STATISTICAL SIGNAL PROCESSING

Statistical DSP

Digital ⟶ discrete-time, sampled, quantized

Signal ⟶ waveform, sequence of measurements or observations

Processing ⟶ analyze, modify, synthesize

Examples of digital signals

- sampled speech waveform
- pixelized image
- Dow-Jones index
- stream of Internet packets
- vector of medical predictors

What kind of processing might be desirable?

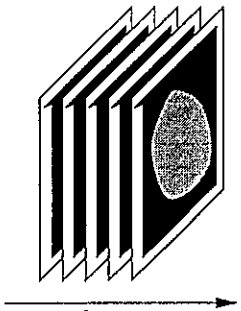
A major difficulty

In many DSP applications, we don't have complete or perfect knowledge of the signals we wish to process. We are faced with many unknowns and uncertainties.

Examples

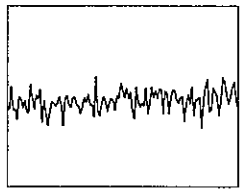
- unknown signal parameters
  - delay of radar return
  - pitch of speech signal
- environmental noise
  - multipath signals in wireless comm.
  - ambient EM waves
  - radar jamming
- sensor noise
  - grainy images
  - old phonograph recordings
- variability inherent in nature
  - stock market
  - Internet

Functional Magnetic Resonance Imaging

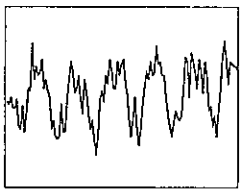


brain image
time-series

$\Rightarrow$



"inactive" pixel
time-series

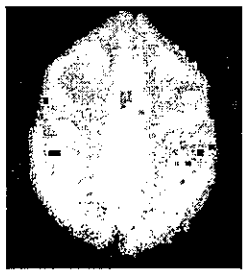


"active" pixel
time-series

$\Downarrow$

statistical signal processing

$\Leftarrow$



activation map

Challenges:

- measurement noise
- intrinsic uncertainties in signal behavior

How can we process signals in the face of such uncertainty?

Can we model the uncertainty and incorporate this model into the processing?

Statistical signal processing is the study of these questions.

Modeling uncertainty

The most widely accepted and commonly used approach to modeling uncertainty

(a) is _____, although alternatives exist such as fuzzy logic.

Probability theory models uncertainty by specifying the chance of observing certain signals.

Alternatively, one can view probability as specifying the degree to which we believe a signal reflects the true state of nature.

Examples of probabilistic models

- sensor noise modeled as an additive Gaussian random variable
- uncertainty in the phase of a sinusoidal signal modeled as a _____ random variable on $[0, 2\pi)$.

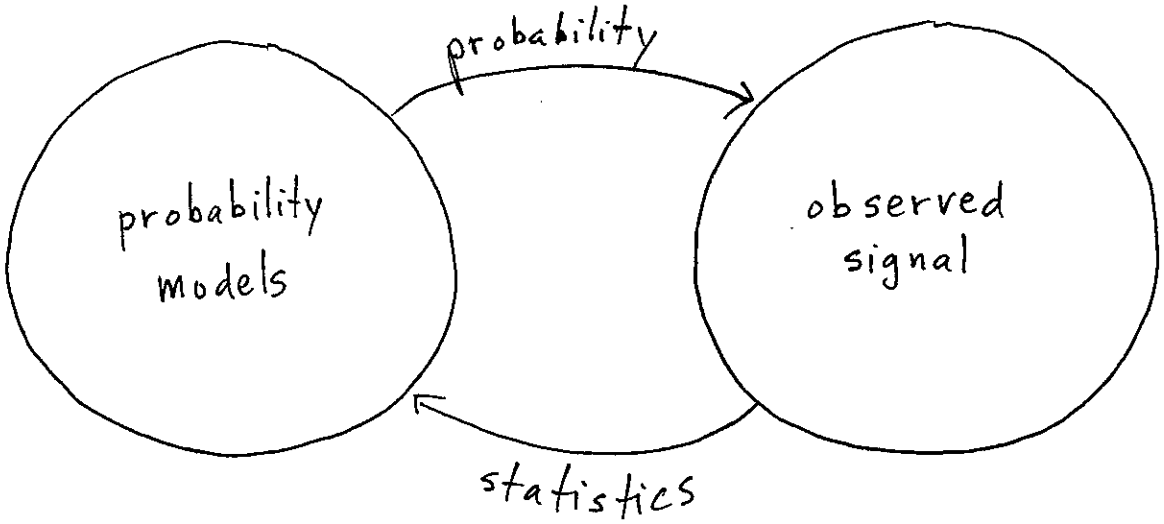
(b)

- uncertainty in the number of photons striking a CCD per unit time modeled as a _____ random variable.

(c)

Probability laws describe the uncertainty in the signals we <u>might</u> observe.

Statistics describe the salient features of the signals we <u>do</u> observe, and allow us to draw conclusions (inferences) about which probability model actually reflects the true state of nature



Statistical inference

A statistic is a function of observed data, and may be scalar or vector valued.

Examples | Suppose we observe n scalar values $x_1, \ldots, x_N$. The following are statistics

- sample mean $\longrightarrow$ $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- the data itself $\longrightarrow$ $[x_1, \cdots, x_N]^T$
- an order statistic $\longrightarrow$ $x_{(1)} = \min \{ x_1, \ldots, x_N \}$
- an arbitrary function $\longrightarrow$ $[x_1^2 - x_2 \sin(x_3), e^{-x_1 x_3}]^T$

A statistic cannot depend on unknown quantities.

Statistical signal processing (in a nutshell)

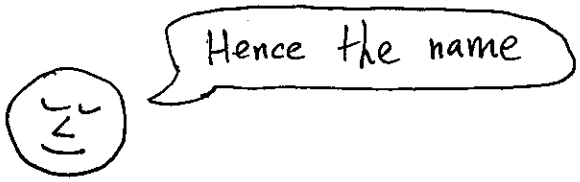
Step 1: Postulate a probability model (or models) that can be expected to reasonably capture the uncertainties in the data.

Step 2: Collect data

Step 3: Formulate statistics that allow us to interpret or understand our probability models.

In this class we will focus on three areas:

- Estimation
- Filtering
- Detection



Estimation

If our probability model has free parameters, can we use the measured signal to infer the actual parameter values?

Examples

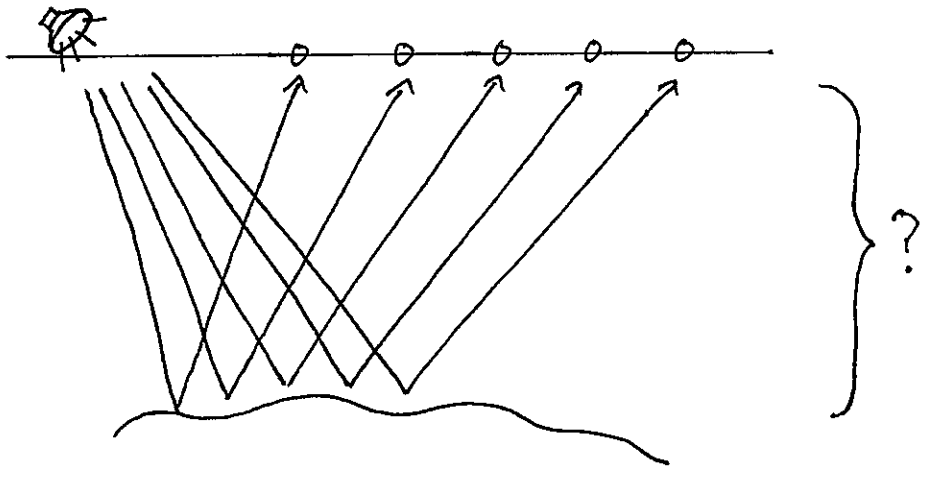
- Signal denoising: If we observe

$$\underline{x} = \underline{s} + \underline{w}$$

where $\underline{s}$ is a signal of interest and $\underline{w}$ is noise, estimate $\underline{s}$.

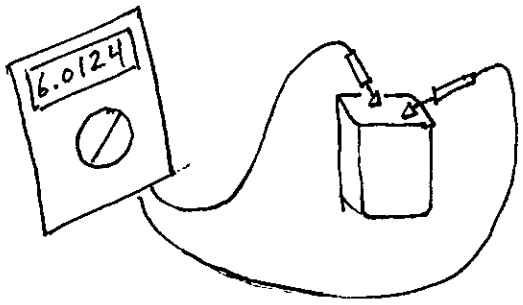
- If $s(n) = A\cos(2\pi f n + \phi)$, estimate A, ϕ, f in signal plus noise model.

- Seismology: estimate depth below ground of an oil pool based on reflected acoustic waves.



Here is a more concrete example:

Example Suppose we measure the voltage A of a battery using a voltmeter. Because the voltmeter tends to pick up noise from nearby objects, we take n measurements in hopes of gaining some accuracy.



Step 1: Assume a Gaussian noise model

$$x_i = A + w_i, \quad i = 1, \ldots, N$$

where

$$w_i \sim N(0, \sigma_w^2)$$

Step 2: Gather data

Step 3: Estimate A via the sample mean

$$\hat{A} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Is this the "right" statistic for this noise model? How accurate is the estimate? What if σ_w^2 is unknown?

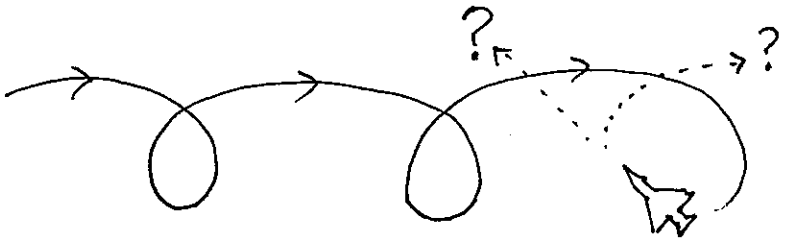
Filtering

Filtering in deterministic DSP refers to modifying a signal by means of a linear operator (usually expressed in terms of convolution).

In statistical DSP, filtering refers to signal estimation by means of a linear function of the data. Thus filtering is a special case of estimation, but it is distinguished by being concerned with online/real-time estimation of streaming data.

Examples

- Denoise a speech/audio signal real-time
- Track a moving target or predict its future location



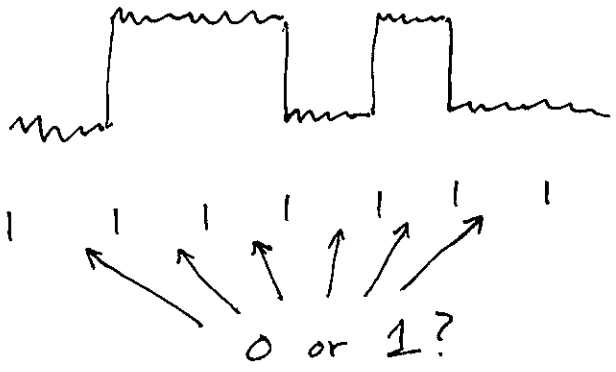
Detection

Given two (or more) probability models, which one best explains the observed signal?

Alternatively, given a single probability model, is it or is it not a valid characterization of the data?

Examples

- Decode a comm. signal into a sequence of 0's and 1's.



- Are other ships present on radar/sonar, and if so, are they friendly?

- Is a supernova exploding somewhere in the field of view of a certain telescope?

A more concrete detection example.

Example | Suppose you are given a coin and are asked to determine whether the following hypothesis is true:

H_0: the coin is fair

Step 1: Assume each toss of the coin is a realization of a _____ random var.

(d)

$$X \sim$$

where $p = \text{Prob}\{\text{heads}\}$.

Step 2: Toss the coin $n = 100$ times

$x_i = 1 \iff$ heads

$x_i = 0 \iff$ tails

Step 3: To assess the hypothesis

$$H_0 : p = \tfrac{1}{2},$$

form the statistic

$$k = \sum_{i=1}^{100} x_i$$

and reject H_0 if $|k - 50| > 10$.

In these examples, we used our intuition and heuristics in Step 3 to solve estimation and detection problems.

In this course we will develop principled and mathematically rigorous approaches to estimation, filtering, and detection using the theoretical framework of probability and statistics.

Summary

DSP = processing digital signals with computer algorithms

SSP = statistical DSP

= processing in the presence of uncertainties and unknowns.

Key

a. probability theory
b. uniform
c. Poisson
d. Bernoulli

SIGNAL SUBSPACES, ORTHOGONAL PROJECTIONS, AND LEAST SQUARES ESTIMATION

The Signal Subspace Model

Let $\underline{a}_1, \ldots, \underline{a}_p \in \mathbb{R}^N$ (or $\mathbb{C}^N$) be linearly independent (so $p \leq N$), and consider the $N \times p$ matrix

$$A = [\underline{a}_1 \cdots \underline{a}_p].$$

Let $\langle A \rangle$ denote the linear span of the columns of A (equivalently, the image of A). Then

$$\dim(\langle A \rangle) = \text{rank}(A) = p$$

$$\dim(\langle A \rangle^{\perp}) = N - p$$

Let $\underline{b}_1, \ldots, \underline{b}_{N-p} \in \mathbb{R}^N$ (or $\mathbb{C}^N$) be a basis for $\langle A \rangle^\perp$ and set

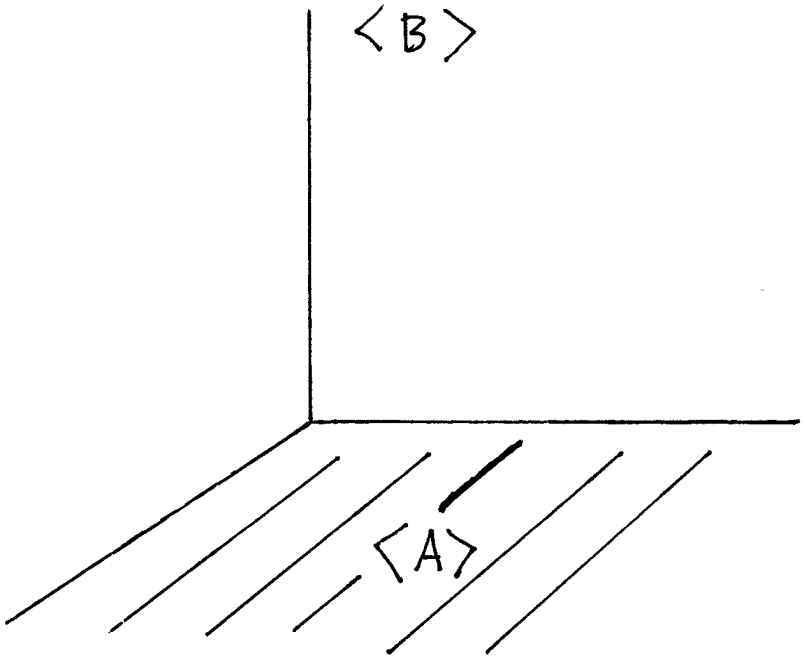
$$B = [\underline{b}_1 \cdots \underline{b}_{N-p}]. \qquad N \times (N-p)$$

Then

$$\langle \underline{a}_i, \underline{b}_j \rangle = 0 \qquad \begin{matrix} i = 1, \ldots, p \\ j = 1, \ldots, N-p. \end{matrix}$$

(a) and $\{\underline{a}_1, \ldots, \underline{a}_p, \underline{b}_1, \ldots, \underline{b}_{N-p}\}$ is a _____.

