2.6 Maximum Likelihood Estimation

- Recall two general estimation approaches
  - $\underline{\theta}$ is random, pdf $p_{\underline{\theta}}(\underline{\theta})$ is known. This produces Bayes parameter estimation, special case linear mmse estimation
  - $\underline{\theta}$ is non random but unknown, or random but no information on $p_{\underline{\theta}}(\underline{\theta})$. ML is a powerful method here.

- Recall MAP estimation:

  choose $\hat{\underline{\theta}}$ to maximize

  $$p(\hat{\theta}|x) = \frac{p(x|\hat{\theta})\,p(\hat{\theta})}{p(x)}, \text{ equivalently } \hat{\theta} = \arg\max_{\theta} p(\underline{x}|\underline{\theta})\,p(\underline{\theta})$$

2.6.1 The ML principle says — if you have no idea how $\underline{\theta}$ is distributed (if it has a distribution), then

choose $\hat{\underline{\theta}}$ to maximize $p(\underline{x}|\underline{\theta})$

Equivalent to making $p_{\underline{\theta}}(\theta)$ uniform — no information.

- ML estimators may or may not be biased.

- Example: Photon Counting. Estimate arrival rate λ from count k. If we have no idea how likely different values of λ are, we use $p_{k|\lambda}(k|\lambda) = e^{-\lambda T}\frac{(\lambda T)^k}{k!}$

$$\frac{\partial}{\partial\lambda} p_{k|\lambda}(k|\lambda) = \frac{T}{k!} e^{-\lambda T}(\lambda T)^{k-1}(-\lambda T + k) \Rightarrow \hat{\lambda} = \frac{k}{T} \quad \text{ML}$$

Compare with mmse when λ exponentially distributed

$$\hat{\lambda} = \frac{b}{bT+1}(k+1) \quad \text{MMSE}$$

- Call $l(\underline{\theta}) = p_{\underline{x}|\underline{\theta}}(\underline{x}|\underline{\theta})$ the likelihood function

Since logarithm is monotonic, it is equivalent to maximize

$$L(\underline{\theta}) = \ln(l(\theta))$$ the log-likelihood (sometimes $\Lambda(\underline{\theta})$)

The log likelihood is convenient when $p(\underline{x}|\underline{\theta})$ has an exponential form (e.g. Gaussian) or if components of $\underline{x}$ are conditionally independent

$$L(\underline{\theta}) = \ln\left[\prod_{i=1}^{M} p_{x_i|\theta}(x_i|\underline{\theta})\right] = \sum_{i=1}^{M} L_i(\underline{\theta})$$

- Since maxima are important, the vector of partial derivatives has a name

$$V(\underline{\theta}) = \frac{d}{d\underline{\theta}} L(\underline{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\theta_1} L(\underline{\theta}) \\ \vdots \\ \frac{\partial L(\underline{\theta})}{\partial \theta_K} \end{bmatrix}$$ the "score"

- Note that $l(\underline{\theta})$, $L(\underline{\theta})$, $V(\underline{\theta})$ all depend on $\underline{x}$, so all are random variables.

- ML estimators have some nice properties. If we make a sequence of observations $\underline{x}_i$, $i = 1, \ldots, N$ that are conditionally independent and identically distributed, we have asymptotic properties as $N \to \infty$:
  - ML estimators are asymptotically unbiased
  - ML estimators are asymptotically consistent
  - ML estimators are asymptotically efficient (unbiased and meet Cramér Rao lower bound)
  - ML estimators are asymptotically Gaussian

2.6.3 A very important special case leads to least squares.

Suppose we observe

$$\underline{x} = \underbrace{A\,\underline{\theta}}_{\underline{s}\ \text{signal}} + \underline{n}$$

where A is a known matrix $N \times M$
$\underline{\theta}$ is to be estimated M
$\underline{n}$ is white Gaussian noise. N
$E[\underline{n}\,\underline{n}^+] = \sigma_n^2 I$

$$\ell(\theta) = p_{x/\theta}(\underline{x}/\underline{\theta}) = \frac{1}{(2\pi)^N \left(\frac{\sigma_n^2}{2}\right)^N} \exp\left[-\frac{1}{\sigma_n^2}\left(\underline{x} - A\,\underline{\theta}\right)^+ \left(\underline{x} - A\,\underline{\theta}\right)\right]$$

$$L(\theta) \sim -\frac{1}{\sigma_n^2}\left|\underline{x} - A\,\underline{\theta}\right|^2 = -\frac{1}{\sigma_n^2}\sum_{i=1}^{N}\left|x_i - \underline{a}_i^+\,\underline{\theta}\right|^2$$

minimize sum of squared errors.

