APPENDIX F

SUFFICIENT STATISTICS

(from class notes, ENSC 810)

2. BASIC ESTIMATION THEORY

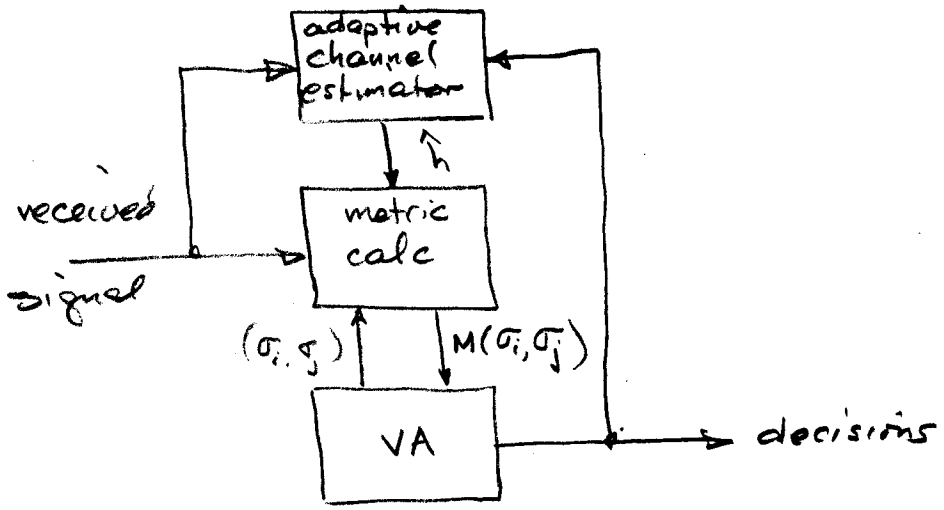
Read Haykin Appendix D

Other references:
- your undergrad text on statistics
- HL van Trees, Detection Estimation and Modulation Theory, Vol. 1, McGraw Hill, late 60's
- Louis L Scharf, Statistical Signal Processing Addison Wesley 91
- S J Orfanidis, Optimum Signal Processing 2nd ed, Macmillan 94

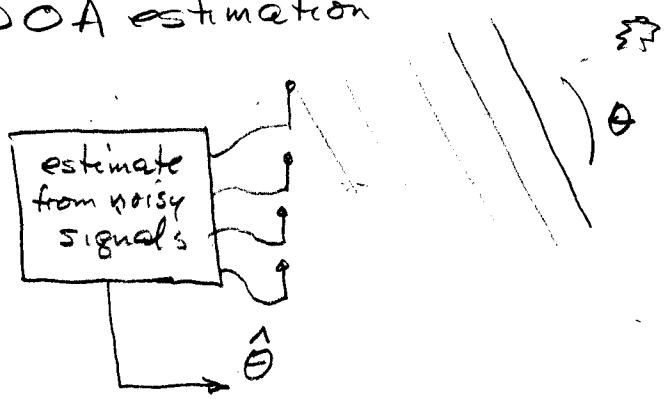
2.1 A Few Examples

- As we start the first significant topic in the course, ask what good it will do you...

- Adaptive Viterbi Equalizer



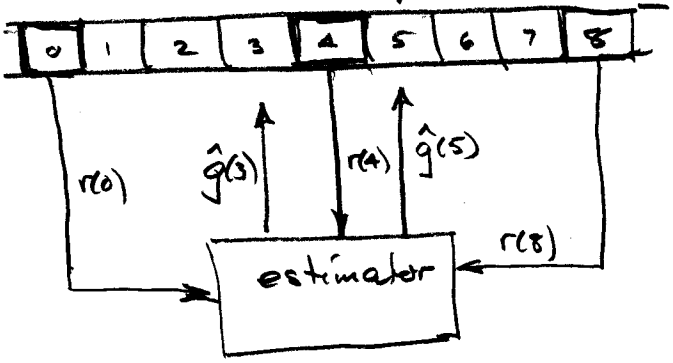
- DOA estimation



- pilot symbol assisted modulation

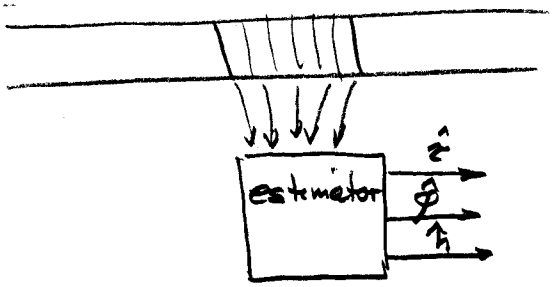
$r(k) = g(k)\,c(k) + n(k)$

fading and noise



some signals c(0), c(4), c(8) are known, so estimate $\hat{g}(3)$, $\hat{g}(5)$ etc from r(0), r(4), r(8) etc