The subspaces $\langle A \rangle$ and $\langle B \rangle$ form an orthogonal decomposition of $\mathbb{R}^N$ (or $\mathbb{C}^N$)



$$\langle A \rangle \oplus \langle B \rangle = \mathbb{R}^N \text{ (or } \mathbb{C}^N)$$

Note) The vectors $\underline{a}_1, \ldots, \underline{a}_p$ are not necessarily orthogonal among themselves. The same goes for $\underline{b}_1, \ldots, \underline{b}_{N-p}$.

In the signal subspace model, we assume our observed signal $\underline{x}$ has the form

$$\underline{x} = \underline{s} + \underline{w}$$

where

$\underline{s} \in \langle A \rangle$ is the signal of interest

$\underline{w}$ is entirely noise.

We use the following terminology:

$\langle A \rangle =$

(b) $\langle B \rangle =$

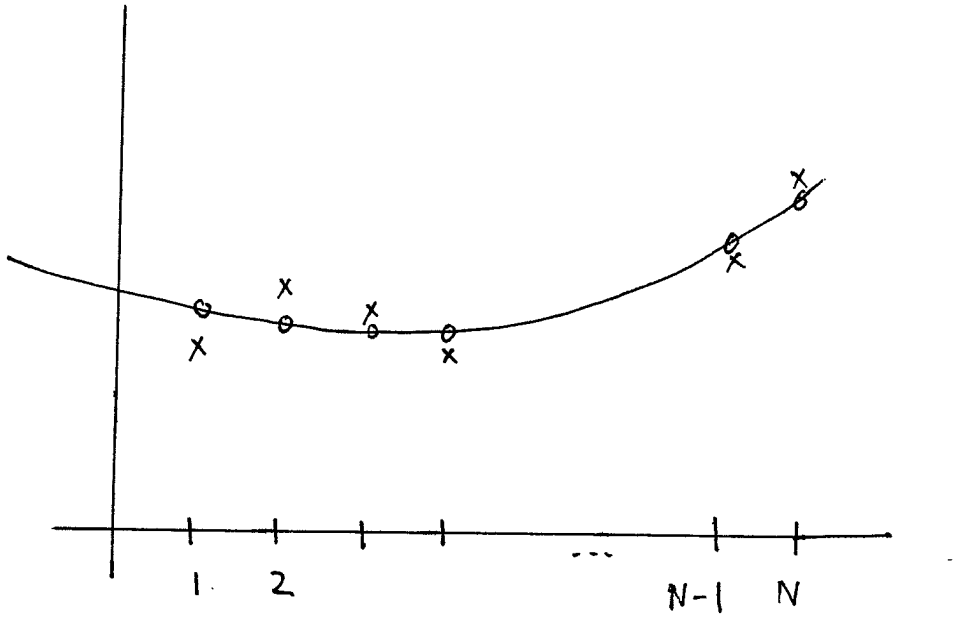
even though $\underline{w} \notin \langle B \rangle$ in general.

Example Quadratic polynomial model

$$s(n) = \theta_2 n^2 + \theta_1 n + \theta_0 \,, \quad n = 1, \ldots, N$$

$$\Rightarrow \quad \underline{s} = A\underline{\theta} \quad \text{where}$$

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ \vdots & \vdots & \vdots \\ 1 & N & N^2 \end{bmatrix} \qquad \underline{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \in \mathbb{R}^3$$



o → clean signal (unobserved)

x → noisy signal (observed)

Exercise Consider the sinusoidal signal

$$s(n) = D \cdot \cos(2\pi f n + \phi), \quad n = 0, \ldots, N-1$$

where f is known but D, ϕ are unknown.

Express $\underline{s} = [s(0) \ \cdots \ s(N-1)]^T$ as an element in a two-dimensional subspace.

That is, write

$$\underline{s} = A \cdot \underline{\theta}$$

where A is a <u>known</u> $N \times 2$ matrix and $\underline{\theta}$ is <u>unknown</u>.

Solution 1 Use $\cos(\alpha) = \frac{e^{j\alpha} + e^{-j\alpha}}{2}$. Then

$$s(n) = \underbrace{\left(\frac{D}{2}e^{j\phi}\right)}_{\theta_1} e^{2\pi j f n} + \underbrace{\left(\frac{D}{2}e^{-j\phi}\right)}_{\theta_2} e^{-2\pi j f n}$$

$$\Rightarrow \underline{s} = \begin{bmatrix} 1 & 1 \\ e^{2\pi j f} & e^{-2\pi j f} \\ e^{4\pi j f} & e^{-4\pi j f} \\ \vdots & \vdots \\ e^{2(N-1)\pi j f} & e^{-2(N-1)\pi j f} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

$$\underline{g}_1, \underline{g}_2 \in \mathbb{C}^N, \quad \underline{\theta} \in \mathbb{C}^2$$

Solution 2 Use $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$

$$s(n) = \underbrace{\left(D\cos(\phi)\right)}_{\theta_1} \cos(2\pi f n) + \underbrace{\left(-D\sin(\phi)\right)}_{\theta_2} \sin(2\pi f n)$$

$$\Rightarrow \underline{s} = \begin{bmatrix} 1 & 0 \\ \cos(2\pi f) & \sin(2\pi f) \\ \cos(4\pi f) & \sin(4\pi f) \\ \vdots & \vdots \\ \cos(2(N-1)\pi f) & \sin(2(N-1)\pi f) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

$$\underline{g}_1, \underline{g}_2 \in \mathbb{R}^N, \quad \underline{\theta} \in \mathbb{R}^2 \quad \text{if} \quad D \in \mathbb{R}$$

Orthogonal Projection

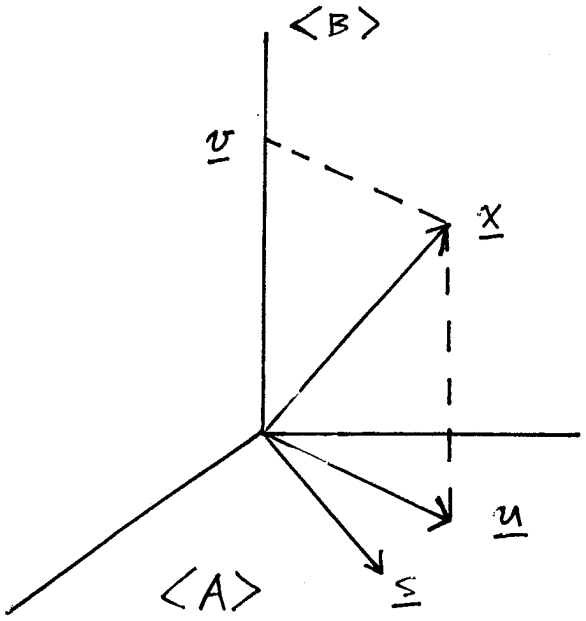
How can we use the knowledge of $\langle A \rangle$ to estimate $\underline{s}$ from $\underline{x} = \underline{s} + \underline{w}$?

Since $\langle A \rangle$ and $\langle B \rangle$ are orthogonal complements, we can uniquely write

$$\underline{x} = \underline{u} + \underline{v}$$

where $\underline{u} \in \langle A \rangle$ and $\underline{v} \in \langle B \rangle$.

Since $\underline{v}$ is pure noise, it makes sense to remove it.



In general, $\underline{s} \neq \underline{u}$ because $\underline{w}$ has some component in $\langle A \rangle$.

The operator that maps $\underline{x} \mapsto \underline{u}$ is called

(c) _____ onto $\langle A \rangle$ and is

denoted Π_A.

Proposition

$$\Pi_A = A (A^H A)^{-1} A^H$$

Proof Let $\underline{x} = \underline{u} + \underline{v}$, $\underline{u} \in \langle A \rangle$, $\underline{v} \in \langle B \rangle$.

We must show $A (A^H A)^{-1} A^H \underline{x} = \underline{u}$. Since the columns of A and B form a basis, we can write

$$\underline{u} = A \underline{\theta}, \quad \underline{\theta} \; p \times 1$$

$$\underline{v} = B \underline{\phi}, \quad \underline{\phi} \; (N-p) \times 1$$

Then

$$A (A^H A)^{-1} A^H \underline{x}$$

$$= A \underbrace{(A^H A)^{-1} A^H A}_{I_{p \times p}} \underline{\theta} + A (A^H A)^{-1} \underbrace{A^H B}_{0_{p \times (N-p)}} \underline{\phi}$$

$$= A \underline{\theta}$$

$$= \underline{u}$$

▨

Properties of projections

- $\Pi_A^H =$ "self-adjoint"
- $\Pi_A^2 =$ "idempotent"
- $\Pi_A + \Pi_B =$
- $\Pi_A \cdot \Pi_B =$
- If $q_1, \ldots, q_p$ are orthonormal, then

$$\Pi_A = A A^H$$

Filtering interpretation

The projection operator is analogous to a bandpass filter; we only retain that information which resides in the passband, which corresponds to the signal subspace.

Least Squares Estimation

To estimate $\underline{s} = A\underline{\theta}$ where

$$\underline{x} = \underline{s} + \underline{w}$$

we use the projection onto $\langle A \rangle$:

$$\hat{\underline{s}} = \Pi_A \underline{x}$$

$$= A\left(A^H A\right)^{-1} A^H \underline{x}.$$

What if we want to estimate $\underline{\theta}$?

An estimate $\hat{\underline{\theta}}$ of $\underline{\theta}$ should satisfy

$$\hat{\underline{s}} = A\hat{\underline{\theta}}.$$

Therefore, an obvious estimate is

$$\hat{\underline{\theta}} =$$

©

It turns out that this is the solution to the least squares problem.

Proposition The unique solution of

$$\min_{\underline{\theta}} \|\underline{x} - A\underline{\theta}\|^2 \qquad (\underline{\theta} \in \mathbb{R}^p \text{ or } \mathbb{C}^p)$$

is $\hat{\underline{\theta}} = (A^H A)^{-1} A^H \underline{x}$.

Proof Write $\underline{x} = \underline{u} + \underline{v}$ where $\underline{u} \in \langle A \rangle$ and $\underline{v} \in \langle A \rangle^{\perp}$. Observe

$$\|\underline{x} - A\underline{\theta}\|^2 = \|\underline{u} - A\underline{\theta} + \underline{v}\|^2$$

$$= \langle \underline{u} - A\underline{\theta} + \underline{v}, \underline{u} - A\underline{\theta} + \underline{v} \rangle$$

$$= \langle \underline{u} - A\underline{\theta}, \underline{u} - A\underline{\theta} \rangle + \langle \underline{v}, \underline{v} \rangle$$

$$+ \underbrace{\langle \underline{u} - A\underline{\theta}, \underline{v} \rangle}_{=0} + \underbrace{\langle \underline{v}, \underline{u} - A\underline{\theta} \rangle}_{=0}$$

$$= \|\underline{u} - A\underline{\theta}\|^2 + \|\underline{v}\|^2.$$

The second term is independent of $\underline{\theta}$. Therefore, to minimize the expression, the best we can do is to make the first term 0 by taking

$$\underline{\theta} = \hat{\underline{\theta}} = (A^H A)^{-1} A^H \underline{x}.$$

Then $A\hat{\underline{\theta}} = \Pi_A \underline{x} = \underline{u}$. To see that $\hat{\underline{\theta}}$ is unique, if $\underline{\theta}'$ is also such that $\|\underline{u} - A\underline{\theta}'\| = 0$, then

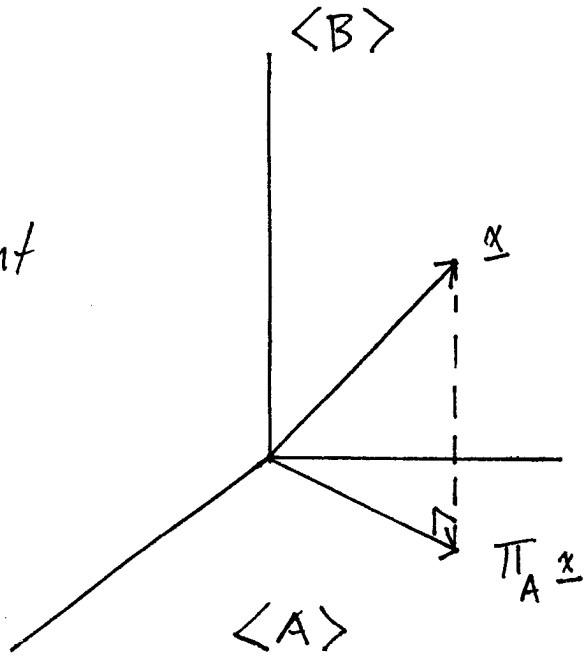
$$A\underline{\theta}' = \underline{u} \implies A\underline{\theta}' = A\hat{\underline{\theta}}$$

$$\implies \underline{\theta}' = \hat{\underline{\theta}}$$

since the columns of A are linearly independent. ▨

Minimum distance property

We may conclude that $\Pi_A \underline{x}$ is the unique point in $\langle A \rangle$ that is closest to $\underline{x}$.



Remark | The operator

$$A^{\#} := (A^H A)^{-1} A^H$$

is called the pseudo-inverse of A.

Example Recall the sinusoid

$$s(n) = D \cdot \cos(2\pi f n + \phi), \quad n = 0, \ldots, N-1$$

where f is known and D, ϕ are unknown.

How can we estimate D and ϕ from $\underline{x}$?