This is a linearly parameterized signal model, and we want the best fit to observations.

Solution: $\frac{d}{d\underline{\theta}}\left(\underline{x} - A\,\underline{\theta}\right)^+\left(\underline{x} - A\,\underline{\theta}\right) = -\left(x^+ - \underline{\theta}^+ A^+\right)A = \underline{0}$

or $A^+A\,\underline{\hat{\theta}} = A^+\underline{x}$, which yields

$$\underline{\hat{\theta}} = \left(A^+A\right)^{-1}A^+\underline{x}$$

"deterministic normal equations"

$$\underline{\hat{s}} = A\,\underline{\hat{\theta}} = A\left(A^+A\right)^{-1}A^+\underline{x}$$

(projection of $\underline{x}$ onto subspace spanned by A cols)

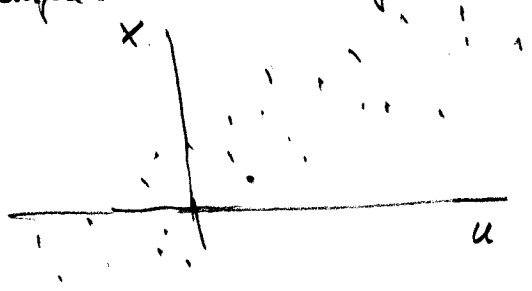
$$\hat{n} = \underline{x} - \hat{s} = \left(I - A\left(A^+A\right)^{-1}A^+\right)\underline{x}$$

(orthog complement — projn of $\underline{x}$ onto subspace orthog to A colums)

Deterministic normal equations:

- basis of signal space is columns of $A = [\underline{a}_1 | \underline{a}_2) \cdots \underline{a}_N]$
- Gram matrix of inner prods $[G]_{i,k} = \underline{a}_k^+\,\underline{a}_i$
  so $G = A^+A$
- Inner prod with observations $[\underline{p}]_k = \underline{a}_k^+\,\underline{x}$
  so $\underline{p} = A^+\underline{x}$
- Therefore $A^+A\,\underline{\hat{\theta}} = A^+\underline{x}$ is just $G\,\underline{\hat{\theta}} = \underline{p}$
  $\hat{\theta} = \left(A^+A\right)^{-1}A^+\underline{x}$ is just $\hat{\theta} = G^{-1}\underline{p}$

example: best straight line



x vs. u

model $x = \theta_1 + \theta_2 u$

$$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad A = \begin{bmatrix} 1 & u_1 \\ 1 & u_2 \\ \vdots & \vdots \\ 1 & u_N \end{bmatrix} \quad \underline{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \qquad \underline{x} = A\underline{\theta} + \underline{n}$$

example: channel sounding (like PSAM)

$$r(k) = g(k)c(k) + n(k)$$

assume g is constant but unknown (including σ_g^2) during pilots

$r(k) = g + n(k)$ what is ML est of g?

Since $n(k)$ is white, Gaussian, ML becomes least squares

$$\underline{x} = \begin{bmatrix} r(1) \\ \vdots \\ r(N) \end{bmatrix} \quad A = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \theta = g \qquad \underline{x} = A\theta + \underline{n}$$

or $\begin{bmatrix} c(1) \\ \vdots \\ c(N) \end{bmatrix}$

More generally, get impulse response

S N Crozier, D D Falconer & S A Mahmoud "Least Sum of Squared Errors Channel Estimation" Proc IEE Part F, vol 138 no 4 August 1991

R A Ziegler & J M Cioffi "Estimation of Time Varying Digital Radio Channels" IEEE Trans Veh Technol vol 41 no 2 pp 134-151, May 1992