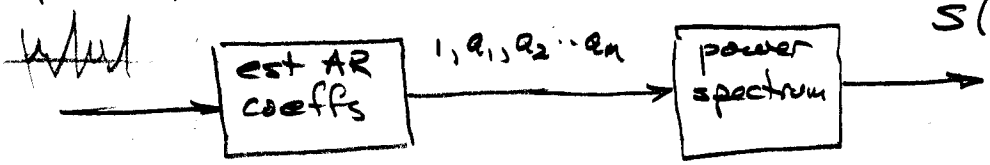
- burst channel estimation



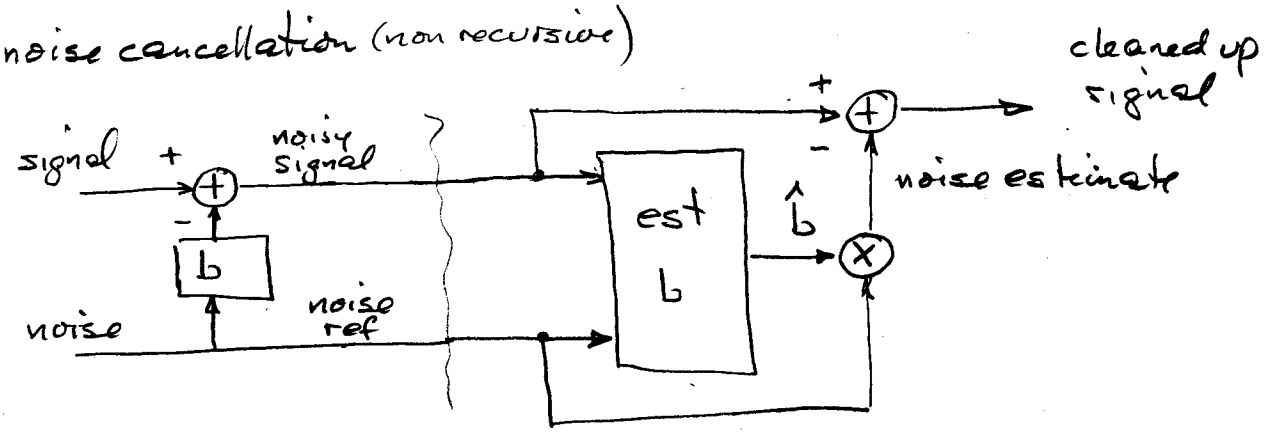
known training sequence

estimate delay, carrier phase, impulse response for use in demodulator

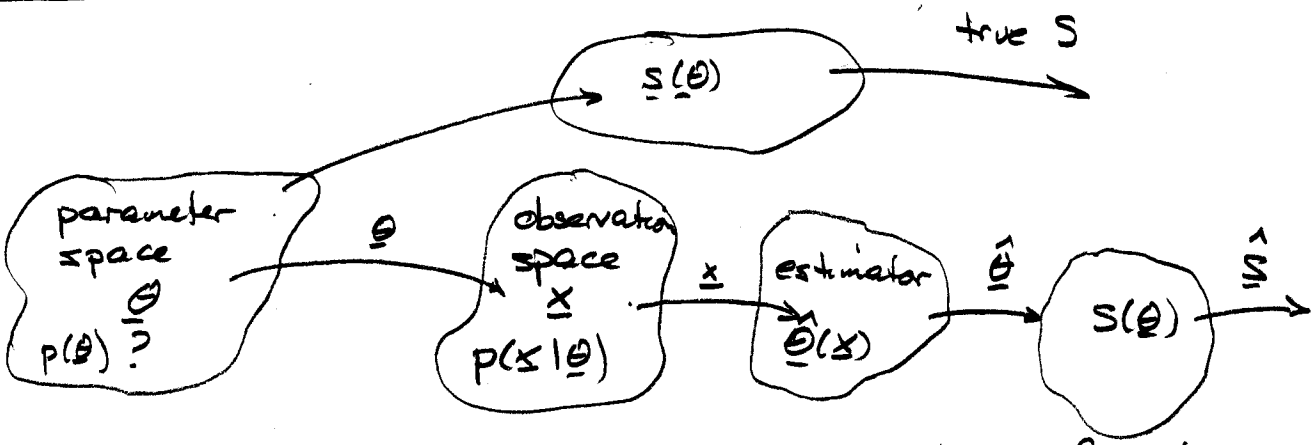
- speech power spectrum estimation



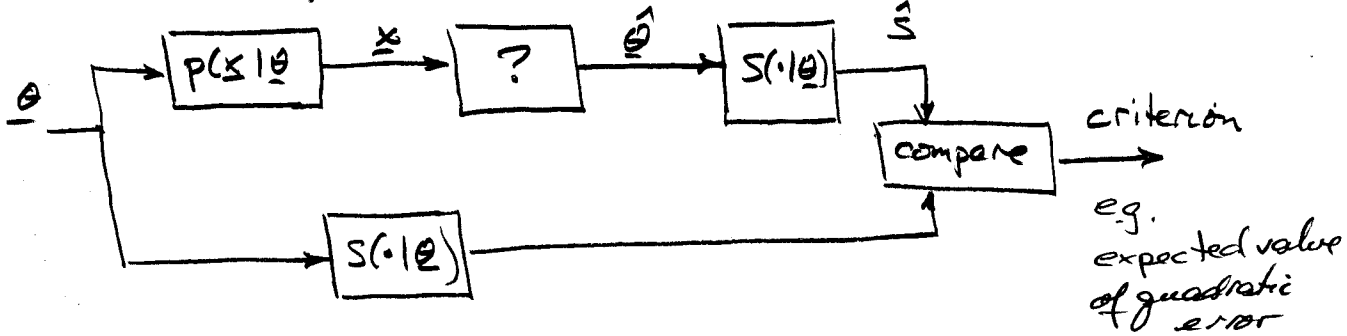
- noise cancellation (non recursive)



2.2 A Model of What We're Doing



We want to estimate a random or deterministic function $S(\cdot|\underline{\theta})$ of an unknown parameter $\underline{\theta}$. All we have is an observation $\underline{x}$ with pdf dependent on $\underline{\theta}$. Use $\underline{x}$ to obtain $\hat{\underline{\theta}}$, hence $\hat{\underline{S}}$, using a criterion of goodness



Examples (Schorf):

→ $S(\cdot|\underline{\theta}) = \underline{\theta}$ Here we estimate $\underline{\theta}$ itself

→ $S(\cdot|\theta) = S(e^{j\omega}|\underline{\theta})$ Estimate power spectral density

Haykin 3.5

- Two important cases
  - $\underline{\theta}$ is random, pdf $p(\underline{\theta})$ known
  - $\underline{\theta}$ is non random, but unknown

- Note this is parameter estimation, not interval estimation: "The poll is considered accurate to within 3% 19 times out of 20."

2.3 Properties of Estimators

There are certain key properties we look for in estimators. Here are some common ones

- Unbiased estimators

  An estimator $\hat{\underline{\theta}}(\underline{x})$ is unbiased if $E[\hat{\underline{\theta}}] = E[\underline{\theta}]$.

  example θ is mean μ, $\underline{x} = (x_1, \cdots, x_N)$ N measurements

  Estimate μ. Easy to show that

  $$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(sample mean)}$$

  is unbiased — just take expectation

  $$E[\hat{\mu}] = \frac{1}{N} N\mu = \mu$$

  example sample variance.

  $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ if we know true mean μ

  $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x - \hat{\mu})^2$ to remove bias if we use sample mean

