Welcome to ECE 531!

This course is a graduate-level introduction to detection and estimation theory, whose goal is to extract information from signals in noise. A solid background in probability and some knowledge of signal processing is needed.

Course Textbook:

Other useful references:
- H. Vincent Poor, Introduction to Signal Detection and Estimation
- Carl Helstrom, Elements of Signal Detection and Estimation. It’s out of print, so here’s my pdf copy.

Notes:
I will follow the course textbooks fairly closely, using a mixture of slides (highlighting the main points and with nice illustrations) and more in-depth blackboard derivations/proofs in class. I will post a pdf version of the slides as they become ready here, but the derivations will be given in class only.

Topics:
Estimation Theory:
- General Minimum Variance Unbiased Estimation, Ch.2, 5
- Cramer-Rao Lower Bound, Ch.3
- Linear Models+Unbiased Estimators, Ch.4, 6
- Maximum Likelihood Estimation, Ch.7
- Least squares estimation, Ch.8
- Bayesian Estimation, Ch.10-12

Detection Theory:
- Statistical Detection Theory, Ch.3
- Deterministic Signals, Ch.4
- Random Signals, Ch.5
- Statistical Detection Theory 2, Ch.6
- Non-parametric and robust detection

Grading:
Weekly homeworks (15%), Exam 1 = max(Exam1, Exam 2, Final) (20%), Exam 2 = max(Exam 2, Final) (20%), Project (15%), Final exam (30%).
Problems - attack!

Find the MVUE for estimating $A$ and $\sigma^2$ from the data

$$x[n] = A + w[n], \quad n = 0, 1, \cdots, N - 1, \quad w[n] \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2).$$


Problems - attack!

Suppose we want to determine which of two channels is being used for communications. We observe the output of the channel $x(n)$ for $n = 0, 1, \cdots, N - 1$. We know that

- If channel 1 is used: $x(n) = a_1 s_1(n) + w_1(n)$
- If channel 2 is used: $x(n) = a_2 s_2(n) + w_2(n)$

(a) Suppose $a_1, a_2, s_1(n), s_2(n)$ are all known and that $w_1 \sim \mathcal{N}(0, \sigma^2 I), w_2 \sim \mathcal{N}(0, \sigma^2 I)$. Describe the detector that minimizes the probability of error.

(b) Suppose $a_1, a_2, s_1(n), s_2(n)$ are all known and that $w_1 \sim \mathcal{N}(0, C_{W_1}), w_2 \sim \mathcal{N}(0, C_{W_2})$. Describe the detector that minimizes the probability of error (you do not need to simplify).

(c) Suppose $a_1, a_2$ are unknown, $s_1(n), s_2(n)$ are known and $w_1 \sim \mathcal{N}(0, C_{W_1}), w_2 \sim \mathcal{N}(0, C_{W_2})$. Determine the generalized likelihood ratio test (again, you do not need to simplify).
Problems - attack!

In the binary communication system shown in Fig. 5-1, messages \( m = 0 \) and \( m = 1 \) occur with a priori probabilities 1/4 and 3/4 respectively. Suppose that we observe \( r \),

\[
r = n + m,
\]

where \( n \) is a continuous valued random variable with the pdf shown in Fig. 5-2. The random variable \( n \) is statistically independent of whether message \( m = 0 \) or \( m = 1 \) occurs.

(a) Find the minimum probability of error detector, and compute the associated probability of error, \( \Pr(m \neq m) \).

(b) (Practice) Suppose that the receiver does not know the a priori probabilities, so it decides to use a maximum likelihood (ML) detector. Find the ML detector and the associated probability of error. Is the ML detector unique? Justify your answer. If your answer is no, find a different ML receiver and the associated probability of error.

Problems - attack!

Consider the estimation of a nonrandom but unknown parameter \( x \) from an observation of the form \( y = x + w \) where \( w \) is a random variable. Two different scenarios are considered in parts (a) and (b) below.

(a) Suppose \( w \) is a zero-mean Laplacian random variable. i.e.,

\[
p_w(w) = \frac{\alpha}{2} e^{-|w|}
\]

for some \( \alpha > 0 \). Does an unbiased estimate of \( x \) exist? Explain. Does an efficient estimate of \( x \) exist? If so, determine \( \hat{x}_{\text{eff}}(y) \). If not, explain.

\[
\text{Hint: } \int_{-\infty}^{\infty} w^2 e^{-|w|} \, dw = \frac{2}{\alpha^2}
\]

(b) Suppose \( w \) is a zero-mean random variable with probability density \( p_w(w) > 0 \) depicted in Figure 6-1. Does an efficient estimate of \( x \) exist? If so, determine \( \hat{x}_{\text{eff}}(y) \). If not, explain.
Problems-attack!

3. Suppose that $h, x, v$ are mutually independent random variables with

$$x \sim \mathcal{N}(3, 4), \quad v \sim \mathcal{N}(0, 4)$$

and where $h$ takes on only two possible values, 1 and 3 with

$$Pr(h = 1) = Pr(h = 3) = 1/2.$$ Let

$$z = hx + v$$

(a) Determine the MMSE estimator of $x$ given $z$ and $h$ (i.e. $\hat{x}_{MMSE}$).
(b) Determine the MMSE achieved by the above estimator.