Complex solution: $\underline{x} = A\underline{\theta}$ where

$$\underline{\theta} = \begin{bmatrix} \frac{D}{2} e^{j\phi} \\ \frac{D}{2} e^{-j\phi} \end{bmatrix}.$$ Use pseudo-inverse to

compute $\hat{\underline{\theta}}$ and form

$$\hat{D} = \sqrt{4 \cdot \hat{\theta}_1 \cdot \hat{\theta}_2}, \quad \hat{\phi} = \frac{1}{2} \angle \left(\frac{\hat{\theta}_1}{\hat{\theta}_2} \right)$$

Real solution: $\underline{x} = A\underline{\theta}$ where

$$\underline{\theta} = \begin{bmatrix} D\cos(\phi) \\ -D\sin(\phi) \end{bmatrix}.$$ Use pseudo-inverse to

compute $\hat{\underline{\theta}}$ and form

$$\hat{D} = \sqrt{\hat{\theta}_1^2 + \hat{\theta}_2^2}, \quad \hat{\phi} = \tan^{-1}\left(-\frac{\hat{\theta}_2}{\hat{\theta}_1} \right)$$

(assuming D to be real)

Summary

- If a signal lies in a subspace, it can be estimated by projection onto that subspace.
- This "filters out" any noise in the noise subspace.
- The projection satisfies the minimum distance property, and is closely related to the least squares problem.
- This approach is <u>non-statistical</u> because no probability model is specified for the noise. Yet it turns out to be equivalent or similar to many methods we will see later.

Key

a. basis b. signal subspace, noise subspace

c. orthogonal projection

d. Π_A, Π_A, $I_{N\times N}$, $O_{N\times N}$

e. $(A^H A)^{-1} A^H \underline{x}$

EIGENDECOMPOSITIONS & THE SPECTRAL THEOREM

The Spectral Theorem

Definition If $U \in \mathbb{C}^{N \times N}$ is such that

$$U^H U = U U^H = I_{N \times N}$$

then U is said to be *unitary*.

If $U \in \mathbb{R}^{N \times N}$ is such that

$$U^T U = U U^T = I_{N \times N}$$

then U is said to be *orthogonal*

[slightly confusing since the columns of U are in fact orthonormal]

Intuitively, such matrices are distance preserving since

$$\begin{aligned} \|U\underline{x} - U\underline{y}\|^2 &= (U\underline{x} - U\underline{y})^H (U\underline{x} - U\underline{y}) \\ &= (\underline{x} - \underline{y})^H U^H U (\underline{x} - \underline{y}) \\ &= \|\underline{x} - \underline{y}\|^2 \end{aligned}$$

Unitary and orthogonal matrices effect a change of coordinate system.

Theorem (Spectral Theorem)

If $A \in \mathbb{C}^{N \times N}$ is Hermitian, then there exist a unitary matrix U and a real diagonal matrix Λ such that

$$A = U \Lambda U^H.$$

If $A \in \mathbb{R}^{N \times N}$ is symmetric, the same result holds where now U is orthogonal.

Proof See Moon and Stirling, *Mathematical Methods and Algorithms for Signal Processing*.

Eigenvalues and eigenvectors

Suppose A is Hermitian/symmetric. Write

$$A = U \Lambda U^H$$

according to the spectral theorem. Let

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{bmatrix}, \quad U = \begin{bmatrix} \underline{u}_1 & \underline{u}_2 & \cdots & \underline{u}_N \end{bmatrix}$$

Since $AU = U\Lambda$, we conclude that the λ_i are eigenvalues with $\underline{u}_i$ the associated eigenvector:

$$A\underline{u}_i = \lambda_i \underline{u}_i, \quad i = 1, \ldots, N.$$

The spectral theorem also gives rise to following <u>spectral decomposition</u> of A:

$$A = \sum_{i=1}^{N} \lambda_i \underline{u}_i \underline{u}_i^H$$

Positive (semi) definite matrices

Let A be a Hermitian / symmetric $N \times N$ matrix.

We say A is *positive definite* (PD) if

$$\underline{x} \neq 0 \implies \underline{x}^H A \underline{x} > 0.$$

We say A is *positive semi-definite* (PSD) if

$$\forall \underline{x} \qquad \underline{x}^H A \underline{x} \geq 0.$$

[PSD is also called nonnegative definite]

Exercise | Show that A is PD (PSD) iff the eigenvalues of A are positive (nonnegative).

Solution By the spectral theorem, $\underline{u}_1, \ldots, \underline{u}_N$ is an orthonormal collection (a basis, in fact) such that

$$A\underline{u}_i = \lambda_i \underline{u}_i.$$

For each $i = 1, \ldots, N$ we have

$$\begin{aligned}
\lambda_i &= \lambda_i \cdot \underline{u}_i^H \underline{u}_i \\
&= \underline{u}_i^H \cdot \lambda_i \underline{u}_i \\
&= \underline{u}_i^H A \underline{u}_i
\end{aligned}
\quad \begin{cases} > 0 & \text{if } A \text{ is PD} \\ \geq 0 & \text{if } A \text{ is PSD.} \end{cases}$$

Conversely, suppose $\lambda_i > 0$ (≥ 0) $\forall i$.

Then for $\underline{x} \neq 0$ we have

$$\begin{aligned}
\underline{x}^H A \underline{x} &= \sum_{i=1}^{n} \lambda_i \underline{x}^H \underline{u}_i \underline{u}_i^H \underline{x} \\
&= \sum_{i=1}^{n} \lambda_i \| \underline{u}_i^H \underline{x} \|^2
\end{aligned}$$

$$\begin{cases} > 0 & \text{if } \lambda_i > 0 \ \forall i \\ \geq 0 & \text{if } \lambda_i \geq 0 \ \forall i \end{cases}$$

THE MULTIVARIATE GAUSSIAN DISTRIBUTION

Density

Let $\underline{\mu} \in \mathbb{R}^N$ and $R \in \mathbb{R}^{N \times N}$ be symmetric and positive definite. A random variable $\underline{X}$ has a multivariate Gaussian distribution if its density is

$$f(\underline{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |R|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T R^{-1} (\underline{x}-\underline{\mu})}$$

Mean and covariance

$$\underline{\mu} = E[\underline{X}]$$

$$R = E[(\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})^T]$$

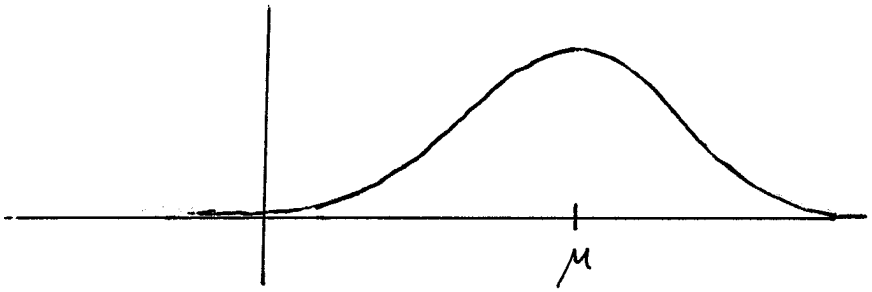
Notation

$$\underline{X} \sim \mathcal{N}(\underline{\mu}, R)$$

Conceptualization

In 1-d, $R = [\sigma^2]$ (1×1) and

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



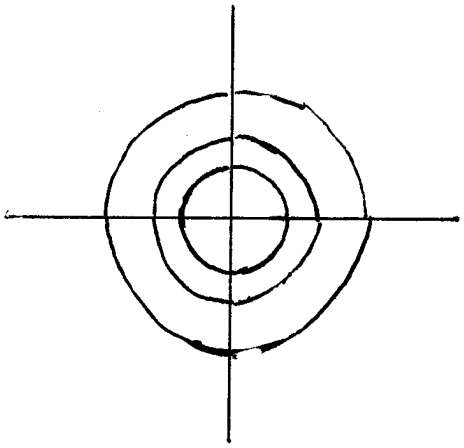
bell curve

In 2-d, let's consider 3 cases:

Case 1: $R = \sigma^2 I_{2\times 2} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$

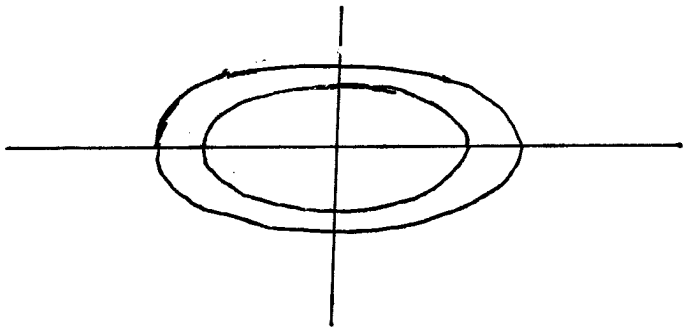
Then a contour of the density is a circle:

$$f(\underline{x}) \equiv \gamma \iff \|\underline{x} - \underline{\mu}\|^2 \equiv \gamma'$$



Case 2: $R = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$, say $\sigma_1 > \sigma_2$

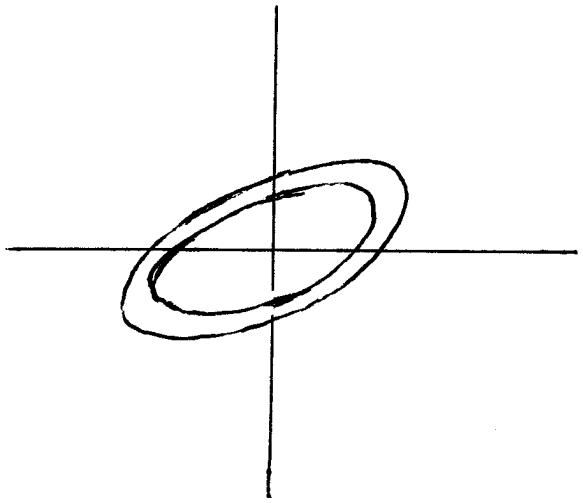
Then the density contours are ellipses whose axes align with the standard basis.



To see this, observe

$$f(\underline{x}) \equiv \gamma \iff \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \equiv \gamma'$$

Case 3: R is arbitrary. Then density contours are ellipses with arbitrary orientation.



To see this, write

$$R = U \Lambda U^T$$

Then

$$(\underline{x} - \underline{\mu})^T R^{-1} (\underline{x} - \underline{\mu})$$

$$= (\underline{x} - \underline{\mu})^T U \Lambda^{-1} U^T (\underline{x} - \underline{\mu})$$

$$= (\underline{x}' - \underline{\mu}')^T \Lambda^{-1} (\underline{x}' - \underline{\mu}')$$

$$\left[\text{where } \underline{x}' = U^T \underline{x}, \ \underline{\mu}' = U^T \underline{\mu} \right]$$

$$= \frac{(x_1' - \mu_1')^2}{\lambda_1} + \frac{(x_2' - \mu_2')^2}{\lambda_2}$$

which defines an ellipse in the rotated coordinate system.

More generally, the MVG distribution

- is symmetric with respect to its mean
- is unimodal
- has ellipsoidal contours: axes $\Leftrightarrow$ eigenvectors of R
  and axis lengths $\Leftrightarrow$ eigenvalues of R

Importance

The MVG model is the most important and widely employed model in statistical signal processing. Some reasons for this include:

- Tractability
- Estimators and detectors with intuitive forms and properties
- Justification in terms of the central limit theorem.

CLT: If $\underline{X} = \frac{1}{n}\sum_{i=1}^{n} \underline{Y}_i$, then

rough paraphrase

$$\underline{X} \to \mathcal{N}(\underline{\mu}, R) \quad \text{as} \quad n \to \infty$$

for some $\underline{\mu}$, R, regardless of the distribution of $\underline{Y}$.

Example In communication systems, electronic noise is due to the aggregate effect of huge numbers of charge carriers undergoing random motion.

Characteristic function

The characteristic function of an N dimensional random variable $\underline{X}$ is defined to be

$$\underline{\Phi}(\underline{\omega}) = E\left[e^{-j\underline{\omega}^T\underline{X}}\right]$$

$$= \int e^{-j\underline{\omega}^T\underline{x}} f(\underline{x})\, d\underline{x}$$

The char. fun. is an N-dim Fourier transform of the density of $\underline{X}$. Thus it uniquely characterizes the random variable. The density may be recovered from Φ by taking the inverse Fourier transform.

For the MVG, $\underline{X} \sim N(\underline{\mu}, R)$ we have

$$\Phi(\underline{\omega}) = E[e^{-j\underline{\omega}^T\underline{X}}]$$

$$= \int e^{-j\underline{\omega}^T\underline{x}} f(\underline{x})\, d\underline{x}$$

$$= \int (2\pi)^{-\frac{N}{2}} |R|^{-\frac{1}{2}} \exp\left\{-j\underline{\omega}^T\underline{x} - \frac{1}{2}(\underline{x}-\underline{\mu})^T R^{-1}(\underline{x}-\underline{\mu})\right\} d\underline{x}$$

complete the square →

$$= e^{-j\underline{\omega}^T\underline{\mu} - \frac{1}{2}\underline{\omega}^T R\underline{\omega}} \times$$

$$\underbrace{\int (2\pi)^{-\frac{N}{2}} |R|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}+jR\underline{\omega})^T R^{-1}(\underline{x}-\underline{\mu}+jR\underline{\omega})\right\} d\underline{x}}$$

Gaussian density $\rightarrow 1$

$$= e^{-j\underline{\omega}^T\underline{\mu} - \frac{1}{2}\underline{\omega}^T R\underline{\omega}}$$

Linear Transformations

Proposition If $\underline{X} \sim \mathcal{N}(\underline{\mu}, R)$ is N-dim, $A \in \mathbb{R}^{M \times N}$, and $\underline{Y} = A\underline{X}$, then

$$\underline{Y} \sim \mathcal{N}(A\underline{\mu}, ARA^T)$$

Proof

$$\begin{aligned}
\underline{\Phi}_Y(\underline{\omega}) &= E\left[e^{-j\underline{\omega}^T \underline{Y}}\right] \\
&= E\left[e^{-j\underline{\omega}^T A\underline{X}}\right] \\
&= E\left[e^{-j(A^T\underline{\omega})^T \underline{X}}\right] \\
&= \underline{\Phi}_X(A^T\underline{\omega}) \\
&= e^{-j\underline{\omega}^T A\underline{\mu} - \frac{1}{2}\underline{\omega}^T ARA^T \underline{\omega}}
\end{aligned}$$

$$\Rightarrow \underline{Y} \sim \mathcal{N}(A\underline{\mu}, ARA^T)$$

since the characteristic function uniquely characterizes a distribution.

Marginals

Proposition Let $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix}$, $\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}$, $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$.