  Since

  $$E\left[\sum_{i=1}^{N}(x_i - \hat{\mu})^2\right] = E\left[\sum_i \left(x_i - \frac{1}{N}\sum_j x_j\right)^2\right]$$

  $$= E\left[\sum_i x_i^2 + \frac{1}{N}\sum_i\sum_j x_i x_j - \frac{2}{N}\sum_i\sum_j x_i x_j\right]$$

  $$= E\left[\sum_i x_i^2 - \frac{1}{N}\sum_i\sum_j x_i x_j\right]$$

  $$= N\overline{x^2} - \frac{1}{N}\left(N\overline{x^2} + N(N-1)\mu^2\right) = (N-1)\overline{x^2} - (N-1)\mu^2$$

  $$= (N-1)\sigma^2 \qquad \text{so divide by } N-1 .$$

  Sometimes we accept a little bias to gain less scatter.

- Consistent estimators
  - unbiasedness is nice, but all it says is that the mean of the estimator is right — doesn't say the estimate is close.
  - estimator is consistent if prob $\hat{\underline{\theta}}$ takes a value farther away from $\underline{\theta}$ than some arbitrary c approaches 0 when $N \to \infty$ (N is # observations).
  - e.g. the sample mean is consistent, since sample variance $\to 0$, and use Chebyshev upper bound
  - we need two properties
    a) $\hat{\underline{\theta}}$ is unbiased
    b) $\text{var}(\hat{\underline{\theta}}) \to \underline{0}$ when $N \to \infty$ (or cov matrix $E\left[(\underline{\theta}-\hat{\underline{\theta}})(\underline{\theta}-\hat{\underline{\theta}})^T\right]$)

- Efficient estimators
  - if there is more than one unbiased consistent estimator available, the one with the smallest variance is said to be the most efficient of them
  - a lower bound on the variance of any unbiased consistent estimator is the Cramér Rao bound (more about it later). If it meets the C-R bound it is an efficient estimate

- Why the restriction to unbiased estimators?
  It's because there is often a tradeoff between bias and variance; e.g.

$$\underline{\hat{\Theta}}(\underline{x}) = \underline{0} \qquad \text{zero variance!}$$

$$E[\underline{\hat{\Theta}}] - \underline{\theta} = -\underline{\theta} \qquad \text{bias is } -\underline{\theta}$$

but it's not much use.