Course outline


- General Minimum Variance Unbiased Estimation, Ch.2, 5
- Cramer-Rao Lower Bound, Ch.3
- Linear Models+Unbiased Estimators, Ch.4, 6
- Maximum Likelihood Estimation, Ch.7
- Least squares estimation, Ch.8
- Bayesian Estimation, Ch.10-12


- Statistical Detection Theory, Ch.3
- Deterministic Signals, Ch.4
- Random Signals, Ch.5
- Statistical Detection Theory 2, Ch.6
- Non-parametric and robust detection
Estimation: General Minimum Variance Unbiased Estimation

- Bias: (expected value of estimator - true value of data)
  \[ \text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = E[\hat{\theta} - \theta] \]

- MVUE:
  Unbiased estimator of minimum variance
  Always exist?

Estimation: Cramer-Rao lower bound

- Lower bound on variance of ANY unbiased estimator!

- Usage:
  - assert whether an estimator is MVUE
  - benchmark against which to measure the performance of an unbiased estimator
  - feasibility studies

- Depends on?

\[ \text{Noise} \]
\[ \text{Nature} \rightarrow \text{Transmission / measurement} \rightarrow \text{Processing} \]
The Cramer-Rao Lower Bound (CRLB)

Theorem: Cramer-Rao Lower Bound: Let \( p(x; \theta) \) satisfy the regularity condition

\[
E \left[ \frac{\partial \ln p(x; \theta)}{\partial \theta} \right] = 0
\]

Then the variance of any unbiased estimator \( \hat{\theta} \) must satisfy

\[
\text{var}(\hat{\theta}) \geq \frac{1}{-E \left[ \frac{\partial^2 \ln p(x; \theta)}{\partial \theta^2} \right]} = \frac{1}{E \left[ \left( \frac{\partial \ln p(x; \theta)}{\partial \theta} \right)^2 \right]}
\]

where the derivative is evaluated at the true value of \( \theta \) and expectation is wrt \( p(x; \theta) \). Furthermore an unbiased estimator may be found that attains the bound for all \( \theta \) if and only if

\[
\frac{\partial \ln p(x; \theta)}{\partial \theta} = I(\theta)(g(x) - \theta)
\]

for some \( g(\cdot) \) and \( I \). That estimator, which is the MVUE, is \( \hat{\theta} = g(x) \) and the minimum variance is \( 1/I(\theta) \).

CRLB examples

Consider estimating the parameter \( \Lambda \) from \( x[0] = \Lambda + w[0] \), where \( w[0] \sim \mathcal{N}(0, \sigma^2) \). What is the CRLB in estimating \( \Lambda \)?

Consider estimating the parameter \( \Lambda \) from the observations \( x[n] = \Lambda + w[n] \), where \( n = 0, 1, 2, \cdots N - 1 \) is White Gaussian Noise of variance \( \sigma^2 \), i.e. \( w[n] \sim \mathcal{N}(0, \sigma^2) \) and are independent. What is the CRLB in estimating \( \Lambda \) now?

Consider estimating the parameter \( \phi \) from the observations

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

in WGN of variance \( \sigma^2 \). The amplitude \( A \) and frequency \( f_0 \) are assumed to be known. What is the CRLB in estimating \( \phi \)?
CRLB T or F

- The CRLB always exists regardless of \( p(x; \theta) \)
- the CRLB applies to unbiased estimators only.
- Determining the CRLB requires statistics of all possible estimators \( \hat{\theta} \).
- The CRLB depends on the observations \( x \).
- The CRLB depends on the parameter to be estimated, \( \theta \).
- The CRLB tells you whether or not a MVUE exists.

all the way to....

The co-variance matrix \( C_\theta \) of any unbiased estimator \( \hat{\theta} \) of \( g(\theta) \), an \( r \)-dimensional function, satisfies:

\[
C_\theta - \frac{\partial g(\theta)}{\partial \theta} \left( I^{-1}(\theta) \frac{\partial g(\theta)}{\partial \theta} \right)^T \succeq 0
\]

Here \( \frac{\partial g(\theta)}{\partial \theta} \) is the \( r \times p \) Jacobian matrix defined by

\[
\left[ \frac{\partial g(\theta)}{\partial \theta} \right]_{ij} = \frac{\partial g_i(\theta)}{\partial \theta_j}
\]
CRLB for General Gaussian Case

- When observations are Gaussian and one knows the dependence of the mean and covariance matrix on the unknown parameters, we know the closed form of the CRLB (or Fisher information matrix):

Given observations
\[ \mathbf{x} \sim \mathcal{N}(\mu(\theta), \mathbf{C}(\theta)), \]
then the Fisher information matrix \( \mathbf{I}(\theta) \) is given by:

\[
[I(\theta)]_{ij} = \left[ \frac{\partial \mu(\theta)}{\partial \theta_i} \right]^T \mathbf{C}^{-1}(\theta) \left[ \frac{\partial \mu(\theta)}{\partial \theta_j} \right] + \frac{1}{2} \text{tr} \left[ \mathbf{C}^{-1}(\theta) \frac{\partial \mathbf{C}^{-1}(\theta)}{\partial \theta_i} \mathbf{C}^{-1}(\theta) \frac{\partial \mathbf{C}^{-1}(\theta)}{\partial \theta_j} \right]
\]

Finding MVUE so far

When the observations and the data are related in a linear way \( \mathbf{x} = \mathbf{H}\theta + \mathbf{w} \) and the noise was Gaussian, then the MVUE was easy to find:

\[ \hat{\theta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \]

with covariance matrix
\[ \mathbf{C}_\hat{\theta} = \sigma^2 (\mathbf{H}^T\mathbf{H})^{-1}. \]

Using the CRLB, if you got lucky then you could write

\[ \frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = \mathbf{I}(\theta) (g(\mathbf{x}) - \theta), \]

where \( \hat{\theta} = g(\mathbf{x}) \) would be your MVUE, which would also be an efficient estimator (meeting the CRLB).