If $\underline{X} \sim N(\underline{\mu}, R)$, then

$$\underline{X}_1 \sim N(\underline{\mu}_1, R_{11}).$$

Exercise Prove this.

Solution Write $\underline{X}_1 = A\underline{X}$ where

$$A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \quad (p \times N)$$

assuming $\underline{X} \in \mathbb{R}^N$, $\underline{X}_1 \in \mathbb{R}^p$.

Then

$$A\underline{\mu} = \underline{\mu}_1$$

$$ARA^T = R_{11}$$

Now apply the previous result.

Conditioning

Proposition Let $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix}$, $\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}$, $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$.

If $\underline{X} \sim N(\underline{\mu}, R)$, then

$$\underline{X}_2 \mid \underline{X}_1 = \underline{x}_1 \sim N(\tilde{\underline{\mu}}, \tilde{R})$$

where

$$\tilde{\underline{\mu}} = \underline{\mu}_2 + R_{21} R_{11}^{-1} (\underline{x}_1 - \underline{\mu}_1)$$

$$\tilde{R} = R_{22} - R_{21} R_{11}^{-1} R_{12}$$

Proof | Write out $f(\underline{x}_2|\underline{x}_1) = \dfrac{f(\underline{x})}{f(\underline{x}_1)}$

and simplify. See Kay (Vol. I) or Moon and Stirling for details.

SUFFICIENT STATISTICS

Suppose the distribution of a random variable $\underline{X}$ is determined by a parameter $\underline{\theta}$:

$$\underline{X} \sim f_{\underline{\theta}}(\underline{x})$$

The functional form of f is known, but $\underline{\theta}$ is unknown.

[Note: We will use f to denote a pdf, p to denote a pmf, and f when it could be either.]

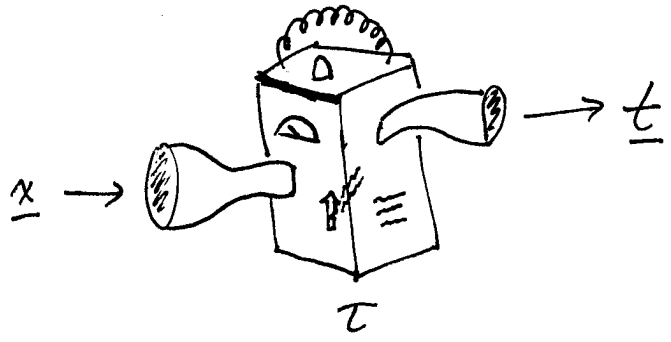
In statistical inference, we observe a realization $\underline{x}$ of $\underline{X}$ and need to answer some question about $\underline{\theta}$, such as

- Is $\underline{\theta} \in \mathcal{H}_1$ or is $\underline{\theta} \in \mathcal{H}_2$? (Detection)
- What is a good guess for $\underline{\theta}$? (Estimation)

If $\underline{x} = [x_1 \cdots x_N]^T$ and $\underline{\theta} = [\theta_1 \cdots \theta_p]^T$ where $p < N$, one might wonder whether it is possible to <u>compress</u> the measurement $\underline{x}$ into a low-dimensional statistic without affecting the quality of the inference about $\underline{\theta}$.

In other words, does there exist $\underline{T} = \tau(\underline{x})$, where the dimension of $\underline{T}$ is $M < N$, such that $\underline{T}$ carries all the useful information about $\underline{\theta}$?

If so, for the purpose of studying $\underline{\theta}$, we could discard the raw measurement $\underline{x}$ and retain only the compressed statistic $\underline{t}$.



Definition Let $\underline{X} \sim f_{\underline{\theta}}(\underline{x})$. The statistic

$\underline{T} = \tau(\underline{X})$ is a <u>sufficient statistic</u> for $\underline{\theta}$ if the conditional distribution of $\underline{X}$ given $\underline{T}$ is independent of $\underline{\theta}$. Equivalently, the functional form of $f(\underline{x}|\underline{t})$ does not involve $\underline{\theta}$.

Interpretations

1. Let $P_{\underline{\theta}}(\underline{x}, \underline{t})$ denote the joint pmf of $(\underline{X}, \underline{T})$. Then

$$P_{\underline{\theta}}(\underline{x}, \underline{t}) = \begin{cases} P_{\underline{\theta}}(\underline{x}) & \text{if } \underline{t} = \tau(\underline{x}) \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\begin{aligned} P_{\underline{\theta}}(\underline{x}) &= P_{\underline{\theta}}(\underline{x}, \tau(\underline{x})) \\ &= P_{\underline{\theta}}(\underline{x} \,|\, \tau(\underline{x})) P_{\underline{\theta}}(\tau(\underline{x})) \\ &= p(\underline{x} \,|\, \tau(\underline{x})) P_{\underline{\theta}}(\tau(\underline{x})) \end{aligned}$$

$\Rightarrow$ the dependence of $P_{\underline{\theta}}(\underline{x})$ on θ is manifested entirely in $P_{\underline{\theta}}(\underline{t})$.

[continuous case requires more care, but same conclusion holds – see Scharf]

2. Given $\underline{t} = \tau(\underline{x})$, full knowledge of $\underline{x}$ brings no additional information about $\underline{\theta}$.

3. Any inference strategy based on $f_{\underline{\theta}}(\underline{x})$ may be replaced by a strategy based on $f_{\underline{\theta}}(\underline{t})$.

Example] Bernoulli trials

Suppose we observe $\underline{x} = [x_1, \cdots x_N]^T$ where

$$X_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

$\theta \in [0,1]$ is unknown.

Recall

(a) $$p_\theta(x) =$$

Since we can assume $x_i \in \{0,1\}$, we may write

$$p_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Therefore

$$p_\theta(\underline{x}) = \prod_{i=1}^{n} p_\theta(x_i)$$

$$= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^k (1-\theta)^{N-k}$$

where $k = \sum_{i=1}^{n} x_i$.

Claim: K is a sufficient statistic for θ.

We must show $p_\theta(\underline{x} \mid k)$ is independent of θ.

From interpretation #1 we know

$$p_\theta(\underline{x} \mid k) = \frac{P_\theta(\underline{x})}{P_\theta(k)}.$$

Exercise | Complete this argument to establish that K is sufficient for θ.

Solution | K is a Binomial (N, θ) random variable.

Therefore

$$P_\theta(k) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

and

$$p(\underline{x} \mid k) = \frac{P_\theta(\underline{x})}{P_\theta(k)}$$

$$= \frac{\theta^k (1-\theta)^{N-k}}{\binom{N}{k} \theta^k (1-\theta)^{N-k}}$$

$$= \frac{1}{\binom{N}{k}}$$

which is independent of θ.

The Fisher-Neyman Factorization Theorem

In the previous example, we had to guess the sufficient statistic and work out the conditional pmf by hand. In general, it is difficult to verify the definition of sufficient statistic directly.

The following theorem allows us to identify and verify sufficient statistics more readily, and can be taken as a working definition of sufficiency.

Theorem Let $f_{\underline{\theta}}(\underline{x})$ be the density or mass function for $\underline{X}$. The statistic $\underline{T} = \tau(\underline{X})$ is sufficient for $\underline{\theta}$ iff there exist functions $g_{\underline{\theta}}(\underline{t})$ and $h(\underline{x})$ such that

$$f_{\underline{\theta}}(\underline{x}) = g_{\underline{\theta}}(\tau(\underline{x})) \cdot h(\underline{x})$$

Note: h is independent of θ.

Example Bernoulli trials revisited

$$P_\theta(\underline{x}) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^k (1-\theta)^{N-k}$$

$$= g_\theta(k) \cdot h(\underline{x})$$

where

$$g_\theta(k) = \theta^k (1-\theta)^{N-k}$$

$$h(\underline{x}) = 1$$

$\Longrightarrow$ K is sufficient for θ.

Proof of Theorem

We will assume $\underline{X}$ is discrete. The continuous case is slightly more involved - see Scharf or Kay, vol I.

First, assume $\underline{T}$ is sufficient for $\underline{\theta}$. Recall

$$P_{\underline{\theta}}(\underline{x}) = p(\underline{x}/\tau(\underline{x})) \cdot P_{\underline{\theta}}(\tau(\underline{x}))$$

Now take

$$g_{\underline{\theta}}(\underline{t}) = P_{\underline{\theta}}(\underline{t})$$

$$h(\underline{x}) = p(\underline{x}/\tau(\underline{x}))$$ ← Independent of $\underline{\theta}$ by sufficiency

For the other direction, assume $p_{\underline{\theta}}(\underline{x})$ may be written

$$p_{\underline{\theta}}(\underline{x}) = g_{\underline{\theta}}(\tau(\underline{x}))h(\underline{x}).$$

We need to show $\underline{T} = \tau(\underline{X})$ is sufficient for $\underline{\theta}$. That is, we need to show $p_{\underline{\theta}}(\underline{x} \mid \underline{t})$ is independent of $\underline{\theta}$.

Again, we will rely on the identity

$$p_{\underline{\theta}}(\underline{x} \mid \underline{t}) = \frac{p_{\underline{\theta}}(\underline{x})}{p_{\underline{\theta}}(\underline{t})}.$$

Since $\underline{X}$ and $\underline{T}$ are discrete, we have

$$p_{\underline{\theta}}(\underline{t}) = \sum_{\underline{x}' : \tau(\underline{x}') = \underline{t}} p_{\underline{\theta}}(\underline{x}').$$

Therefore

$$p_{\underline{\theta}}(\underline{x} \mid \underline{t}) = \frac{g_{\underline{\theta}}(\underline{t}) \cdot h(\underline{x})}{\sum_{\underline{x}' : \tau(\underline{x}') = \underline{t}} g_{\underline{\theta}}(\underline{t}) h(\underline{x}')}$$

$$= \frac{h(\underline{x})}{\sum_{\underline{x}' : \tau(\underline{x}') = \underline{t}} h(\underline{x}')} \leftarrow \boxed{\text{Independent of } \underline{\theta}}$$

∎

Note The FNFT gives us a formula for $P(\underline{x}|\underline{t})$, namely

$$P(\underline{x}|\underline{t}) = \frac{h(\underline{x})}{\sum_{\underline{x}': \tau(\underline{x}') = \underline{t}} h(\underline{x}')}$$

Example Bernoulli trials, part III

$h(\underline{x}) = 1$, so

$$p(\underline{x}|k) = \frac{1}{\sum_{\underline{x}': \sum_{i=1}^{N} x_i' = k} 1}$$

$$= \frac{1}{\#\{\underline{x}': \sum_{i=1}^{N} x_i' = k\}}$$

$$= \frac{1}{\binom{N}{k}}$$

Let look at an example of the FNFT for the continuous case:

Example] Gaussian with unknown mean

We are given $\underline{x} = [x_1 \cdots x_N]^T$ where

$$x_i \overset{iid}{\sim} N(\theta, \sigma^2)$$

and σ^2 is known.

$$f_\theta(\underline{x}) = \prod_{i=1}^{N} f_\theta(x_i)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\}$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\theta)^2\right\}$$

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{N} x_i^2 - 2\theta\sum_{i=1}^{N} x_i + N\theta^2\right]\right\}$$

(b)

where $t =$ $\Rightarrow$ t is sufficient for θ.

Example Gaussian w/ unknown mean and variance.

Now assume

$$x_i \overset{iid}{\sim} N(\theta_1, \theta_2), \quad i = 1, \ldots, N$$

where $\underline{\theta} = [\theta_1 \ \theta_2]^T$ is unknown. Then

$$f_{\underline{\theta}}(\underline{x}) = \underbrace{\left(\frac{1}{2\pi\theta_2}\right)^{\frac{N}{2}} \exp\left\{-\frac{1}{2\theta_2}\left[\sum_{i=1}^{N} x_i^2 - 2\theta_1 \sum_{i=1}^{N} x_i + N\theta_1^2\right]\right\}}_{g_{\underline{\theta}}(\underline{t})} \cdot \underbrace{1}_{h(\underline{x})}$$

where $\underline{t} = \left[\sum_{i=1}^{N} x_i \quad \sum_{i=1}^{N} x_i^2\right]^T$ is sufficient.

If an invertible function is applied to a sufficient statistic, the result is again a sufficient statistic.

For example, in the Gaussian iid model:

- $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is sufficient for θ_1
- $\left[\bar{x} \quad s^2\right]^T$ is sufficient for $[\theta_1, \theta_2]^T$

  ↳ $= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

The Rao-Blackwell Theorem

The power and importance of sufficient statistics is reflected in the following famous result:

Theorem Let $\underline{X}$ be a random variable with pdf/pmf $f_{\underline{\theta}}(\underline{x})$ and let $\underline{T} = \tau(\underline{X})$ be a sufficient statistic. Let $\hat{\underline{\theta}}_1(\underline{x})$ be an estimator of $\underline{\theta}$ and define the mean-squared error

$$MSE(\hat{\underline{\theta}}_1) := E\left[\left\|\hat{\underline{\theta}}_1(\underline{X}) - \underline{\theta}\right\|^2\right].$$

Next define

$$\hat{\underline{\theta}}_2(\underline{x}) = E\left[\hat{\underline{\theta}}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x})\right]$$

Then

$$MSE(\hat{\underline{\theta}}_2) \leq MSE(\hat{\underline{\theta}}_1)$$

with equality iff

$$\hat{\underline{\theta}}_1(\underline{X}) = \hat{\underline{\theta}}_2(\underline{X})$$

with probability one (almost surely).

Observations / Interpretations

1. $\hat{\underline{\theta}}_2$ is a function of the sufficient statistic.

2. Given any estimator $\hat{\underline{\theta}}_1$ that is <u>not</u> a function of a sufficient statistic, there exists a better estimator (with respect to MSE).

3. We may restrict our search for estimators to functions of a sufficient stat.

4. The conditional expectation

$$E\left[\hat{\underline{\theta}}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x})\right]$$

averages out (or removes) non-informative components in $\hat{\underline{\theta}}_1$. We can view the conditional expectation operator as a <u>filter</u> that eliminates unnecessary components of the data.

