} + b[2]z^{-2}$.

Assume that $\mathcal{B}(z)$ is a minimum phase filter (so that the zeros are inside the unit circle.). Let the zeros be z_1, z_2.

In order to estimate $\theta = \begin{bmatrix} b[1] \\ b[2] \end{bmatrix}$, we need to invert the Toeplitz data covariance matrix.

$$\ln p(x;\theta) = \frac{-N}{2} \ln(2\pi) - \frac{N}{2} \int_{-1/2}^{1/2} \left[\ln P_{xx}(f) + \frac{I(f)}{P_{xx}(f)} \right] df$$

The first term is a constant (w.r.t θ). Since $\mathcal{B}(z)$ is min-phase

$\int_{-1/2}^{1/2} \ln P_{xx}(f)\, df = 0$ (using Cauchy Residue Theorem; see P 7.22).

Then $\hat{\theta}_{ML} = \arg\min_{\theta} \int_{-1/2}^{1/2} \frac{I(f)}{P_{xx}(f)} df$. Substituting $P_{xx}(f) = \left| \mathcal{B}(z) \big|_{z=e^{j2\pi f}} \right|^2$

$$= \arg\min_{\theta} \int_{-1/2}^{1/2} \frac{I(f)}{|1 - z_1 e^{-j2\pi f}|^2 \cdot |1 - z_2 e^{-j2\pi f}|^2} df$$

where θ and $\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ are related through $\begin{matrix} b[1] = -(z_1 + z_2) \\ b[2] = z_1 z_2 \end{matrix}$

Ex | Range Estimation

In (active) radar or sonar, a signal is transmitted and the roundtrip delay τ_0 for the received echo is used to calculate range. If $s(t)$ is transmitted, $x(t) = s(t-\tau_0) + w(t)$, $0 \le t \le T$ is received. Using a bandlimited signal and assuming bandlimited Gaussian noise (PSD of $w(t)$ is $N_0/2$ for $F(Hz) \in [-B, B]$, 0 o.w.), we can sample the received cont. time signal every $\Delta = \frac{1}{2B}$ (~~or faster~~)

to get $x(n\Delta) = s(n\Delta - \tau_0) + w(n\Delta)$, $n = 0, 1, \ldots, (N-1)$.

In DT notation $x[n] = s(n\Delta - \tau_0) + w[n]$. Here $w[n]$ is WGN (since we sample exactly at the Nyquist rate) with var $\sigma^2 = r_{ww}(0) = N_0 B$.

The signal is nonzero only over $\tau_0 \le t \le \tau_0 + T_s$, so

$$x[n] = \begin{cases} w[n] & 0 \le n \le n_0 - 1 \\ s(n\Delta - \tau_0) + w[n] & n_0 \le n \le n_0 + M - 1 \\ w[n] & n_0 + M \le n \le N - 1 \end{cases}$$

where M is the length of the sampled version of $s(t-\tau_0)$ and $n_0 = \tau_0/\Delta$ (assuming τ_0/Δ is an integer). We want to estimate n_0.

$$p(x; n_0) = \prod_{n=0}^{n_0-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} x^2[n]} \prod_{n=n_0}^{n_0+M-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x[n] - s[n-n_0])^2} \prod_{n=n_0+M}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} x^2[n]}$$

$$= \underbrace{\frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x^2[n]}}_{\text{does not depend on } n_0} \prod_{n=n_0}^{n_0+M-1} e^{-\frac{1}{2\sigma^2}\left(-2x[n]s[n-n_0] + s^2[n-n_0]\right)}$$

$$\hat{n}_{0_{ML}} = \arg\min_{n_0} \sum_{n=n_0}^{n_0+M-1} \left[-2x[n]s[n-n_0] + s^2[n-n_0]\right]$$

since $\sum_{n=n_0}^{n_0+M-1} s^2[n-n_0] = \underbrace{\sum_{n=0}^{N-1} s^2[n]}_{\text{does not depend on } n_0}$

$$= \arg\max_{n_0} \sum_{n=n_0}^{n_0+M-1} x[n]s[n-n_0]$$

Matched filter.

Ex| Sinusoidal Parameter Estimation

$$p(x;\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A\cos(2\pi f_0 n+\phi))^2}$$

where $A>0$, $0<f_0<1/2$. Estimate A, f_0, ϕ. $\theta \triangleq \begin{bmatrix} A \\ f_0 \\ \phi \end{bmatrix}$

Maximizing the log-likelihood is equivalent to minimizing

$$J(\theta) = \sum_{n=0}^{N-1}(x[n]-A\cos(2\pi f_0 n+\phi))^2 \quad \text{(least squares)}$$

$$= \sum_{n=0}^{N-1}\left(x[n] - A\cos\phi\cos(2\pi f_0 n) + A\sin\phi\sin(2\pi f_0 n)\right)^2$$