Even if no efficient estimator exists, a MVUE may exist. In this section we try to find a systematic way of determining the MVUE if it exists.
Finding the MVUE

We wish to estimate the parameter(s) $\theta$ from the observations $x$.

1. Determine (if possible) a sufficient statistic $T(x)$ for the parameter to be estimated $\theta$. This may be done using the Neyman-Fisher factorization theorem.

2. Determine whether the sufficient statistic is also complete. This is generally hard to do. If it is not complete, we can saying nothing more about the MVUE. If it is, continue to step 3.

3. Find the MVUE $\hat{\theta}$ from $T(x)$ is one of two ways using the Rao-Blackwell-Lehmann-Scheffe (RBLs) theorem:
   - Find a function $g(\cdot)$ of the sufficient statistic that yields an unbiased estimator $\hat{\theta} = g(T(x))$, the MVUE! By definition of completeness of the statistic, this will yield the MVUE.
   - Find any unbiased estimator $\bar{\theta}$ for $\theta$, and then determine $\hat{\theta} = E[\bar{\theta}|T(x)]$. This is usually very tedious/difficult to do. The expectation is taken over the distribution $p(\bar{\theta}|T(x))$.

Estimation: linear models

- What’s a linear model and why is it useful?

\[ x = H\theta + w, \]

where

- $\theta =$ vector parameter to be estimated
- $x =$ received signal from which to estimate $\theta$
- $H =$ (known) observation matrix
- $w =$ noise of statistical characterization $\mathcal{N}(0, \sigma^2 I)$

- What can be said?

- Best Linear Unbiased Estimators (BLUE)
MVUE for the linear model

If the data follows the linear model specified above (notice the condition on the rank of the matrix $H$!) then the MVUE is given by

$$\hat{\theta} = (H^T H)^{-1} H^T x$$

and the covariance matrix of $\hat{\theta}$ is

$$C_{\hat{\theta}} = \sigma^2 (H^T H)^{-1}.$$ 

In fact, we have a full statistical description of the estimator $\hat{\theta}$ as we know its pdf, which is

$$\hat{\theta} \sim N(\theta, \sigma^2 (H^T H)^{-1}).$$

Examples

Consider $x[n] = A + Bn + w[n]$, for $n = 0, 1, \cdots, N - 1$ where $w[n]$ is WGN and $A, B$ are to be estimated. Formulate the linear model and determine the MVUE and its statistics when estimating $A$ and $B$. What does the rank condition on $H$ mean?

Consider $x[n] = A + r^n + w[n]$ for $n = 0, 1, \cdots, N - 1$ where $w[n]$ is WGN, $r$ is known, and $A$ is to be estimated. Find the MVUE and its statistics when estimating $A$. 

Examples

Consider the following polynomial curve fitting problem, where we wish to find \( \theta_1, \theta_2, \ldots, \theta_p \) so as to best fit the experimental data points \( (t_n, x(t_n)) \) for \( n = 0, 1, \ldots, N - 1 \) by the polynomial curve

\[
x(t_n) = \theta_1 t_n^1 + \theta_2 t_n^2 + \cdots + \theta_p t_n^{p-1} + w(t_n)
\]

for \( w(t_n) \) are WGN samples. Find the MVUE and its statistics when estimating \( \theta_1, \theta_2, \ldots, \theta_p \).

---

**General linear model**

\[
x = \mathbf{H} \theta + \mathbf{w}:
\]

- \( \mathbf{x} : N \times 1 \) observation vector
- \( \mathbf{H} : N \times p \) known observation matrix with \( N > p \) and rank \( p \)
- \( \theta : p \times 1 \) vector of unknown parameters to be estimated
- \( \mathbf{w} : N \times 1 \) noise vector, \( \sim \mathcal{N}(0, \mathbf{C}_w) \)

Re-do the derivation

**Key idea:** *whiten* the observations, then apply previous results
BLUE - vector

- Gauss-Markov theorem for BLUES:
  If the data are of the general linear model form

\[ x = H\theta + w \]

where \( H \) is a known \( N \times p \) matrix, \( \theta \) is a \( p \times 1 \) vector of parameters to be estimated and \( w \) is a \( N \times 1 \) noise vector with zero mean and covariance \( C \) (the PDF of \( w \) is otherwise arbitrary), then the BLUE of \( \theta \) is

\[ \hat{\theta} = (H^T C^{-1} H)^{-1} H^T C^{-1} x \]

and the minimum variance of \( \theta_i \) is

\[ \text{var}(\theta_i) = [(H^T C^{-1} H)^{-1}]_{ii}. \]

In addition, the covariance matrix of \( \hat{\theta} \) is

\[ C_{\hat{\theta}} = (H^T C^{-1} H)^{-1}. \]

---

Estimation: Maximum Likelihood Estimation

- Alternative to MVUE which is hard to find in general

- Easy to compute - very widely used and practical

- What is the MLE?