Proof of Theorem

$$MSE(\underline{\hat{\theta}}_2) = E\left[\|\underline{\hat{\theta}}_2(\underline{X}) - \underline{\theta}\|^2\right]$$

$$= E\left[\|E[\underline{\hat{\theta}}_1(\underline{X}) \mid \underline{T}] - \underline{\theta}\|^2\right]$$

$$= E\left[\|E[\underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta} \mid \underline{T}]\|^2\right]$$

$$\leq E\left[E\left[\|\underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta}\|^2 \mid \underline{T}\right]\right] \quad \boxed{1}$$

$$= E\left[\|\underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta}\|^2\right] \quad \boxed{2}$$

$$= MSE(\underline{\hat{\theta}}_1)$$

1 Consider the random variable

$$\underline{Y} = \underline{Y}(\underline{t}) = \underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta} \mid T(\underline{X}) = \underline{t}$$

By Jensen's inequality,

$$\varphi(E[\underline{Y}]) \leq E[\varphi(\underline{Y})]$$

where φ is the convex function

$$\varphi(\underline{y}) = \|\underline{y}\|^2 = \underline{y}^T\underline{y}.$$

Moreover, since φ is <u>strictly</u> convex, equality holds iff $\underline{Y}$ is a deterministic function of $\underline{t}$ almost surely.

But then

$$\hat{\underline{\theta}}_1(\underline{X}) = \hat{\underline{\theta}}_2(\underline{X})$$

almost surely.

[2] This follows from the "law of total expectation" which states that

$$E_{\underline{u}}[\psi(\underline{u})] = E_{\underline{v}}[E_{\underline{u}|\underline{v}}[\psi(\underline{u})|\underline{v}]]$$

<u>Remark</u>

The result still holds if we replace $\varphi(\underline{y}) = \underline{y}^T\underline{y}$ with any convex function.

However, the condition on equality may need to be modified if φ is not <u>strictly</u> convex (e.g., $\varphi(\underline{y}) = \sum_{i=1}^{p} |y_i|$)

Minimal Sufficient Statistics

If we observe $\underline{x}$, then $\underline{x}$ itself is a sufficient statistic, albeit not a very interesting one. When is a suff. stat. as compressed as it can possibly be?

Definition A sufficient statistic is *minimal* if it is a function of every other sufficient statistic.

Example

For iid Gaussian observations with unknown mean, the following statistics are sufficient:

- $\underline{x} = [x_1, \cdots, x_n]^T$
- $[x_1 + x_3 + \cdots \;,\; x_2 + x_4 + \cdots]^T$
- $[\bar{x} \;,\; s^2]^T$
- $\bar{x}$

However, the first 3 are not minimal because they are not functions of the 4th.

Since $\bar{x}$ is 1-dimensional, it is minimal.

Remark | When we say "is a function of every other sufficient statistic," we exclude functions that increase dimensionality. Otherwise

$$[\bar{x}, \bar{x}]^T$$

would be sufficient in the previous example.

The dimension of a minimal suff. stat. cannot be less than the dimension of $\underline{\theta}$. If we are lucky the dimensions will be equal. Sometimes, the dim. of a minimal suff. stat. is as large as N. See, for example, the Cauchy distribution.

Proving Minimality

Proposition | $\underline{T} = \tau(\underline{X})$ is a minimal suff. stat. if

$$\frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y})} \text{ is independent of } \underline{\theta} \iff \tau(\underline{x}) = \tau(\underline{y})$$

Proof First let's show $\underline{T}$ is a sufficient stat. under the given assumption.

For each $\underline{t}$ in the range of τ, assign a vector $\underline{y}(\underline{t})$ such that $\tau(\underline{y}(\underline{t})) = \underline{t}$.

Then

$$f_{\underline{\theta}}(\underline{x}) = \frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y}(\tau(\underline{x})))} \cdot f_{\underline{\theta}}(\underline{y}(\tau(\underline{x})))$$

$$= h(\underline{x}) \cdot g_{\theta}(\tau(\underline{x}))$$

$\uparrow$ independent of θ since $\tau(\underline{x}) = \tau(\underline{y}(\tau(\underline{x})))$.

To show $\underline{T}$ is minimal, we must show that for any other suff. stat. $\underline{T}' = \tau'(\underline{X})$, $\underline{T}$ is a function of $\underline{T}'$.

So suppose $\underline{T}'$ takes on the value $\underline{t}'$. We must show that $\underline{T}$ is uniquely determined by $\underline{t}'$.

That is, if $\underline{x}$ and $\underline{y}$ are such that $\tau'(\underline{x}) = \tau'(\underline{y}) = \underline{t}'$, then $\tau(\underline{x}) = \tau(\underline{y})$.

So suppose $\underline{x}$ any $\underline{y}$ are such that $\tau'(\underline{x}) = \tau'(\underline{y}) = \underline{t}'$.

Then

$$\frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y})} = \frac{g'_{\underline{\theta}}(\tau'(\underline{x})) \cdot h'(\underline{x})}{g'_{\underline{\theta}}(\tau'(\underline{y})) \cdot h'(\underline{y})} = \frac{h'(\underline{x})}{h'(\underline{y})}$$

which is independent of $\underline{\theta}$. Therefore $\tau(\underline{x}) = \tau(\underline{y})$ ▨

Exercise | Show that $[\Sigma X_i^2, \Sigma X_i]^T$ is a minimal suff. stat. for $[\mu, \sigma^2]^T$, where $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

Solution | $\underline{\theta} = [\mu, \sigma^2]^T$.

$$\frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y})} = \frac{\exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - 2\mu \sum x_i + n\mu^2\right]\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\left[\sum y_i^2 - 2\mu \sum y_i + n\mu^2\right]\right\}}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - \sum y_i^2\right] + \frac{\mu}{\sigma}\left[\sum x_i - \sum y_i\right]\right\}$$

which is independent of $\underline{\theta} \iff \left[\sum x_i^2 \;\; \sum x_i\right]^T = \left[\sum y_i^2 \;\; \sum y_i\right]^T$

$$\iff \tau(\underline{x}) = \tau(\underline{y}).$$

Since $[\bar{X} \;\; S^2]^T$ is a function of $\left[\sum X_i^2 \;\; \sum X_i\right]^T$, it is also minimal. ▨

A second way to show that a suff. stat. is minimal is to show that it is <u>complete</u>.

Complete Sufficient Statistics

Definition A sufficient statistic $T = \tau(\underline{X})$ is **complete** iff for all real-valued functions ϕ

$$\left(E_{\underline{\theta}}[\phi(T)] = 0 \quad \forall \underline{\theta} \right) \Rightarrow \left(P_{\underline{\theta}}[\phi(T) = 0] = 1 \ \forall \underline{\theta} \right)$$

Example Bernoulli trials, part IV

Consider N independent Bernoulli trials

$$X_i \overset{iid}{\sim} \text{Bernoulli}(\theta) \qquad \theta \in [0,1]$$

Recall $K = \sum_{i=1}^{N} X_i$ is sufficient for θ.

Suppose $E_\theta[\phi(K)] = 0 \quad \forall \theta$. But

$$E[\phi(K)] = \sum_{k=0}^{N} \phi(k) \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$$= \text{poly}(\theta)$$

where $\text{poly}(\theta)$ is an N^{th} degree polynomial.

Then $\text{poly}(\theta) = 0 \quad \forall \theta \in [0,1]$

$\Rightarrow$ $\text{poly}(\theta)$ is the zero polynomial

$\Rightarrow \phi(k) = 0 \quad \forall k$

$\Rightarrow K$ is complete

Note The definition of completeness depends on the parameter space Θ. In the last example we had $\Theta = [0,1]$. What would happen if $\Theta = \{\frac{1}{3}, \frac{2}{3}\}$?

Proposition Under very general conditions, if T is a complete S.S., then T is minimal.

Proof Let $T' = \tau'(X)$ be any S.S.

We need to show T is determined completely by T'.

Define

$$\psi(t') := E_\theta[T \mid T' = t'].$$

We will show in particular that $T = \psi(T')$.

Introduce

$$\rho(t) := E_\theta[\psi(T') \mid T = t]$$

Note that

$$\begin{aligned} E_\theta[T] &= E[E[T \mid T']] \quad \leftarrow \text{(law of total expectation)} \\ &= E[\psi(T')] \\ &= E[E[\psi(T') \mid T]] \\ &= E[\rho(T)] \end{aligned}$$

By completeness, we deduce

$$P_{\underline{\theta}}[\underline{T} = \rho(\underline{T})] = 1 \quad \forall \underline{\theta} \tag{1}$$

This implies

$$\psi(\underline{t}') = E[\underline{T} \mid \underline{T}' = \underline{t}'] = E[\rho(\underline{T}) \mid \underline{T}' = \underline{t}'] \tag{2}$$

Now recall the "law of total variance":

$$\mathrm{Var}[Y] = E[\mathrm{Var}[Y \mid Z]] + \mathrm{Var}[E[Y \mid Z]]$$

for scalar random variables Y and Z. Using the subscript "i" to denote the i^{th} component, we have

$$\mathrm{Var}[\rho_i(\underline{T})] = E[\mathrm{Var}[\rho_i(\underline{T}) \mid \underline{T}']] + \mathrm{Var}[E[\rho_i(\underline{T}) \mid \underline{T}']]$$

$$= \text{"} \quad + \mathrm{Var}[\psi_i(\underline{T}')] \quad \text{(by (2))}$$

$$= \text{"} \quad + E[\mathrm{Var}[\psi_i(\underline{T}') \mid \underline{T}]] + \mathrm{Var}[E[\psi_i(\underline{T}') \mid \underline{T}]]$$

$$= \text{"} \quad + \text{"} \quad + \mathrm{Var}[\rho_i(\underline{T})]$$

Since $\mathrm{Var}[\text{anything}] \geq 0$ we deduce

$$\mathrm{Var}[\rho_i(\underline{T}) \mid \underline{T}'] = 0 \quad \text{with prob. 1}$$

$$\mathrm{Var}[\psi_i(\underline{T}') \mid \underline{T}] = 0 \quad \text{with prob. 1}$$

How does this imply the desired result?

$\rho_i(\underline{I})$ is a deterministic function of $\underline{I}'$

In particular (returning to vector notation)

$$\rho(\underline{I}) = E[\rho(\underline{I}) \mid \underline{I}']$$

$$= \psi(\underline{I}')$$

Since $\underline{I} = \rho(\underline{I})$ with prob. 1, we conclude

$$\underline{I} = \rho(\underline{I}) = \psi(\underline{I}') \qquad \text{w.p. 1}$$

as was to be shown. ▨

The "general conditions" under which the result is valid are that the various means and variances used throughout the proof are well-defined and finite.

The Exponential Family

In general, sufficient statistics, especially ones that are minimal and complete, can be difficult to find (if they even exist).

For a special family of distributions, however, we can immediately identify a complete and minimal suff. stat.

Definition We say the distribution of $\underline{X}$ belongs to the exponential family of distributions if its pdf/pmf can be written

$$f_{\underline{\theta}}(\underline{x}) = a(\underline{\theta})\, b(\underline{x}) \exp\left\{ c(\underline{\theta})^T \tau(\underline{x}) \right\}$$

for some a, b, c and τ, where the dimension p of $\underline{\theta}$ is also the dimension of $c(\underline{\theta})$ and $\tau(\underline{x})$.

Note My definition is slightly different from that given in Prof. Hero's notes.

- c instead of $-c$
- defined for vectors, not just scalars

Example | Bernoulli trials, part IV

$$p_\theta(\underline{x}) = \theta^k (1-\theta)^{N-k} \qquad \left[k = \sum_{i=1}^{N} x_i\right]$$

$$= \exp\left\{ \log\left(\theta^k (1-\theta)^{N-k} \right) \right\}$$

$$= \exp\left\{ k \log\theta + (N-k)\log(1-\theta) \right\}$$

$$= \underbrace{\exp\left\{ N \log(1-\theta) \right\}}_{a(\theta)} \cdot \underbrace{\exp\left\{ [\log\theta - \log(1-\theta)]}_{c(\theta)} \cdot \underbrace{k \right\}}_{\tau(x)}$$

$$[b(\underline{x}) = 1]$$

Many common distributions belong to the exponential family, including Gaussian w/ unknown mean and/or variance, Poisson, exponential, gamma, binomial, and multinomial.

Exercise Suppose $\underline{X} = [X_1, \ldots, X_N]^T$ where

$$X_i \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$$

Recall that if $X \sim \text{Gamma}(\alpha, \beta)$, then

$$f_{\underline{\theta}}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\beta x}, \quad x \geq 0$$

where $\underline{\theta} = [\alpha \ \beta]^T$. Show that the distribution of $\underline{X}$ belongs to the exponential family.

Solution

$$f_{\underline{\theta}}(\underline{x}) = \prod_{i=1}^{N} f_{\underline{\theta}}(x_i)$$

$$= \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{N} \cdot \prod_{i=1}^{N} x_i^{\alpha-1} \exp\left\{-\beta \sum_{i=1}^{N} x_i\right\}$$

$$= \left(\frac{\beta}{\Gamma(\alpha)}\right)^{N} \exp\left\{\log\left[\left(\prod x_i\right)^{\alpha-1}\right] - \beta \sum x_i\right\}$$

$$= \left(\frac{\beta}{\Gamma(\alpha)}\right)^{N} \exp\left\{(\alpha-1) \sum_{i=1}^{N} \log x_i - \beta \sum x_i\right\}$$

$$\Rightarrow = \underbrace{\left(\frac{\beta}{\Gamma(\alpha)}\right)^{N}}_{a(\underline{\theta})} \underbrace{\left(\prod_{i=1}^{N} x_i\right)^{-1}}_{b(\underline{x})} \exp\left\{\underbrace{[\alpha \;\; -\beta]}_{c(\underline{\theta})^T} \underbrace{\begin{bmatrix} \sum \log x_i \\ \sum x_i \end{bmatrix}}_{\tau(\underline{x})}\right\}$$

[Note: representation is not unique.]

Proposition If the distribution of $\underline{X}$ belongs to the exponential family, then $\underline{T} = \tau(\underline{X})$ is a sufficient statistic for $\underline{\theta}$.

Proof | $\underline{T}$ is sufficient for $\underline{\theta}$ by the FNFT:

$$f_{\underline{\theta}}(\underline{x}) = a(\underline{\theta})\, b(\underline{x}) \exp\left\{ c(\underline{\theta})^T \tau(\underline{x}) \right\}$$

$$= \underbrace{a(\underline{\theta}) \exp\left\{ c(\underline{\theta})^T \tau(\underline{x}) \right\}}_{g_\theta(\tau(\underline{x}))} \cdot \underbrace{b(\underline{x})}_{h(\underline{x})}$$

Proposition | Under certain "reasonable" conditions, $\underline{T} = \tau(\underline{X})$ is a complete and minimal sufficient statistic for $\underline{\theta}$.

Sketch of proof | Before we argued that the pdf/pmf of $\underline{X}$ depends on $\underline{\theta}$ only through $f_{\underline{\theta}}(\underline{t})$.