Letting $\alpha_1 = A\cos\phi$, $\alpha_2 = -A\sin\phi$, which is an invertible parameter transformation (with inverse $A = (\alpha_1^2+\alpha_2^2)^{1/2}$, $\phi = \arctan\left(\frac{-\alpha_2}{\alpha_1}\right)$) and also letting $c = [1, \cos(2\pi f_0), \ldots, \cos(2\pi f_0(N-1))]^T$ and $s = [0, \sin(2\pi f_0), \ldots, \sin(2\pi f_0(N-1))]^T$, we get

$$J'(\alpha_1, \alpha_2, f_0) = (x - \alpha_1 c - \alpha_2 s)^T(x - \alpha_1 c - \alpha_2 s)$$

$$= (x - H\alpha)^T(x - H\alpha) \quad \text{where } \alpha = \begin{bmatrix}\alpha_1 \\ \alpha_2\end{bmatrix},\ H = [c, s].$$

Given H, the optimal α is found, as in the linear model with $C = I$, using $\hat{\alpha} = (H^TH)^{-1}H^Tx$. Then

$$J'(\hat{\alpha}, f_0) = (x - H\hat{\alpha})^T(x - H\hat{\alpha})$$

$$= (x - H(H^TH)^{-1}H^Tx)^T(x - H(H^TH)^{-1}H^Tx)$$

$$= x^T\left(I - H(H^TH)^{-1}H^T\right)x = x^TAx.$$

Note that $A^2 = A$ (A is idempotent).

$$\hat{f}_0 = \arg\min_{f_0} x^TAx = \arg\max_{f_0} x^TH(H^TH)^{-1}H^Tx$$

$$x^TH(H^TH)^{-1}H^Tx = \begin{bmatrix} c^Tx \\ s^Tx \end{bmatrix}^T \begin{bmatrix} c^Tc & c^Ts \\ s^Tc & s^Ts \end{bmatrix}^{-1} \begin{bmatrix} c^Tx \\ s^Tx \end{bmatrix}$$

Once $\hat{f}_0$ is found, then $\hat{\alpha}$ can be calculated directly.
From $\hat{\alpha}$, $\hat{A}$ and $\hat{\phi}$ can be obtained.

If f_0 is not near 0 or 1/2, then

$$\frac{1}{N} c^T s = \frac{1}{N} \sum_{n=0}^{N-1} \cos(2\pi f_0 n) \sin(2\pi f_0 n)$$

$$= \frac{1}{2N} \sum_{n=0}^{N-1} \sin(4\pi f_0 n) \approx 0 \quad \text{for large } N.$$

Similarly, $\frac{1}{N} c^T c \approx \frac{1}{2}$ and $\frac{1}{N} s^T s \approx \frac{1}{2}$. Then

$$\begin{bmatrix} c^T x \\ s^T x \end{bmatrix}^T \begin{bmatrix} N/2 & 0 \\ 0 & N/2 \end{bmatrix} \begin{bmatrix} c^T x \\ s^T x \end{bmatrix} = \frac{2}{N} \left[(c^T x)^2 + (s^T x)^2 \right]$$

$$= \frac{2}{N} \left[\left(\sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n) \right)^2 + \left(\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n) \right)^2 \right]$$

$$= \frac{2}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-j 2\pi f_0 n} \right|^2 = 2 I(f_0)$$

periodogram.

Then $\hat{f}_0 = \arg\max_{f_0} I(f_0)$. Immediately after

$$\hat{\alpha} \approx \frac{2}{N} \begin{bmatrix} \hat{c}^T x \\ \hat{s}^T x \end{bmatrix} = \begin{bmatrix} \frac{2}{N} \sum_{n=0}^{N-1} x[n] \cos(2\pi \hat{f}_0 n) \\ \frac{2}{N} \sum_{n=0}^{N-1} x[n] \sin(2\pi \hat{f}_0 n) \end{bmatrix}$$

$$\therefore \hat{A} = (\hat{\alpha}_1^2 + \hat{\alpha}_2^2) = \frac{2}{N} \left| \sum_{n=0}^{N-1} x[n] e^{-j 2\pi \hat{f}_0 n} \right|$$

$$\hat{\phi} = \arctan \frac{-\sum_{n=0}^{N-1} x[n] \sin(2\pi \hat{f}_0 n)}{\sum_{n=0}^{N-1} x[n] \cos(2\pi \hat{f}_0 n)}$$

where $\hat{A}$, $\hat{f}_0$, $\hat{\phi}$ are (approximate) MLE.

Ex) AR Parameter Estimation

$$\underset{\text{WGN}}{u[n]} \longrightarrow \boxed{\frac{1}{A(z)}} \longrightarrow x[n] \qquad P_{xx}(f) = \frac{\sigma_u^2}{|A(f)|^2}$$

$A(z)$ min phase

$$\ln p(x; a, \sigma_u^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\int_{-1/2}^{1/2}\left[\ln \frac{\sigma_u^2}{|A(f)|^2} + \frac{I(f)}{\left(\frac{\sigma_u^2}{|A(f)|^2}\right)}\right] df$$

$A(z)$ is min phase, since $\frac{1}{A(z)}$ is assumed to be stable (causal).