If \( \theta \) is the parameter to be estimated and \( x \) is the observation, then the MLE estimator \( \hat{\theta}_{MLE} \) is:

\[ \hat{\theta}_{MLE} = \arg \max p(x; \theta) \text{ for fixed (given) } x \]

- Properties?
Properties of the MLE

- the MLE is asymptotically unbiased (i.e. consistent: converges in probability to the true parameter value)
- the MLE is asymptotically efficient (meets the CRLB)
- the MLE is asymptotically distributed as $\mathcal{N}(\theta, I^{-1}(\theta))$, where $I(\theta)$ is the Fisher information.
- if an efficient estimator exists, the MLE finds it

The above statements hold under the regularity conditions:

- The existence of $\frac{\partial \ln p(x; \theta)}{\partial \theta}$ and $\frac{\partial^2 \ln p(x; \theta)}{\partial \theta^2}$
- $E \left[ \frac{\partial \ln p(x; \theta)}{\partial \theta} \right] = 0$

Invariance property

The MLE also has what's known as an invariance property, i.e. the MLE of the parameter $\alpha = g(\theta)$ where pdf $p(x; \theta)$ parameterized by $\theta$ is

$$\hat{\alpha} = g(\hat{\theta}),$$

where $\hat{\theta}$ is the MLE of $\theta$. The MLE of $\hat{\theta}$ is obtained by maximizing $p(x; \theta)$. If $g(\cdot)$ is not a 1:1 function, then $\hat{\alpha}$ maximizes the modified likelihood function

$$p_T(x; \theta) = \max_{\theta: \alpha = g(\theta)} p(x; \theta)$$
Examples

Find the ML estimate of $\theta$ given a single observation $x$ if $p(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, for $x, \theta \geq 0$. Is it unbiased? Is it efficient?

Consider the data

$$x[n] = A + w[n], \quad n = 0, 1, \cdots, N - 1,$$

where $w[n]$ is WGN with variance $\sigma^2$ and the vector parameter $\theta = [A \sigma^2]^T$ is to be estimated. Find the MLE and its asymptotic statistics.

MLE for linear model

When we have a linear model, meaning $x = H\theta + w$, where $H$ is a known $N \times p$ matrix with $N > p$ and rank $p$, and $w \sim \mathcal{N}(0, C)$, then the MLE of $\theta$ is

$$\hat{\theta} = (H^T C^{-1} H)^{-1} H^T C^{-1} x, \quad \hat{\theta} \sim \mathcal{N}(0, (H^T C^{-1} H)^{-1})$$

Thus, in the linear model, the MLE is efficient and MVUE!
Numerical methods

Finding the MLE boils down to finding the maximum of the likelihood function. This may be possible analytically; if not, numerical methods may be used. Numerically, one can find the maximum through a grid search over the possible parameter space: this may be good enough for small or finite parameter spaces, but becomes problematic once the parameter space extends to infinity. In these situations, iterative methods for finding the maximum are more useful.

We will see three iterative methods, where we start off with an initial guess of the MLE $\theta_0$, and at iteration $k \geq 0$ update the estimate as follows:

1. Newton-Raphson method: update $\theta_{k+1} = \theta_k - \left( \frac{\partial^2 \ln p(x; \theta)}{\partial \theta^2} \right)^{-1} \left. \frac{\partial \ln p(x; \theta)}{\partial \theta} \right|_{\theta=\theta_k}$

2. Scoring method: update $\theta_{k+1} = \theta_k + I^{-1}(\theta) \left. \frac{\partial \ln p(x; \theta)}{\partial \theta} \right|_{\theta=\theta_k}$

3. Expectation-Maximization (EM)

Estimation: Least Squares

• Alternative estimator with no general optimality properties, but nice and intuitive and no probabilistic assumptions on data are made - only need a signal model

The least squares estimator $\hat{\theta}_{LS}$ is equal to the value of $\theta$ that minimizes

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

where $s[n, \theta]$ is the sent signal (or nature) if $\theta$ for given parameter $\theta$.

• Advantages?

• Disadvantages?
Examples

Assume the signal model is $s[n] = A \cos(2\pi f_0 n)$, $\forall n = 0, 1, 2, \ldots, N - 1$. Find the LSE of $A$ if $f_0$ is known, then do the reverse.

Say we are given data $(x, y) : (1, 6), (2, 5), (3, 7), (4, 10)$ : and we want to find the line

$y = \beta_1 + \beta_2 x$

that best fits the data in the least squares sense. Find $\beta_1$ and $\beta_2$.

Example: Find the LSE of $\{A_1, A_2, A_3, r\}$ if $0 < r < 1$ in the signal model

$s[n] = A_1 r^n + A_2 r^{2n} + A_3 r^{3n}$

---

Linear LSE: theory

The signal model for linear LSE is

$s(\theta) = H\theta$

where $\theta$ is a $p \times 1$ vector of unknown parameters, $H$ is an $N \times p$ known matrix with $N > p$ and rank $p$.

**Linear model without any noise assumptions!**

In this case, determining the linear LSE boils down to solving the following $p$ normal equations:

$H^T H \theta = H^T x$

We can show the following about the LSE:

$\hat{\theta} = (H^T H)^{-1} H^T x$

$J_{min} = x^T (I - H(H^T H)^{-1} H^T) x$
Linear LSE vs other estimators

**Model**

- \( x = H_0 + e \)  
  - No Probability Model Needed
- \( x = H_0 + w \)  
  - PDF Unknown, White
- \( x = H_0 + w \)  
  - PDF Gaussian, White

**Estimate**

- \( \hat{\theta}_{LS} = \left( H^T H \right)^{-1} H^T x \)
- \( \hat{\theta}_{BLUE} = \left( H^T H \right)^{-1} H^T x \)
- \( \hat{\theta}_{ML} = \left( H^T H \right)^{-1} H^T x \)
- \( \hat{\theta}_{MVU} = \left( H^T H \right)^{-1} H^T x \)

If you assume Gaussian & apply these... BUT you are WRONG... you at least get the LSE!