Thus

$$f_{\underline{\theta}}(\underline{t}) \propto \exp\left\{ c(\underline{\theta})^T \underline{t} \right\}$$

Suppose ϕ is a real-valued function such that

$$E_{\underline{\theta}}\left\{ \phi(\underline{T}) \right\} = 0 \quad \forall \theta$$

We must show $P_\theta\left\{ \phi(\underline{T}) = 0 \right\} = 1 \quad \forall \theta$.

For each $\underline{\theta}$ we can write

$$0 = E\{\phi(\underline{T})\}$$

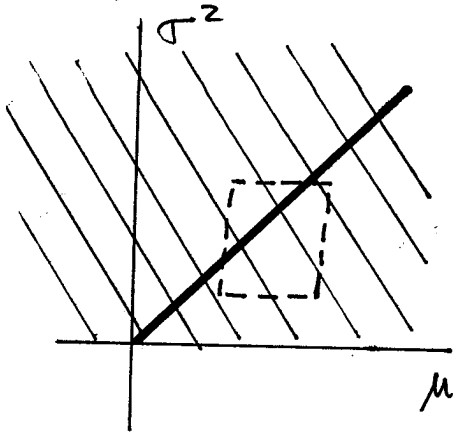
$$= \int \phi(\underline{t}) f_{\underline{\theta}}(\underline{t}) \, d\underline{t}$$

$$\propto \int \phi(\underline{t}) \exp\{c(\underline{\theta})^T \underline{t}\} \, d\underline{t}$$

which is the Laplace transform of ϕ at $c(\underline{\theta})$. Inverting the Laplace transform we find $\phi \equiv 0$. ∎

The "reasonable" conditions under which the above arguments hold are needed to ensure the uniqueness (invertibility) of the Laplace transform:

- The parameter space Θ must contain an open rectangle: think back to Bernoulli example. As another example, in Gaussian case with $\mu = \sigma^2$, $\underline{T}$ is not complete.



- The image of $c(\underline{\theta})$, $\underline{\theta} \in \Theta$, should have full dimensionality.

Note that minimality of $\underline{T} = \tau(\underline{X})$ for the exponential model follows from completeness by a previous result.

How could we show minimality directly? What condition would need to be satisfied?

Summary

- Sufficient statistic $\underline{T}$ for $\underline{\theta}$: contains all the information about $\underline{\theta}$ present in $\underline{X}$.
- Fisher-Neyman Factorization Theorem: $\underline{T}$ is sufficient

$$\iff f_{\underline{\theta}}(\underline{x}) = g_{\underline{\theta}}(\tau(\underline{x})) \cdot h(\underline{x})$$

- Rao-Blackwell Theorem: Can improve any estimator by conditioning on a suff. stat.
- Minimal suff. stat: function of any other suff. stat. Intuitively, maximal compression of information.
- Completeness: technical condition that ensures minimality. Will also be important in our study of minimum variance unbiased estimators.
- Exponential family: Broad class for which a complete and minimal suff. stat. is easily identified.

Key

a. $\begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \\ 0 & \text{else} \end{cases}$

b. $\exp\left\{\frac{-1}{2\sigma^2}\left[-2\theta\sum_{i=1}^{N} x_i + N\theta^2\right]\right\}$ $\underbrace{\qquad\qquad}_{g_\theta(t)}$ $(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{\frac{-\sum x_i^2}{2\sigma^2}\right\}$ $\underbrace{\qquad\qquad}_{h(\underline{x})}$

$t = \sum_{i=1}^{N} x_i$

ESTIMATION THEORY

What is Estimation?

The main actors in the story of estimation theory are

$\mathcal{X} \subseteq \mathbb{R}^N$	the sample space
$\underline{X} \in \mathcal{X}$	a random vector of measurements
$\Theta \subseteq \mathbb{R}^p$	the parameter space
$\underline{\theta} \in \Theta$	the unknown parameter
$f_{\underline{\theta}}(\underline{x})$	the pdf/pmf of $\underline{X}$

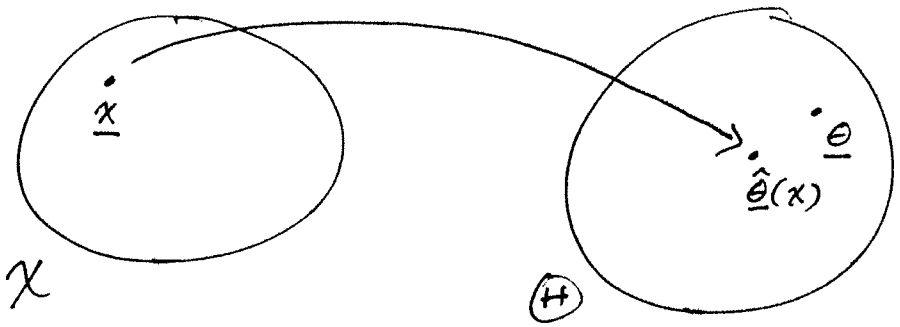
The basic plot is a follows:

A realization $\underline{x} \in \mathcal{X}$ of $\underline{X}$ is observed. The distribution of $\underline{X}$ is specified entirely by the unknown parameter $\underline{\theta}$. $\underline{x}$ must be used to estimate $\underline{\theta}$.

The protagonist of our story is a function

$$\underline{\hat{\theta}} : \mathcal{X} \longrightarrow \textcircled{H}$$

called an <u>estimator</u>. Estimation theory studies various kinds of estimators and their properties.



The notation $\underline{\hat{\theta}}$ will be used in several ways

- $\underline{\hat{\theta}}(\underline{X})$ is a random variable
- $\underline{\hat{\theta}}$ may also denote the same random variable
- for a particular $\underline{x}$, $\underline{\hat{\theta}}(\underline{x})$ is the <u>estimate</u> of $\underline{\theta}$

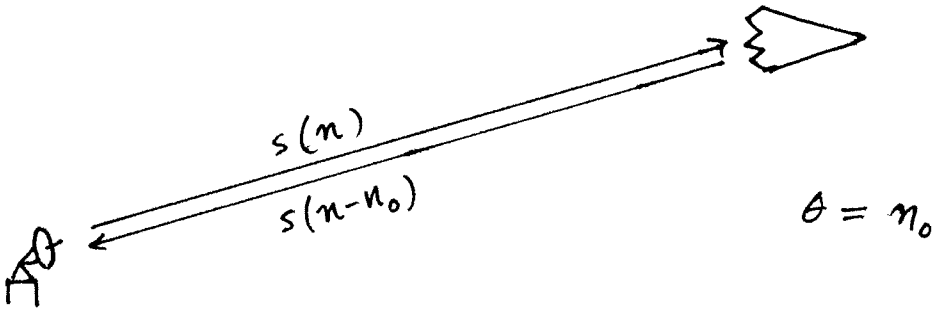
The meaning will be clear from the context.

The methods we will study will be quite general, but we will emphasize applications in signal processing.

The methods to be covered might also be covered in a statistics course on "parameter estimation" or "point estimation."

Examples

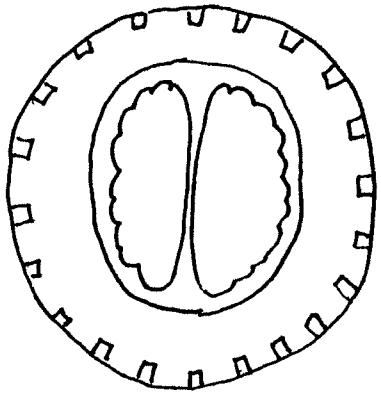
1. signal delay estimation

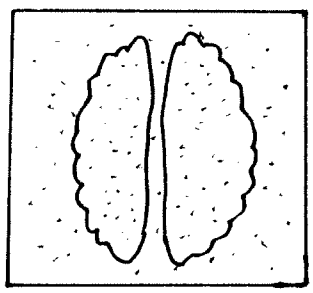


$$x(n) = s(n-n_0) + w(n), \quad n = 0, 1, \ldots, N-1$$

2. Image reconstruction

$\underline{\theta}$ = image of brain





radial projections $\Longrightarrow$ fMRI brain image

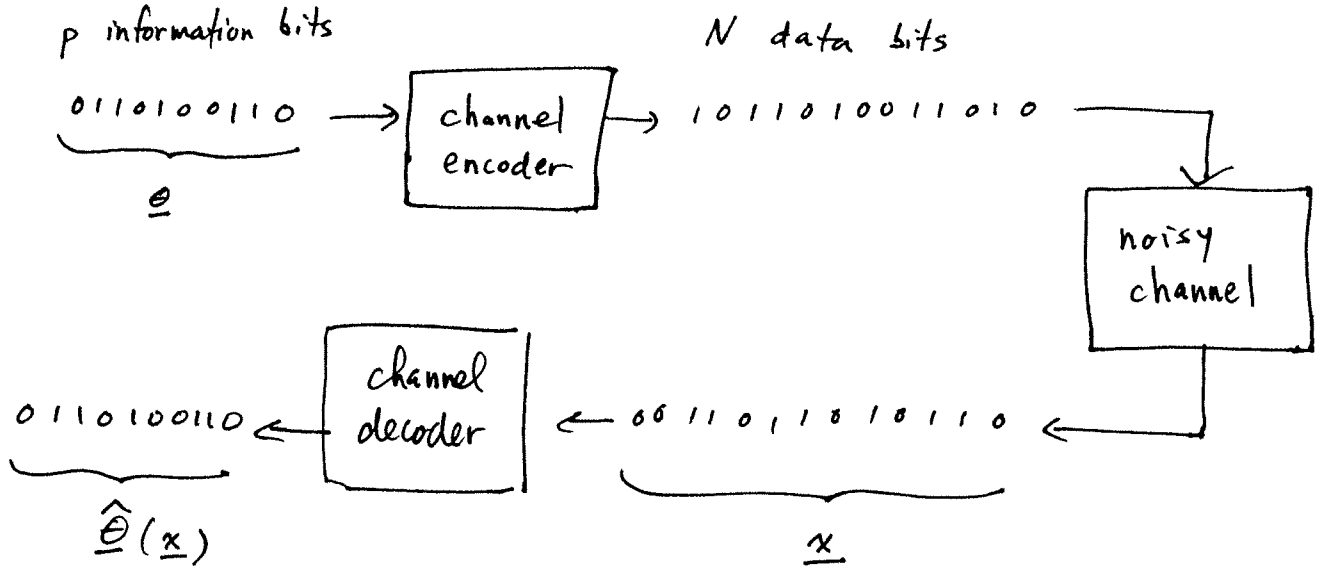
3. Signal / Image Denoising

old record with static $\Rightarrow$ clean recording with crackles and pops removed

$$x(n) = s(n) + w(n) \Rightarrow \hat{s}(n)$$

$$\underline{\theta} = [s(0) \cdots s(N-1)]^T.$$

4. Communication / Transmission over noisy channels



5. Many, many others

Estimation Categories

There are several ways to classify estimators and estimation strategies. Here are some of the main ones.

I. Optimality Criterion

The primary element in an optimality criterion is whether the unknown parameter is viewed as random or nonrandom.

A θ nonrandom : classical (frequentist) estimation

- minimum variance unbiased estimation
- maximum likelihood estimation
- method of moments
- least squares (nonstatistical)

B θ random : Bayesian estimation

- minimum mean squared error
- minimum absolute deviation
- maximum a posteriori

Depending on the problem, sometimes different criteria are optimized by the same estimator.

II. Form

The primary distinction regarding the form of an estimator is whether it is linear or nonlinear.

A linear estimator has the form

$$\hat{\underline{\theta}}(\underline{x}) = \underline{c}^T \underline{x}.$$

Such estimators arise frequently in conjunction with the multivariate Gaussian distribution.

Because of their simplicity, sometimes we will optimize our criteria while restricting the estimator to be linear.

III. Offline vs. Online

When we study filtering we will focus on estimators that can efficiently update their estimate as data streams in.

Classical Estimation: Basic Notions

Our study of estimation will begin with classical estimators. Thus, assume $\underline{\theta}$ is nonrandom, that is, unknown but fixed.

Definitions Let $\hat{\underline{\theta}}$ be an estimator of $\underline{\theta}$.

The mean squared error of $\hat{\underline{\theta}}$ is

$$\text{MSE}(\hat{\underline{\theta}}) = E\left[\|\hat{\underline{\theta}} - \underline{\theta}\|^2\right]$$

$$= E\left[\|\hat{\underline{\theta}}(\underline{x}) - \underline{\theta}\|^2\right]$$

The variance of $\hat{\underline{\theta}}$ is

$$\text{Var}(\hat{\underline{\theta}}) = E\left[\|\hat{\underline{\theta}} - E[\hat{\underline{\theta}}]\|^2\right]$$

The bias of $\hat{\underline{\theta}}$ is

$$\text{Bias}(\hat{\underline{\theta}}) = E[\hat{\underline{\theta}}] - \underline{\theta}$$

We say $\hat{\underline{\theta}}$ is unbiased if

$$\text{Bias}(\hat{\underline{\theta}}) = \underline{0} \qquad \forall \theta \in (H)$$

Otherwise, we say $\hat{\underline{\theta}}$ is biased.

Let $\{\underline{\hat{\theta}}_N\}_{N=1}^{\infty}$ be a family of estimators.

We say $\{\underline{\hat{\theta}}_N\}$ is <u>asymptotically unbiased</u> if

$$\text{Bias}(\underline{\hat{\theta}}_N) \to \underline{0} \quad \text{as} \quad N \to \infty \qquad \forall \underline{\theta} \in \Theta$$

and <u>consistent</u> if

$$\text{MSE}(\underline{\hat{\theta}}_N) \to 0 \quad \text{as} \quad N \to \infty \qquad \forall \theta \in \Theta$$

<u>Example</u> Suppose $\underline{X} = [X_1, \ldots, X_N]^T$ where

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2), \qquad i = 1, \ldots, N$$

Consider the estimator of μ given by

$$\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

What is the bias of $\hat{\mu}$?

$$E\{\hat{\mu}\} = E\left\{\frac{1}{N}\sum_{i=1}^{N} X_i\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\{X_i\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mu$$

$$= \mu$$

$\Rightarrow$ $\hat{\mu}$ is unbiased.

Exercise | Find the variance of $\hat{\mu}$. Is $\hat{\mu}$ consistent?

