E9: Method of Moments

This estimation method does not have any optimality properties, but it usually yields consistent estimators. For large number of samples, it could give an easy-to-implement estimate, which could also be used as an initial estimate (for MLE, for instance).

Ex] Suppose that we observe $x[n] \overset{iid}{\sim}$ Gaussian Mixture Model for $n = 0, 1, \ldots, (N-1)$. For illustration, let the pdf be a mixture of two Gaussians

$$p(x[n]; \epsilon) = \frac{1-\epsilon}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{x^2[n]}{2\sigma_1^2}} + \frac{\epsilon}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{x^2[n]}{2\sigma_2^2}}$$

$\boxed{0 < \epsilon < 1}$

$$= (1-\epsilon)\phi_1(x[n]) + \epsilon\, \phi_2(x[n])$$

where $\phi_i(x[n])$ is the i^{th} Gaussian component $(i = 1, 2)$. Assume that σ_1^2 and σ_2^2 are known, and ϵ is to be estimated. Consider the 2^{nd} moment of $x[n]$:

$$E[x^2[n]] = \int_{-\infty}^{\infty} x^2[n] \left[(1-\epsilon)\phi_1(x[n]) + \epsilon\, \phi_2(x[n])\right] dx[n]$$

since ϕ_i are zero-mean Gaussians

$$= (1-\epsilon)\sigma_1^2 + \epsilon\, \sigma_2^2$$

Using the sample estimator $\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$, we get

$$\hat{\epsilon} = \left(\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] - \sigma_1^2\right) \Big/ (\sigma_2^2 - \sigma_1^2)$$

In this example, $E[\hat{\epsilon}] = \epsilon$ and

$$\text{var}(\hat{\epsilon}) = \frac{1}{(\sigma_2^2-\sigma_1^2)} \text{var}\left(\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right) = \frac{\text{var}(x^2[n])}{N(\sigma_2^2-\sigma_1^2)^2}$$

$$= \frac{E(x^4[n]) - E^2(x^2[n])}{N(\sigma_2^2-\sigma_1^2)^2} = \frac{3(1-\epsilon)\sigma_1^4 + 3\epsilon\sigma_2^4 - [(1-\epsilon)\sigma_1^2 + \epsilon\sigma_2^2]^2}{N(\sigma_2^2-\sigma_1^2)^2}$$

Clearly $\lim_{N\to\infty} \text{var}(\hat{\epsilon}) = 0$ if $\sigma_1^2 \neq \sigma_2^2$ since $E[x^4[n]] = (1-\epsilon)3\sigma_1^4 + \epsilon 3\sigma_2^4$

Method of Moments for a Scalar Parameter

Let $\mu_k \triangleq E(x^k[n]) = h(\theta)$. If h^{-1} exists, then $\theta = h^{-1}(\mu_k)$.

Using the sample average $\hat{\mu}_k = \frac{1}{N}\sum_{n=0}^{N-1} x^k[n]$, $\hat{\theta} = h^{-1}\left(\frac{1}{N}\sum_{n=0}^{N-1} x^k[n]\right)$.

Ex) Exponential pdf : $p(x[n];\lambda) = \lambda e^{-\lambda x[n]} u(x[n])$ ← unit step.

$$\mu_1 = E(x[n]) = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}\int_0^\infty \xi e^{-\xi}\,d\xi = \lambda^{-1}.$$

$$\Rightarrow \lambda = \mu_1^{-1} \text{ and } \hat{\lambda} = \left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right)^{-1}.$$

Extension to a Vector Parameter

Let θ be a $p\times 1$ parameter vector for the data pdf.

$$\left.\begin{aligned} \mu_1 &= h_1(\theta_1 \cdots \theta_p) \\ \mu_2 &= h_2(\theta_1 \cdots \theta_p) \\ &\vdots \\ \mu_p &= h_p(\theta_1 \cdots \theta_p) \end{aligned}\right\}$$

$\mu = h(\theta)$ where $h: \mathbb{R}^p \to \mathbb{R}^p$ (or $\mathbb{C}^p \to \mathbb{C}^p$) would be needed to solve for θ.

$$\theta = h^{-1}(\mu)$$

Then $\hat{\theta} = h^{-1}(\hat{\mu})$ where $\hat{\mu}$ is the vector of sample (average) estimators of the moments.

Ex] GMM

Consider the mixture of two Gaussians we had earlier. Now $\epsilon, \sigma_1^2, \sigma_2^2$ are all unknown. Then, using

$$\mu_2 = E\left(x^2[n]\right) = (1-\epsilon)\sigma_1^2 + \epsilon\sigma_2^2$$

$$\mu_4 = E\left(x^4[n]\right) = 3(1-\epsilon)\sigma_1^4 + 3\epsilon\sigma_2^4$$

$$\mu_6 = E\left(x^6[n]\right) = 15(1-\epsilon)\sigma_1^6 + 15\epsilon\sigma_2^6$$

we can estimate $\theta = [\epsilon, \sigma_1^2, \sigma_2^2]^T$. Let $u = \sigma_1^2 + \sigma_2^2$, $v = \sigma_1^2\sigma_2^2$.

Then $u = \dfrac{\mu_6 - 5\mu_4\mu_2}{5\mu_4 - 15\mu_2^2}$ and $v = \mu_2 u - \dfrac{\mu_4}{3}$. Determining $\hat{u}, \hat{v}$ in this order then yields $\hat{\sigma}_1$ and $\hat{\sigma}_2$ from

$\sigma_1^2 = \dfrac{u + \sqrt{u^2 - 4v}}{2}$ and $\sigma_2^2 = \dfrac{v}{\sigma_1^2}$. Finally, $\epsilon = \dfrac{\mu_2 - \sigma_1^2}{\sigma_2^2 - \sigma_1^2}$, all evaluated using the sample estimators of the moments.