Estimation: Bayesian Estimation

- Parameter to be estimated is assumed to be random, according to some prior distribution which models our knowledge of it

- Bayesian Minimum Mean Squared Error (MMSE):

Select the estimator \( \hat{A} \) to minimize \( BMSE(\hat{A}) = \int \int (A - \hat{A})^2 p(x, A) dx dA \)

Obtain the famous mean of the posterior pdf, i.e.

\( \hat{A} = E[A|x] \)

- Applications to Gaussian noise / linear model

http://www.ws.binghamton.edu/fowler/fowler%20personal%20page/EE522.htm
Bayesian estimation on Gaussian priors/noise

We now state (proofs in Appendix A of Ch.10) some useful theorems on jointly Gaussian distributions.

(Theorem 10.1) If \( x, y \) are jointly Gaussian with mean vector \( [E[x] \ E[y]]^T \) and covariance matrix

\[
C = \begin{bmatrix}
    \text{var}(x) & \text{cov}(x,y) \\
    \text{cov}(y,x) & \text{var}(y)
\end{bmatrix}
\]

then the conditional pdf \( p(y|x) \) is also Gaussian with mean and variance

\[
E(y|x) = E(y) + \frac{\text{cov}(x,y)}{\text{var}(x)}(x - E(x))
\]

\[
\text{var}(y|x) = \text{var}(y) - \frac{\text{cov}(x,y)^2}{\text{var}(x)}.
\]

This is crucial to know!

Bayesian estimation on Gaussian priors/noise

(Theorem 10.2) If \( x \) and \( y \) are jointly Gaussian where \( x \) is \( k \times 1 \) and \( y \) is \( l \times 1 \), with mean vector \( [E(x)^T \ E(y)^T]^T \) and partitioned covariance matrix

\[
C = \begin{bmatrix}
    C_{xx} & C_{xy} \\
    C_{yx} & C_{yy}
\end{bmatrix} = \begin{bmatrix}
    k \times k & k \times l \\
    l \times k & l \times l
\end{bmatrix},
\]

then the conditional pdf \( p(y|x) \) is also Gaussian and

\[
E(y|x) = E(y) + C_{yx}C_{xx}^{-1}(x - E(x))
\]

\[
C_{y|x} = C_{yy} - C_{yx}C_{xx}^{-1}C_{xy}
\]

This is crucial to know!
Bayesian linear model

(Theorem 10.3) Under the Bayesian linear model described above $\theta$ and $x$ are jointly Gaussian with a posterior pdf $p(\theta|x)$ that is also Gaussian with mean and covariance:

$$E[\theta|x] = \mu_\theta + C_\theta H^T (HC_\theta H^T + C_w)^{-1} (x - H \mu_\theta)$$

$$C_{\theta|x} = C_\theta - C_\theta H^T (HC_\theta H^T + C_w)^{-1} HC_\theta$$

There are alternative forms of this mean and variance which may be more useful, depending on the application, given by:

$$E[\theta|x] = \mu_\theta + (C^{-1}_\theta + H^T C^{-1}_w H)^{-1} H^T C^{-1}_w (x - H \mu_\theta)$$

$$C_{\theta|x} = (C^{-1}_\theta + H^T C^{-1}_w H)^{-1}$$

Examples: MMSE in Gaussian noise

Find the MMSE estimator of $A$ when given $x = A + w$, $A \sim N(\mu_A, \sigma^2_A)$, $w \sim N(0, \sigma^2)$

$$\hat{A}_{MMSE} = \mu_A + \frac{\sigma^2_A}{\sigma^2_A + \sigma^2} (x - \mu_A)$$
Examples: MMSE in Gaussian noise

Find the MMSE estimator of $A$ when given $x(n) = A + w(n)$, for $n = 0, 1, \cdots N - 1$, $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, $w \sim \mathcal{N}(0, \sigma^2 I)$

$$\hat{A}_{MMSE} = \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N} (\bar{x} - \mu_a) = \alpha \bar{x} + (1 - \alpha) \mu_A, \text{ where } \alpha = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/N}.$$

Examples: MMSE in Gaussian noise

Find the MMSE estimator of $\theta$ when given $x = H\theta + w$, $\theta \sim \mathcal{N}(\mu_\theta, C_\theta)$, $w \sim \mathcal{N}(0, C_w)$

$$E[\theta|x] = \mu_\theta + C_\theta H^T (HC_\theta H^T + C_w)^{-1} (x - H\mu_\theta)$$

$$C_{\theta|x} = C_\theta - C_\theta H^T (HC_\theta H^T + C_w)^{-1} HC_\theta$$
Properties of MMSE

- The MMSE estimator for the Bayesian model becomes the MVUE estimator for the classical linear model as the prior distribution becomes uninformative.
- The MMSE commutes over affine transformations.
- When $x = [x_1, x_2]^T$ and $\theta$ are jointly Gaussian, the MMSE is additive for independent data sets $x_1, x_2$, i.e.
  \[
  \hat{\theta}_{MMSE} = E[\theta] + C_{\theta x_1} C_{x_1 x_1}^{-1} (x_1 - E[x_1]) + C_{\theta x_2} C_{x_2 x_2}^{-1} (x - E[x_2])
  \]