Then $\int_{-1/2}^{1/2} \ln|A(f)|^2 df = 0$.

$$\ln p(x; a, \sigma_u^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma_u^2 - \frac{N}{2\sigma_u^2}\int_{-1/2}^{1/2}|A(f)|^2 I(f)\, df$$

$\partial / \partial \sigma_u^2$ $\rightsquigarrow$ $= 0$

$$\frac{-N}{2\sigma_u^2} + \frac{N}{2\sigma_u^4}\int_{-1/2}^{1/2}|A(f)|^2 I(f)\, df = 0 \Rightarrow \hat{\sigma}_u^2 = \int_{-1/2}^{1/2}|A(f)|^2 I(f)\, df$$

Then $\ln p(x; a, \hat{\sigma}_u^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\hat{\sigma}_u^2 - \frac{N}{2}$.

To find $\hat{a}$, we need to maximize $\ln p(x; a, \hat{\sigma}_u^2)$ or equivalently minimize $\hat{\sigma}_u^2$. So

$$\hat{a} = \arg\min \underbrace{\int_{-1/2}^{1/2}|A(f)|^2 I(f)\, df}_{J(a)}$$

Note that $J(a)$ is quadratic in a, with a pos. def Hessian.

$$\frac{\partial J(a)}{\partial a[k]} = \int_{-1/2}^{1/2}\left[A(f)\frac{\partial A^*(f)}{\partial a[k]} + \frac{\partial A(f)}{\partial a[k]}A^*(f)\right] I(f)\, df$$

$$= \int_{-1/2}^{1/2}\left[A(f)e^{j2\pi fk} + A^*(f)e^{-j2\pi fk}\right] I(f)\, df$$

since $A(-f) = A^*(f)$ & $I(-f) = I(f)$

$$= 2\int_{-1/2}^{1/2} A(f) I(f) e^{j2\pi fk}\, df = 0 \quad \text{at } \hat{a}.$$

$$\therefore \int_{-1/2}^{1/2}\left(1 + \sum_{l=1}^{p} a[l] e^{-j2\pi fl}\right) I(f) e^{j2\pi fk}\, df = 0 \qquad k = 1, 2, \ldots, p$$

$$\Rightarrow \sum_{l=1}^{p} a[l] \int_{-1/2}^{1/2} I(f) e^{j2\pi f(k-l)}\, df = -\int_{-1/2}^{1/2} I(f) e^{j2\pi fk}\, df$$

Since $\int_{-1/2}^{1/2} I(f) e^{j2\pi fk} df$ is the inverse Fourier transform of the periodogram, evaluated at k, we have

$$\sum_{l=1}^{p} \hat{a}[l] \hat{r}_{xx}[k-l] = -\hat{r}_{xx}[k] \qquad k=1,\dots,p$$

where $\hat{r}_{xx}[k] = \begin{cases} \frac{1}{N} \sum_{n=0}^{N-1-|k|} x[n]x[n+|k|] & , |k| \le N-1 \\ 0 & , |k| \ge N \end{cases}$

or in matrix form $R\hat{a} = -p$ where $R_{ij} = \hat{r}_{xx}[i-j]$ $i,j=1,\dots,p$

and $\hat{a} = [\hat{a}[1] \dots \hat{a}[p]]^T$. $\qquad p_i = \hat{r}_{xx}[i]$ $i=1,\dots,p$

This is called the (estimated) Yule-Walker equation and results in linear prediction with least squares objective. A recursive solution (using Levinson's algorithm) is possible, so model order can be recursively searched over and optimized (see Brooks, DSP).

$$\hat{\sigma}_u^2 = \int_{-1/2}^{1/2} |\hat{A}(f)|^2 I(f) df = \int_{-1/2}^{1/2} \hat{A}(f) I(f) \hat{A}^*(f) df$$

$$= \sum_{k=0}^{p} \hat{a}[k] \int_{-1/2}^{1/2} \hat{A}(f) I(f) e^{j2\pi fk} df$$

but $\int_{-1/2}^{1/2} \hat{A}(f) I(f) e^{j2\pi fk} df = 0$ $k=1,2,\dots,p$ $\left(\text{from } \frac{\partial J(a)}{\partial a}\Big|_{a=\hat{a}} = 0\right)$.

$\therefore$ $\hat{\sigma}_u^2 = \int_{-1/2}^{1/2} \hat{A}(f) I(f) df$ (only the $\hat{a}[0]=1$ term remains for $k=0$)

$$= \sum_{k=0}^{p} \hat{a}[k] \int_{-1/2}^{1/2} I(f) e^{-j2\pi fk} df = \sum_{k=0}^{p} \hat{a}[k] \hat{r}_{xx}[-k]$$

$$= \sum_{k=0}^{p} \hat{a}[k] \hat{r}_{xx}[k].$$

Suggested Problems: 3, 5, 6, 9, 12, 17, 22, 23, 26