- Sufficient estimator
  - This is linked to sufficient statistics — and to data reduction.
  - example: When you estimate μ from a set of measurements, do you really need all of $x_1, \cdots x_N$ in detail, or is the sample mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ sufficient?
  - An estimator $\underline{\hat{\Theta}}$ is sufficient if it uses all the info relevant to estimation of $\underline{\theta}$ in the observations.

    An informal example is an easier way to understand than mathematics (that comes later).

    - Measurements $\underline{x}$ are useful for estimating θ only to the extent that θ influences the pdf of $\underline{x}$. If $p_{\underline{x}|\theta}(\underline{x}|\theta)$ is not a function of θ, then $\underline{x}$ gives no information about θ

    - Consider N Bernoulli trials, $x_n = \begin{cases} 1 & \text{prob } \theta \\ 0 & \text{prob } 1-\theta \end{cases}$

      In $\underline{x}$, get k ones.

- Question: Does the sample mean $\hat{\Theta} = k/N$ capture all the info about Θ in $\underline{x}$?

  Equivalent question: Given k, does the conditional pdf of $\underline{x}$ still depend on Θ?

- Example. Suppose $N = 4$ and you know there are k ones. Then the possible patterns are

  1100 1010 1001

  0110 0101 0011

  Nothing about Θ would make one of them more probable than another, so they are equiprobable. Consequently, knowing which $\underline{x}$ pattern it is tells nothing about Θ. Therefore, k alone is sufficient, not the detailed pattern.

Formally,

$\hat{\Theta}(\underline{x})$ is a sufficient estimator of $\underline{\Theta}$ if $p(\underline{x}/\underline{\Theta})$ does not depend on the true value $\underline{\Theta}$.

In effect, the value of $\hat{\Theta}$ summarizes all the information in $\underline{x}$ relevant to estimation of $\underline{\Theta}$

- Example: N Bernoulli trials, observe $x_n = \begin{cases} 1, & \text{prob } \Theta \\ 0, & \text{prob } 1-\Theta \end{cases}$

$$p_{\underline{x}|\Theta}(\underline{x}|\Theta) = \Theta^k (1-\Theta)^{N-k} \qquad k = \#\text{ of ones}$$

$$p_k(k) = \binom{N}{k} \Theta^k (1-\Theta)^{N-k} \qquad \text{prob of } k \text{ ones}$$

Then k is a sufficient statistic for Θ, and

$\hat{\Theta} = k/N$ is a sufficient estimator.

---

why?

$$p_{\underline{x}|\Theta}(\underline{x}|\hat{\Theta}=k) = \frac{p_{\underline{x},\hat{\Theta}}(\underline{x}, k)}{p_k(k)} = \frac{\Theta^k (1-\Theta)^{N-k}}{\binom{N}{k} \Theta^k (1-\Theta)^{N-k}} = \frac{1}{\binom{N}{k}}$$

a uniform pdf independent of Θ

- Example: binomial trials with same mean Θ

N_1 has x_1 ones ; N_2 has x_2 ones.

To estimate mean, we don't need x_1, x_2 individually — the sum will do:

$$\hat{\Theta} = \frac{x_1 + x_2}{N_1 + N_2}$$

$$p_{\underline{x}|\Theta}\left(\underline{x}|\hat{\Theta} = \tfrac{x_1+x_2}{N_1+N_2}\right) = \frac{p_{\underline{x},x_1+x_2}(\underline{x}, x_1+x_2)}{p_{\hat{\Theta}}\left((x_1+x_2)/(N_1+N_2)\right)}$$

$$= \frac{p_{\underline{x}}(x_1, x_2)}{p_{x_1+x_2}(x_1+x_2)} = \frac{\binom{N_1}{x_1}\Theta^{x_1}(1-\Theta)^{N_1-x_1}\binom{N_2}{x_2}\Theta^{x_2}(1-\Theta)^{N_2-x_2}}{\binom{N_1+N_2}{x_1+x_2}\Theta^{x_1+x_2}(1-\Theta)^{N_1+N_2-x_1-x_2}}$$

$$= \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}}{\binom{N_1+N_2}{x_1+x_2}}$$

Θ has no further influence