Estimation: Bayesian Estimation

- General risk functions - arbitrary “cost” functions
  \[
  R = \int \int C(\theta - \hat{\theta}) p(\theta) \, dx \, d\theta
  \]
- Maximum a posteriori (MAP) estimation
  \[
  \hat{\theta} = \arg \max_{\theta} p(\theta|x)
  \]
- Linear MMSE: constrain estimator to be linear - very practical
  \[
  \hat{\theta} = \sum_{n=0}^{N} a_n x[n] + a_N
  \]
  where we choose the weighting coefficients $a_n$ to minimize the Bayesian MSE
  \[
  BMSE(\hat{\theta}) = E[(\theta - \hat{\theta})^2]
  \]
MAP properties

Some interesting properties of the MAP estimator:

- As $N \to \infty$ the MAP becomes the Bayesian ML estimator.
- If $x$ and $\theta$ are jointly Gaussian then the MAP = MMSE estimator.
- The MAP commutes over invertible linear transformations, but in contrast to ML estimators, does not commute over non-linear functions in general.

Sequential LMMSE

Let’s consider the estimation of $A$ from data $x(n) = A + w(n)$, where $w(n)$ is white Gaussian noise of variance $\sigma^2$ and $A$ is uniform on $[-A_0, A_0]$. Say we have $N$ data points $x(0), x(1), \cdots x(N-1)$ from which we form an LMMSE estimate of the parameter at time $N-1$,

$$\hat{A}(N-1) = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma^2}{N}} \overline{x}.$$ 

Now we obtain a new data point $x(N)$ and wish to update our estimate of $A$ accordingly. For each new data point thereafter we wish to update the estimator; this procedure is called sequential LMMSE. We consider only the simple example of estimating the DC level in noise, a general form of the sequential LMMSE for the Bayesian linear model form is given in the textbook pg. 397-399.

Algebraic approach  Geometric approach
LMMSE properties

- We only need the first and second order moment of the parameter and the data: \( E(\theta), E(x) \), first order statistics and \( C_{\theta \theta}, C_{x \theta}, C_{xx} \), the covariance matrices of the parameters and data.

- The LMMSE yields the same form of the estimator as the Gaussian MMSE estimator except that we do not assume the noise or prior parameter pdf is Gaussian.

- The LMMSE is sub-optimal except when \( E[\theta|x] \) happens to be linear (which is the case for jointly Gaussian \( x, \theta \)). This means that under the jointly Gaussian assumption, LMMSE = MMSE.

- LMMSE, like the MMSE estimator, commutes over affine transformations.

- If \( x(0) \) and \( x(1) \) are orthogonal (uncorrelated) observations, then \( \hat{\theta}_{LMMSE} = \hat{\theta}_{LMMSE}(x(0)) + \hat{\theta}_{LMMSE}(x(1)) \).

Course outline


- General Minimum Variance Unbiased Estimation, Ch.2, 5
- Cramer-Rao Lower Bound, Ch.3
- Linear Models+Unbiased Estimators, Ch.4, 6
- Maximum Likelihood Estimation, Ch.7
- Least squares estimation, Ch.8
- Bayesian Estimation, Ch.10-12


- Statistical Detection Theory, Ch.3
- Deterministic Signals, Ch.4
- Random Signals, Ch.5
- Statistical Detection Theory 2, Ch.6
- Non-parametric and robust detection
Problems - attack!

. The random variables $X$ and $Y$ are uniformly distributed over the trapezoidal region shown in the figure below, described by the equations

$$p(x, y) = \begin{cases} 1 & |y| < 1/2, \; y - 1/2 < x < y + 1/2 \\ 0 & \text{else} \end{cases}$$

(a) (5 points) Find the MMSE of $X$ given $Y$ and the corresponding minimum mean squared error.
(b) (3 points) Find the Linear MMSE of $X$.
(c) (2 points) Define the Quadratic MMSE as the estimator $\hat{X} = aY^2 + bY + c$ where the constants $a, b, c$ are chosen to minimize the mean squared error. Find the quadratic MMSE estimator of $X$ based on $Y$ and the corresponding MMSE. HINT: think first!
(d) (5 points) Find the MAP estimate of $X$ based on $Y$.

Problems - attack

- T/F: The performance of a detector in colored noise is always worse than that of a detector in white noise.
- T/F: The Wiener filter is a Bayesian filter than optimizes the MMSE criterion.
- T/F: The Cramer-Rao bound seen in class only applies to unbiased estimators.
- Explain the difference between MAP and ML estimators.
- Explain the difference between MAP and ML detectors.
- T/F: The sample mean is the MVUE for the DV level of any signal in noise.
Course outline


- General Minimum Variance Unbiased Estimation, Ch.2, 5
- Cramer-Rao Lower Bound, Ch.3
- Linear Models+Unbiased Estimators, Ch.4, 6
- Maximum Likelihood Estimation, Ch.7
- Least squares estimation, Ch.8
- Bayesian Estimation, Ch.10-12


- Statistical Detection Theory, Ch.3
- Deterministic Signals, Ch.4
- Random Signals, Ch.5
- Statistical Detection Theory 2, Ch.6
- Non-parametric and robust detection

Detection: Statistical Detection Theory

- Binary hypothesis testing
  \[ H_0 \text{ or } H_1? \]

- \( P(H_0; H_0) = \text{prob}(\text{decide } H_0 \text{ when } H_0 \text{ is true}) = \text{prob of correct non-detection} \)
- \( P(H_0; H_1) = \text{prob}(\text{decide } H_0 \text{ when } H_1 \text{ is true}) = \text{prob of missed detection} \) \( := P_M \)
- \( P(H_1; H_0) = \text{prob}(\text{decide } H_1 \text{ when } H_0 \text{ is true}) = \text{prob of false alarm} \) \( := P_{FA} \)
- \( P(H_1; H_1) = \text{prob}(\text{decide } H_1 \text{ when } H_1 \text{ is true}) = \text{prob of detection} \) \( := P_D \)
Detection: Statistical Detection Theory

Neyman-Pearson (NP): maximize $P_D$ subject to a desired fixed $P_{FA}$.