Solution 1

Approach 1 :

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= E\left\{(\hat{\mu}-\mu)^2\right\} \\
&= E\left\{\left(\frac{1}{N}\sum X_i - \mu\right)^2\right\} \\
&= E\left\{\sum_{i=1}^{N}\left(\frac{X_i-\mu}{N}\right)^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N} E\left\{(X_i-\mu)^2\right\} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

Approach 2

Since $\underline{X} \sim N(\underline{1}\mu, \sigma^2\underline{I})$ and $\bar{X} = A\cdot\underline{X}$,

$A = \left[\frac{1}{N} \cdots \frac{1}{N}\right]$, we deduce

$$\begin{aligned}
\hat{\mu} &\sim N\left(A\cdot\underline{1}\mu,\ A\cdot\sigma^2\underline{I}\cdot A^T\right) \\
&\sim N\left(\mu, \frac{\sigma^2}{N}\right)
\end{aligned}$$

As for consistency, note

$$\begin{aligned} MSE(\hat{\mu}) &= E\{(\hat{\mu}-\mu)^2\} \\ &= E\{(\hat{\mu}-E\hat{\mu})^2\} \\ &= Var(\hat{\mu}) \\ &= \frac{\sigma^2}{N} \longrightarrow 0 \quad \text{as} \quad N \to \infty. \end{aligned}$$

Example | Note that for an estimator to be unbiased, its expected value must be the true value for all $\underline{\theta} \in (H)$.

In the previous example, suppose we take

$$\hat{\mu} = \frac{2}{N}\sum_{i=1}^{N} X_i.$$

Then

$$E\{\hat{\mu}\} = \mu \quad \text{if} \quad \mu = 0$$

$$E\{\hat{\mu}\} \neq \mu \quad \text{if} \quad \mu \neq 0.$$

$\Rightarrow \hat{\mu}$ is biased.

MINIMUM VARIANCE UNBIASED ESTIMATION

Bias - Variance Tradeff

The MSE of an estimator $\underline{\hat{\theta}}$ can be broken down into two components, the bias of $\underline{\hat{\theta}}$ and the variance of $\underline{\hat{\theta}}$.

In particular, the following bias-variance decomposition of the MSE holds:

$$MSE_{\underline{\theta}}(\underline{\hat{\theta}}) = \left\| Bias_{\underline{\theta}}(\underline{\hat{\theta}}) \right\|^2 + Var_{\underline{\theta}}(\underline{\hat{\theta}}).$$

Let's prove this for the case of a scalar parameter (the vector case is left as an exercise).

$$MSE_\theta(\hat{\theta}) = E\{(\hat{\theta}-\theta)^2\}$$

$$= E\{(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2\}$$

$$= E\{(\hat{\theta}-E\hat{\theta})^2\} + 2E\{(\hat{\theta}-E\hat{\theta})\cdot(E\hat{\theta}-\theta)\}$$

$$+ E\{(E\hat{\theta}-\theta)^2\}$$

The middle term is

$$E\{(\hat{\theta}-E\hat{\theta})(E\hat{\theta}-\theta)\}$$

(a) $=$

This leaves

$$MSE_\theta(\hat{\theta}) = E\{(\hat{\theta}-E\hat{\theta})^2\} + E\{(E\hat{\theta}-\theta)^2\}$$

$$= E\{(\hat{\theta}-E\hat{\theta})^2\} + (E\hat{\theta}-\theta)^2$$

$$= Var_\theta(\hat{\theta}) + Bias_\theta^2(\hat{\theta})$$

▨

When designing an estimator, you can typically trade off bias and variance. Decreasing the bias of your estimator will increase the variance, while increasing the bias will decrease the variance.

Example Suppose $\underline{x} = [x_1, \ldots, x_N]^T$ where

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2), \qquad i = 1, \ldots, N.$$

Consider the class of estimators for μ

$$\hat{\mu}_\alpha = \frac{\alpha}{N} \sum_{i=1}^{N} X_i, \qquad \alpha \in \mathbb{R}.$$

Let's see how α affects the bias-variance tradeoff.

Note that

$$\hat{\mu}_\alpha = \alpha \cdot \overline{X}$$

Recall

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

Thus

$$\mathrm{Var}_\mu(\hat{\mu}_\alpha) = \mathrm{Var}(\alpha \bar{X})$$

(b)

$$=$$

The bias of $\hat{\mu}_\alpha$ is

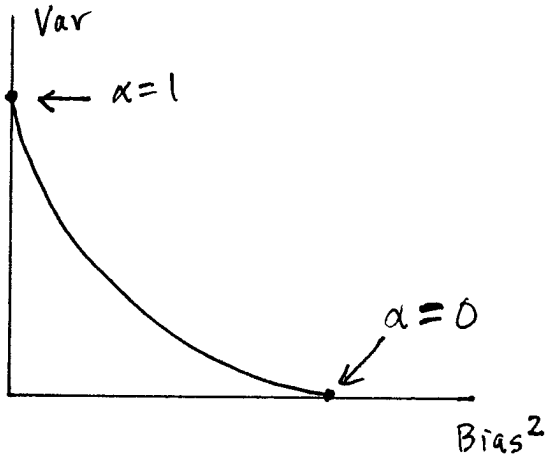
$$\mathrm{Bias}_\mu(\hat{\mu}_\alpha) = E\hat{\mu}_\alpha - \mu$$

$$= E\{\alpha \bar{X}\} - \mu$$

$$=$$

Therefore the MSE is

$$\mathrm{MSE}_\mu(\hat{\mu}_\alpha) = (\alpha-1)^2\mu^2 + \frac{\alpha^2\sigma^2}{N}$$



Minimum MSE?

How practical is the MSE as a *design* criterion?

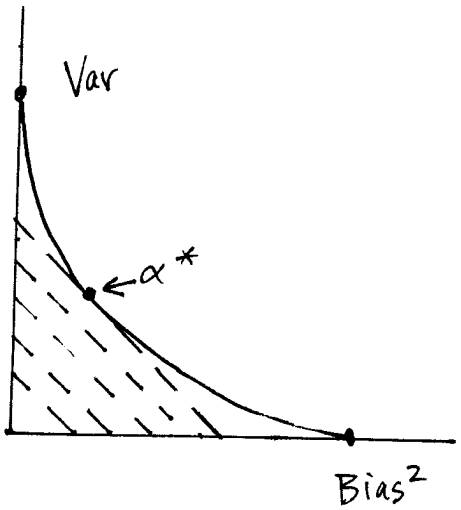
In the previous example, the MSE is minimized when

$$\frac{\partial \, MSE_{\mu}(\hat{\mu}_{\alpha})}{\partial \alpha} = 2(\alpha-1)\mu^2 + 2\alpha\frac{\sigma^2}{N} = 0$$

$$\Rightarrow \alpha^* = \frac{\mu^2}{\mu^2 + \frac{\sigma^2}{N}}.$$

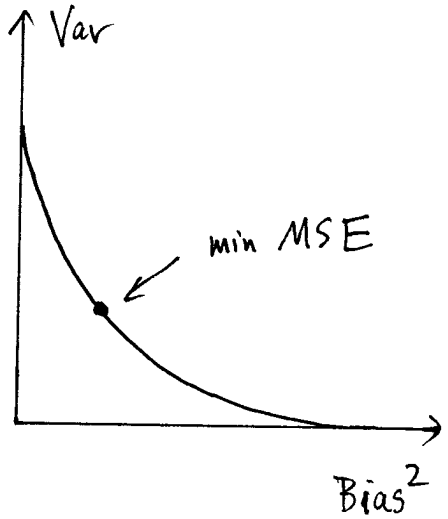
Unfortunately the solution depends on the unknown parameter μ. Therefore the estimator is not *realizable*.

This phenomenon occurs for many classes of problems, and therefore we need an alternative to direct MSE minimization.



Minimum Variance Unbiased Estimation

In general the minimum MSE estimator has nonzero bias and variance.



However, in many situations only the bias depends on the unknown parameter.

In our example, recall we had

$$\text{Bias}_\mu(\hat{\mu}_\alpha) = (\alpha - 1)\mu$$

$$\text{Var}_\mu(\hat{\mu}_\alpha) = \frac{\alpha^2 \sigma^2}{N}.$$

This suggests the following alternative:

- constrain the estimator to be unbiased and minimize the variance
- equivalently, minimize the MSE among all unbiased estimators.

Definition $\hat{\underline{\theta}}$ is said to be a (uniform) minimum variance unbiased estimator (MVUE) for $\underline{\theta}$ if

1. $E\,\hat{\underline{\theta}} = \underline{\theta} \qquad \forall\, \underline{\theta} \in \Theta$
2. If $E\,\hat{\underline{\theta}}' = \underline{\theta} \qquad \forall\, \underline{\theta} \in \Theta$, then
$$Var_{\underline{\theta}}(\hat{\underline{\theta}}) \leq Var_{\underline{\theta}}(\hat{\underline{\theta}}') \qquad \forall\, \underline{\theta} \in \Theta$$

Remark Notice that the MVUE criterion requires an estimator to be optimal for all values of $\underline{\theta}$.

This highlights another drawback of MSE minimization: it is absurd to require an estimator to have minimal MSE for all $\underline{\theta}$.

For example, the estimator $\hat{\theta} = 28$ can't be beat when $\theta = 28$, but it is terrible elsewhere.

By restricting the class of possible estimators, estimators with uniformly best MSE become possible.

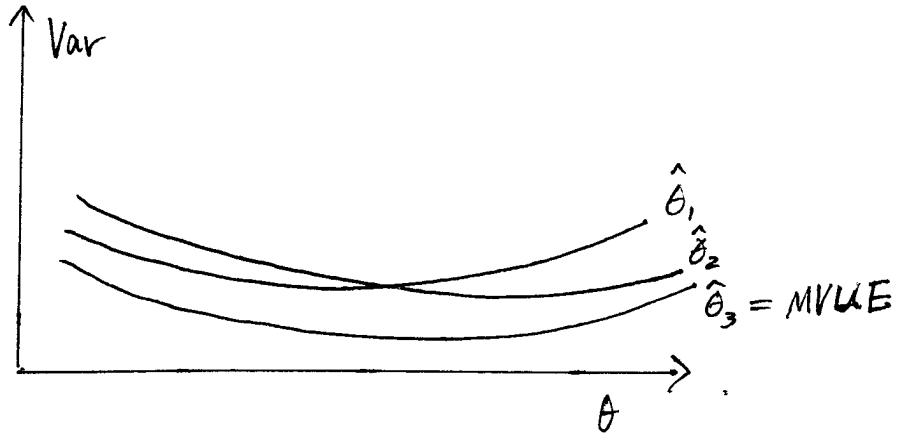
Existence of MVUE

Despite all of this talk, the MVUE does not always exist.

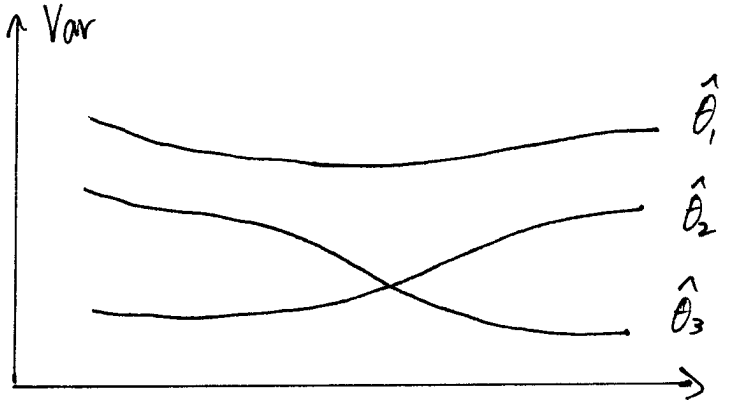
Suppose there are three unbiased estimators, $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$.

There are two possibilities

(1) One estimator has uniformly smaller variance



(2) No estimator has uniformly smaller variance



No MVUE exists!

In fact, sometimes there may not exist even a single unbiased estimator!

Exercise | Suppose we observe a single scalar realization x of

$$X \sim \text{unif}\left(0, \frac{1}{\theta}\right), \quad \theta > 0.$$

Show that an unbiased estimator of θ does not exist.

Solution / The density of X is

$$f_\theta(x) = \theta \cdot I_{[0, \frac{1}{\theta}]}(x)$$

If $\hat{\theta}$ is unbiased then

$$\forall \theta > 0, \quad \theta = E\{\hat{\theta}\}$$

$$= \int_0^{\frac{1}{\theta}} \hat{\theta}(x) \cdot \theta \, dx$$

$$\Rightarrow \int_0^{\frac{1}{\theta}} \hat{\theta}(x) \, dx = 1 \qquad \forall \theta > 0.$$

$$\Rightarrow \hat{\theta}\left(\frac{1}{\theta}\right) = 0 \qquad \forall \theta > 0 \quad (\text{FTC}), \text{ a contradiction}$$

Finding the MVUE

Unfortunately there is no systematic procedure. We will discuss three potential ways of finding the MVUE

- Calculate the CRLB and see if some estimator achieves the bound
- Apply the Rao-Blackwell theorem with a complete sufficient statistic
- Further restrict the class of possible estimators to be linear.

Summary

- MSE = Bias^2 + Variance
- Minimizing MSE often requires knowledge of θ
- MVUE := uniformly best MSE/variance among all unbiased est.
  - may not always exist

Key

a. $= (E\hat{\theta} - \theta)\, E\{(\hat{\theta} - E\hat{\theta})\}$

$= (E\hat{\theta} - \theta) \cdot (E\hat{\theta} - E\hat{\theta}) = 0$

b. $= \alpha^2 \text{Var}(\bar{X})$

$= \frac{\alpha^2 \sigma^2}{N}$

c. $= \alpha\mu - \mu$

$= (\alpha - 1)\mu$

THE CRAMER-RAO LOWER BOUND

The CRLB is a lower bound on the variance of any unbiased estimator of a parameter $\underline{\theta}$.

It is useful in many ways:

1. If $\hat{\underline{\theta}}$ achieves the CRLB for all $\underline{\theta} \in \Theta$, the $\hat{\underline{\theta}}$ is a MVUE.

2. The CRLB provides a benchmark against which we can compare the performance of any unbiased estimator. We're doing well if our estimator is "close" to the CRLB

3. The CRLB allows us to rule out impossible estimators. We know it is impossible to find an estimator that beats the CRLB. This is useful in feasibility studies.

4. The theory behind the CRLB tells us precisely when the bound is achievable.

CRLB: Scalar Parameter

Theorem Consider $\underline{X} \sim f_\theta(\underline{x}) = f(\underline{x}; \theta)$ where θ is fixed but unknown. Assume $f(x; \theta)$ satisfies

$$E\left\{ \frac{\partial \log f(\underline{X}; \theta)}{\partial \theta} \right\} = 0$$

where the expectation is with respect to $f(\underline{x}; \theta)$.