Statistical Evaluation

We have $\hat{\theta} = h^{-1}(\hat{\mu}) = \tilde{g}(x) = g(T(x))$ where $T(x) = \begin{bmatrix} T_1(x) \\ \vdots \\ T_r(x) \end{bmatrix}$

where $\hat{\mu}_i = T_i(x)$. Then (with $E[T] = \mu$) using 1st order Taylor expansion $\hat{\theta} = g(T) \approx g(\mu) + \sum_{k=1}^{r} \dfrac{\partial g}{\partial T_k}\Big|_{T=\mu} (T_k - \mu_k)$, which would hold under some circumstances (unbiased, small variance), we get $E[\hat{\theta}] = g(\mu)$.

$$\mathrm{Var}\,[redacted](\hat{\theta}) = E\left[\left(g(\mu) + \frac{\partial g}{\partial T}\Big|_{T=\mu}^{T}(T-\mu) - E(\hat{\theta})\right)^2\right]$$

(for a scalar θ)

$$= \frac{\partial g}{\partial T}\Big|_{T=\mu}^{T} C_T \frac{\partial g}{\partial T}\Big|_{T=\mu}$$

since $E(\hat{\theta}) = g(\mu)$, $C_T = \mathrm{Cov}(T(x))$

Ex) Exponential pdf

Recall that we had $\hat{\lambda} = \left(\frac{1}{N} \sum_{n=0}^{N-1} x[n] \right)^{-1}$.

Let $T_1 = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$. Then $\hat{\lambda} = g(T_1)$ where $g(T_1) = \frac{1}{T_1}$

$$\mu_1 = E(T_1) = \frac{1}{N} \sum_{n=0}^{N-1} E(x[n]) = E[x[n]] = \frac{1}{\lambda}.$$

$$\text{Var}(T_1) = \text{var}\left(\frac{1}{N} \sum_{n=0}^{N-1} x[n] \right) = \frac{1}{N} \text{var}(x[n])$$

$$= \frac{1}{N} \left[\frac{2}{\lambda^2} - \frac{1}{\lambda^2} \right] = \frac{1}{N\lambda^2}$$

$$g(T_1) = \frac{1}{T_1} \Rightarrow \left. \frac{\partial g}{\partial T_1} \right|_{T_1 = \mu_1} = -\frac{1}{\mu_1^2} = -\lambda^2$$

Then $E\{\hat{\lambda}\} = g(\mu_1) = \frac{1}{(1/\lambda)} = \lambda$ and

$$\text{var}(\hat{\lambda}) = \left. \frac{\partial g}{\partial T_1} \right|_{T_1 = \mu_1} \text{var}(T_1) \left. \frac{\partial g}{\partial T_1} \right|_{T_1 = \mu_1} = (-\lambda^2) \frac{1}{N\lambda^2} (-\lambda^2) = \frac{\lambda^2}{N}$$

$\therefore$ In this case $\hat{\lambda}$ is unbiased and consistent for large N.
(as far as we can tell from the linearized g).

Note: The analysis method above is based on the premise that g is approximately linear for values of T where $p(T;\theta)$ is nonzero. This is reasonably accurate when

i) N is large so that $p(T;\theta)$ is concentrated around its mean,

ii) The data distribution is already tight (due to low noise, for instance).

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

Here $w[n]$ is zero mean white noise with var σ^2.
We need to estimate f_0. Assume that $\phi \sim \mathrm{Unif}[0, 2\pi)$.
Let $s[n] = A\cos(2\pi f_0 n + \phi)$. Then

$$E\{s[n]\} = E[A\cos(2\pi f_0 n + \phi)] = \int_0^{2\pi} A\cos(2\pi f_0 n + \phi)\frac{1}{2\pi}\,d\phi = 0.$$

$$r_{ss}[k] = E[s[n]s[n+k]] = E[A^2\cos(2\pi f_0 n + \phi)\cos(2\pi f_0(n+k)+\phi)]$$

$$= A^2 E\left[\tfrac{1}{2}\cos(4\pi f_0 n + 2\pi f_0 k + 2\phi) + \tfrac{1}{2}\cos(2\pi f_0 k)\right]$$

$$= \frac{A^2}{2}\cos(2\pi f_0 k)$$

Then $r_{xx}[k] = r_{ss}[k] + r_{ww}[k] = \frac{A^2}{2}\cos(2\pi f_0 k) + \sigma^2\delta[k]$.

If A is known (for simplicity, take $A = \sqrt{2}$),

$$r_{xx}[k] = \cos(2\pi f_0 k) + \sigma^2\delta[k]$$

$$\therefore\ r_{xx}[1] = \cos(2\pi f_0) \Rightarrow \hat{f}_0 = \frac{1}{2\pi}\arccos \hat{r}_{xx}[1].$$

For instance, $\hat{r}_{xx}[1] = \frac{1}{N-1}\sum_{n=0}^{N-2} x[n]x[n+1]$.

If the SNR is low (or N is small), we might get $|\hat{r}_{xx}[1]| > 1$, in which case $\hat{f}_0$ is meaningless.

With some effort, we can show that $E[\hat{f}_0] = f_0$

and $$\mathrm{var}(\hat{f}_0) = \frac{\sigma^2}{(2\pi)^2(N-1)^2\sin^2(2\pi f_0)}\left[s^2[1] + 4\cos^2(2\pi f_0)\sum_{n=1}^{N-2} s^2[n] + s^2[N-2]\right]$$

for large SNR. Here $\mathrm{var}(\hat{f}_0)$ decays as $\frac{1}{N^2}$, while CRLB as $\frac{1}{N^3}$.

Suggested Problems: 1, 2, 4, 5, 8, 9