Receiver Operating Characteristics (ROC) curves

Generalized Bayesian risk which includes as special cases

- Minimum probability of error (min $P_E$) or maximum a posteriori (MAP): $C_{ii} = 0, C_{ij} = 1$ for $i \neq j$.
- Maximum likelihood (ML): $C_{ij} = 0, C_{ij} = 1$ for $i \neq j$ AND all priors are equal, i.e. $P(\mathcal{H}_i) = P(\mathcal{H}_j), \forall i, j$.

Neyman-Pearson hypothesis testing

Neyman-Pearson Theorem 3.1 (pp.65)

To maximize $P_D$ for a given $P_{FA} = \alpha$, decide $\mathcal{H}_1$ if

$$L(x) := \frac{p(x; \mathcal{H}_1)}{p(x; \mathcal{H}_0)} > \gamma,$$

where the threshold $\gamma$ is found from

$$P_{FA} = \int_{\{x: L(x) > \gamma\}} p(x; \mathcal{H}_0) dx = \alpha$$

$L(x)$ is the likelihood ratio, and comparing $L(x)$ to a threshold is termed the likelihood ratio test.
Useful problem 2.1

If \( T \sim \mathcal{N}(\mu, \sigma^2) \), then

\[
\Pr\{T > \gamma\} = Q\left( \frac{\gamma - \mu}{\sigma} \right)
\]

Bayesian risk

The Bayesian risk detection framework encompasses:

- Minimum probability of error (min \( P_E \)) or maximum a posteriori (MAP) (same): \( C_{ii} = 0, C_{ij} = 1 \) for \( i \neq j \). These detectors decide \( \mathcal{H}_1 \) if
  
  \[
  \min P_E : \frac{p(x|\mathcal{H}_1)}{p(x|\mathcal{H}_0)} > \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} = \gamma
  \]

- MAP: \( P(\mathcal{H}_1|x) > P(\mathcal{H}_0|x) \)

- Maximum likelihood (ML): \( C_{ij} = 0, C_{ij} = 1 \) for \( i \neq j \) AND all priors are equal, i.e. \( P(\mathcal{H}_i) = P(\mathcal{H}_j) \), \( \forall i, j \). This detector decides \( \mathcal{H}_1 \) if
  
  - ML: \( P(x|\mathcal{H}_1) > P(x|\mathcal{H}_0) \)
Multiple hypothesis testing

Once again, the Bayes risk is a generalization of MAP and ML detectors, which in the multiple hypothesis case reduce to:

- Minimum probability of error (min $P_E$) or maximum a posteriori (MAP) (same): $C_{ii} = 0, C_{ij} = 1$ for $i \neq j$. These detectors decide $\mathcal{H}_i$ if $P(\mathcal{H}_i|x) > P(\mathcal{H}_j|x), \forall j$

- Maximum likelihood (ML): $C_{ij} = 0, C_{ij} = 1$ for $i \neq j$ AND all priors are equal, i.e. $P(\mathcal{H}_i) = P(\mathcal{H}_j), \forall i, j$. This detector decides $\mathcal{H}_1$ if $P(x|\mathcal{H}_i) > P(x|\mathcal{H}_j), \forall j$

Detection: Deterministic Signals

- How to detect known signals in noise?

  $\mathcal{H}_0 : x[n] = w[n]$
  $\mathcal{H}_1 : x[n] = s[n] + w[n],$

- The famous matched filter!

- Generalized matched filter
- $> 2$ hypotheses
Performance of matched filter

\[ \mathcal{H}_0 : x[n] = w[n] \]
\[ \mathcal{H}_1 : x[n] = s[n] + w[n], \]

where \( w[n] \) is for now assumed to be white with variance \( \sigma^2 \).

The test statistic \( T(x) = x^T s \) has pdf \( \mathcal{N}(0, \sigma^2 \mathcal{E}) \) under \( \mathcal{H}_0 \) and pdf \( \mathcal{N}(\mathcal{E}, \sigma^2 \mathcal{E}) \) under \( \mathcal{H}_1 \). We can obtain \( P_{FA} \) and \( P_Q \) as follows:

\[
P_{FA} = Pr\{T > \gamma' ; \mathcal{H}_0\} \Rightarrow \gamma' = \sqrt{\sigma^2 \mathcal{E}} Q^{-1}(P_{FA})
\]

\[
P_D = Pr\{T > \gamma' ; \mathcal{H}_1\} = \frac{Q\left(\frac{\gamma' - \mathcal{E}}{\sqrt{\sigma^2 \mathcal{E}}}\right)}{Q\left(\frac{Q^{-1}(P_{FA}) - \sqrt{\mathcal{E}}}{\sigma^2}\right)}
\] 

Performance of generalized matched filters

\[ \mathcal{H}_0 : x[n] = w[n] \]
\[ \mathcal{H}_1 : x[n] = s[n] + w[n], \]

where \( w[n] \) is for now assumed to be correlated Gaussian noise with covariance matrix \( C \).

Performance of generalized matched filter: the test statistic \( T(x) = X^T C^{-1} s \) has pdf \( \mathcal{N}(0I, s^T C^{-1} s) \) under \( \mathcal{H}_0 \) and pdf \( \mathcal{N}(s^T C^{-1} s, s^T C^{-1} s) \) under \( \mathcal{H}_1 \). We can obtain \( P_{FA} \) and \( P_Q \) as follows:

\[
P_{FA} = Pr\{T > \gamma' ; \mathcal{H}_0\} \Rightarrow \gamma' = (s^T C^{-1} s)^{1/2} Q^{-1}(P_{FA})
\]

\[
P_D = Pr\{T > \gamma' ; \mathcal{H}_1\} = \frac{Q\left(\frac{\gamma' - s^T C^{-1} s}{(s^T C^{-1} s)^{1/2}}\right)}{Q\left(\frac{Q^{-1}(P_{FA}) - \sqrt{s^T C^{-1} s}}{\sigma^2}\right)}
\]

From the performance metric \( P_D \) we see that in colored noise, unlike in white Gaussian noise, the shape of the signal \( s \) can affect the performance.
Designing signals

From the performance metric $P_D$ we see that in colored noise, unlike in white Gaussian noise, the shape of the signal $s$ can affect the performance. The question is now how to design $s$ to maximize $P_D$, subject to a given desired power of the signal $\mathcal{E} = s^T s$ (else best performance will result from taking a signal of infinite energy).

Attack?

Using Lagrangians, we find that if $\lambda_{\text{min}}$ is the smallest eigenvalue of $C$ with corresponding normalized eigenvector $\mathbf{v}_{\text{min}}$ then the signal $s$ that maximizes $P_D$ under an energy constraint of $\mathcal{E}$ is $s = \sqrt{\mathcal{E}} \mathbf{v}_{\text{min}}$.

Be able to show this!

Multiple Deterministic Signals in Gaussian Noise

We are interested in simple ML detection in white Gaussian noise, i.e. selecting the hypothesis $\mathcal{H}_i$ for which $p(x|\mathcal{H}_i)$ is maximal. When you want to determine which of many signals was sent the general approach is to correlate the received signal with each of the possible hypothesis signals $s_i = [s_i[0], s_i[1], \ldots s_i[N-1]]^T$, adjust for the signal energy $\mathcal{E}_i = \sum_{n=0}^{N-1} s_i[n]^2$, and select the hypothesis which resulted in the maximal correlator output. That is, pick the $\mathcal{H}_i$ for which the following is maximum:

$$T_i(x) := \sum_{n=0}^{N-1} x[n]s_i[n] - \frac{1}{2} \mathcal{E}_i$$  \hspace{1cm} (1)

Let’s derive this decision rule!

Geometric meaning?
Detection: Random Signals

• What if s[n] is random?

\[ H_0 : x[n] = w[n] \]
\[ H_1 : x[n] = s[n] + w[n], \]

• Key idea behind estimator-correlator:

Estimate the signal first, then matched-filter the estimate

• Linear model simplifies things again...

The problem

Consider a binary hypothesis testing model of the following form:

\[ H_0 : x[n] = w[n] \]
\[ H_1 : x[n] = s[n] + w[n], \]

where \( w \sim \mathcal{N}(0, C_w) \) and \( s \sim \mathcal{N}(\mu_s, C_s) \) and \( s, w \) are independent. We have \( n = 0, 1, \cdots N - 1 \) (N samples).

We thus can discriminate between the two hypothesis based on both their means and covariances. Taking the likelihood ratio and simplifying, our test statistic \( T(x) \) can be shown to be:

\[
T(x) = \frac{1}{2} x^T \left[ C_w^{-1} C_s (C_s + C_w)^{-1} \right] x + x^T (C_s + C_w)^{-1} \mu_s
\]  

(1)

The test statistic has a quadratic term in \( x \) (intuitively account for the different variances) as well as a linear term in \( x \) accounting for the different means.
Example: Linear Model

Linear model? We now have

\[ H_0 : \mathbf{x} = \mathbf{w} \]
\[ H_1 : \mathbf{x} = \mathbf{H}\theta + \mathbf{w}, \]

where \( \mathbf{w} \sim \mathcal{N}(0, \mathbf{C}_w) \), \( \mathbf{H} \) is a known \( N \times p \) observation matrix, and \( \theta \sim \mathcal{N}(0, \mathbf{C}_\theta) \) and \( \theta, \mathbf{w} \) are independent. We have \( n = 0, 1, \ldots, N - 1 \) (\( N \) samples).

The test statistic becomes

\[
T(\mathbf{x}) = \mathbf{x}^T \left[ \mathbf{C}_s (\mathbf{C}_s + \mathbf{C}_w)^{-1} \right] \mathbf{x} \\
= \mathbf{x}^T \mathbf{H} \mathbf{C}_\theta \mathbf{H}^T (\mathbf{H} \mathbf{C}_\theta \mathbf{H}^T + \mathbf{C}_w)^{-1} \mathbf{x}
\]

Detection: Statistical Decision Theory II

We now ask the question: “how can we detect a signal presence when the probability density functions under hypotheses \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) are not exactly known?” We do assume that the pdfs are parametrized and that these parameters may or may not be known. More specifically, under \( \mathcal{H}_0 \) the set of unknown parameters is \( \theta_0 \), while under \( \mathcal{H}_1 \) the set of parameters is \( \theta_1 \).

- Uniformly most powerful test
- Generalized likelihood ratio test
- Bayesian approach
- Wald test
- Rao test
Further resources

http://www.ece.uic.edu/~devroye/courses/ECE531/