Then the variance of any unbiased estimator $\hat{\theta}$ satisfies

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the **Fisher information**

$$I(\theta) := E\left\{ \left(\frac{\partial \log f(\underline{X}; \theta)}{\partial \theta} \right)^2 \right\}$$

Here $I(\theta)$ is evaluated at the true value of the unknown parameter, and the expectation is w.r.t $f(\underline{x}; \theta)$

Furthermore, the bound holds with equality iff

$$\frac{\partial \log f(\underline{x}; \theta)}{\partial \theta} = I(\theta) \cdot (\hat{\theta}(x) - \theta) \qquad \forall x \in \mathcal{X}$$

In this case we say $\hat{\theta}$ is **efficient**.

Remarks 1. $\underline{X}$ can be continuous or discrete; we only need differentiability w.r.t. θ.

2. When viewed as a function of θ, $f(\underline{x};\theta)$ is called the *likelihood* of θ, and $\log f(\underline{x};\theta)$ the *log-likelihood*.

3. The function

$$\frac{\partial \log f(\underline{x};\theta)}{\partial \theta}$$

is called the *score function*.

4. The condition

$$E\left\{\frac{\partial \log f(\underline{X};\theta)}{\partial \theta}\right\} = 0$$

is called a *regularity condition*.

5. Using integration by parts, the Fisher information can be rewritten

(a) $$I(\theta) =$$

6. An efficient estimator does not always exist.

7. If

$$\frac{\partial \log f(\underline{x};\theta)}{\partial \theta} = I(\theta)\left(\hat{\theta}(\underline{x}) - \theta\right) \qquad \forall\theta\ \forall\underline{x}$$

then $\hat{\theta}$ is a MVUE.

Example Suppose $\underline{x} = [x_1, \ldots, x_N]^T$ where

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, N$$

with $\theta = \mu$ (assume σ^2 is known)

Let's compute the CRLB for θ.

First, we need to check the condition

$$E\left\{ \frac{\partial \log f(\underline{x}; \theta)}{\partial \theta} \right\} = 0$$

- $\log f(\underline{x}; \mu) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$
- $\frac{\partial \log f(\underline{x}; \mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum (x_i - \mu)$
- $E\left\{ \frac{1}{\sigma^2} \sum (x_i - \mu) \right\} = \frac{1}{\sigma^2} \sum (\underbrace{E x_i}_{\mu} - \mu) = 0$

Now let's compute the CRLB

- $-\frac{\partial^2}{\partial \mu^2} \log f(\underline{x}; \mu) = -\frac{\partial}{\partial \mu} \left\{ \frac{1}{\sigma^2} \sum (x_i - \mu) \right\} = \frac{N}{\sigma^2}$

$$\Rightarrow I(\mu) = E\left\{ \frac{N}{\sigma^2} \right\} = \frac{N}{\sigma^2}.$$

$\Rightarrow$ If $\hat{\mu}$ is any unbiased estimator of μ, then $\mathrm{Var}\{\hat{\mu}\} \geq 1/I(\mu) = \frac{\sigma^2}{N}$. ▨

Recall the sample mean, $\hat{\mu} = \bar{X} = \frac{1}{N}\sum X_i$. Previously we saw $E\hat{\mu} = \mu$ and $Var\{\hat{\mu}\} = \frac{\sigma^2}{N}$

$\Rightarrow$ $\bar{X}$ is efficient and a MVUE.

Regularity Condition

$$E\left\{ \frac{\partial \log f(\underline{X};\theta)}{\partial\theta} \right\} = \int \frac{\partial \log f(\underline{x};\theta)}{\partial\theta} \cdot f(\underline{x};\theta)\, d\underline{x}$$

$$= \int \frac{\left(\frac{\partial f(\underline{x};\theta)}{\partial\theta}\right)}{f(\underline{x};\theta)} \cdot f(\underline{x};\theta)\, d\underline{x}$$

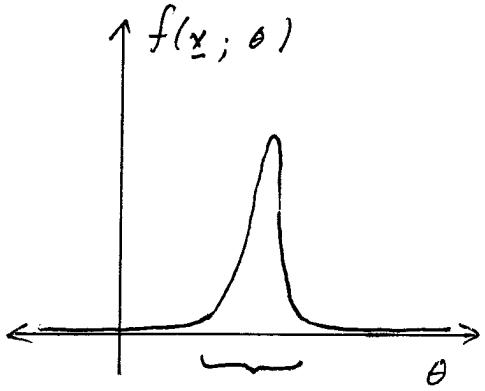
$$= \int \frac{\partial f(\underline{x};\theta)}{\partial\theta}\, d\underline{x}$$

$$= \frac{\partial}{\partial\theta} \int f(\underline{x};\theta)\, d\underline{x}$$

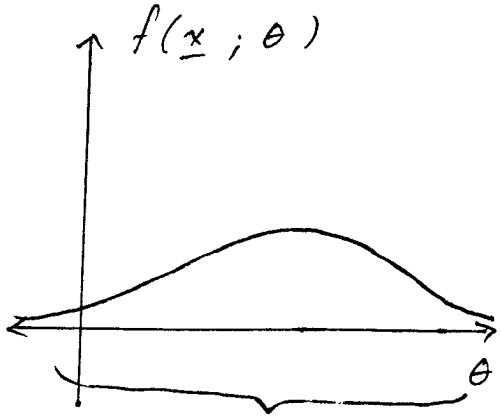
$$= \frac{\partial}{\partial\theta}\{1\} = 0$$

So the regularity condition holds provided we can interchange $\frac{\partial}{\partial\theta}$ and $\int$ (or Σ for discrete $\underline{X}$). This is true for many distributions. A case where it is not true is when the <u>support</u> of $\underline{X}$ depends on θ. For example, $X \sim unif(0,\theta)$.

Fisher Information and Average Curvature

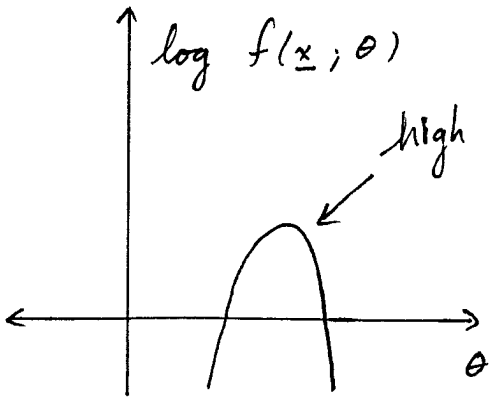


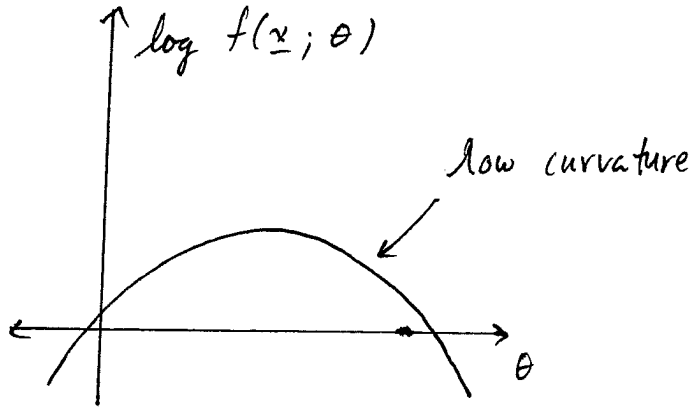
highly likely that θ is in this range



harder to pinpoint θ

The operator $-\frac{\partial^2}{\partial\theta^2}$ measures curvature





So $I(\theta)$ reflects the average curvature of the log-likelihood $\log f(\underline{x};\theta)$

Conclusion: θ is easy to estimate

$\iff$ $f(\underline{x};\theta)$ is "peaky" near θ (on average)

$\iff$ $\log f(\underline{x};\theta)$ has high curvature at θ (on average)

$\iff$ $I(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\log f(\underline{x};\theta)\right\}$ is large

$\iff$ CRLB is small

Exercise | Consider the correlated bivariate Gaussian

$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

where ρ is known. Find the CRLB. Hint:

$$\log f(\underline{x}; \mu) = \frac{-1}{1+\rho} \left(\mu^2 - \mu(x_1 + x_2) \right) + C$$

Does an efficient estimator exist?

Solution | Let's check the regularity condition first.

$$\frac{\partial \log f(\underline{x};\mu)}{\partial \mu} = -\frac{(2\mu - (x_1 + x_2))}{1+\rho}$$

$$E\left\{-\frac{(2\mu - (X_1+X_2))}{(1+\rho)}\right\} = -\frac{(2\mu - (\mu+\mu))}{1+\rho} = 0$$

Because $EX_1 = EY_2 = \mu$. Now

$$-\frac{\partial^2}{\partial \mu^2} \log f(\underline{x};\mu) = \frac{2}{1+\rho}$$

$$\Rightarrow I(\mu) = E\left\{\frac{2}{1+\rho}\right\} = \frac{2}{1+\rho}.$$

$\Rightarrow$ For any unbiased $\hat{\mu}$, $\mathrm{Var}(\hat{\mu}) \geq \frac{1+\rho}{2} = \mathrm{CRLB}$.

Notice that

$$\frac{\partial \log f(\underline{x};\mu)}{\partial \mu} = \frac{2}{1+\rho}\left(\frac{x_1+x_2}{2} - \mu\right)$$

$$= I(\theta)\cdot(\hat{\theta}(\underline{x}) - \theta)$$

$\Rightarrow \hat{\mu} = \frac{x_1+x_2}{2}$ is efficient for all μ, and hence an MVUE.

Proof of CRLB: Scalar Case

Let $\hat{\theta}$ be an unbiased estimator.

Consider the random variables

$$Y = \hat{\theta}(\underline{X}) - \theta$$

$$Z = \frac{\partial \log f(\underline{X}; \theta)}{\partial \theta}$$

Note that both have zero mean.

We will show

$$E\{Y \cdot Z\} = 1.$$

The CRLB then follows by application of the Cauchy-Schwarz inequality:

$$1 = E\{Y \cdot Z\} = (E\{Y \cdot Z\})^2$$

$$\leq E\{Y^2\} \cdot E\{Z^2\}$$

$$= E\{(\hat{\theta}(\underline{X}) - \theta)^2\} \cdot E\left\{\left(\frac{\partial \log f(\underline{X}; \theta)}{\partial \theta}\right)^2\right\}$$

$$= \mathrm{Var}_\theta(\hat{\theta}) \cdot I(\theta)$$

$$\Rightarrow \mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

We now show $E\{YZ\} = 1$. Since $\hat{\theta}$ is unbiased

(b) $\theta =$

Differentiating both sides w.r.t. θ

$$1 = \frac{\partial}{\partial \theta} \int \hat{\theta}(\underline{x}) f(\underline{x}; \theta) \, d\underline{x}$$

$$= \int \hat{\theta}(\underline{x}) \frac{\partial}{\partial \theta} f(\underline{x}; \theta) \, d\underline{x} \qquad (1)$$

$$= \int \hat{\theta}(\underline{x}) \frac{\left(\frac{\partial f(\underline{x}; \theta)}{\partial \theta}\right)}{f(\underline{x}; \theta)} f(\underline{x}; \theta) \, dx$$

$$= \int \hat{\theta}(\underline{x}) \frac{\partial \log f(\underline{x}; \theta)}{\partial \theta} f(\underline{x}; \theta) \, d\underline{x}$$

$$= \int (\hat{\theta}(\underline{x}) - \theta) \frac{\partial \log f(\underline{x}; \theta)}{\partial \theta} f(\underline{x}; \theta) \, d\underline{x} \qquad (2)$$

$$= E\{Y \cdot Z\}.$$

Remarks

- Technically, (1) requires an additional assumption on it being valid to exchange $\frac{\partial}{\partial \theta}$ and $\int$ here
- (2) follows from the regularity condition

Equality holds in the Cauchy-Schwarz inequality iff $\exists\, k(\theta)$ (a constant not depending on $\underline{x}$) such that

$$\frac{\partial \log f(\underline{x};\theta)}{\partial \theta} = k(\theta)(\hat{\theta}(\underline{x}) - \theta) \qquad \forall \underline{x} \in \mathcal{X}$$

Taking the derivative w.r.t θ of both sides

$$\frac{\partial^2}{\partial \theta^2} \log f(\underline{x};\theta) = -k(\theta) + k'(\theta)\cdot(\hat{\theta}(\underline{x}) - \theta),$$

and taking $-E\{\cdot\}$ we get

$$I(\theta) = -E\left\{\frac{\partial^2}{\partial \theta^2} \log f(\underline{X};\theta)\right\}$$

$$= k(\theta) - k'(\theta) \underbrace{E\{\hat{\theta}(\underline{X}) - \theta\}}_{=0}$$

$$= k(\theta).$$

∎

CRLB: Vector Parameter

The vector CRLB has the form

$$\mathrm{Cov}_{\underline{\theta}}(\underline{\hat{\theta}}) \geq I(\underline{\theta})^{-1}$$

where

$$\mathrm{Cov}_{\underline{\theta}}(\underline{\hat{\theta}}) = E\left\{(\underline{\hat{\theta}}-\underline{\theta})(\underline{\hat{\theta}}-\underline{\theta})^T\right\}$$

$$= \begin{bmatrix} \mathrm{Var}(\hat{\theta}_1) & \mathrm{Cov}(\hat{\theta}_1,\theta_2) & \cdots & \mathrm{Cov}(\hat{\theta}_1,\hat{\theta}_p) \\ \mathrm{Cov}(\hat{\theta}_2,\hat{\theta}_1) & \mathrm{Var}(\hat{\theta}_2) & & \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(\hat{\theta}_p,\hat{\theta}_1) & & & \mathrm{Var}(\hat{\theta}_p) \end{bmatrix}$$

is the <u>covariance matrix</u> of $\underline{\hat{\theta}}$ and

$$I(\underline{\theta}) = E\left\{\left(\frac{\partial}{\partial \underline{\theta}} \log f(\underline{X};\underline{\theta})\right)\left(\frac{\partial}{\partial \underline{\theta}} \log f(\underline{X};\underline{\theta})\right)^T\right\}$$

is the <u>Fisher information matrix</u> of $\underline{\theta}$, and

$$\mathrm{Cov}_{\underline{\theta}}(\underline{\hat{\theta}}) \geq I(\underline{\theta})^{-1}$$

means

$$\mathrm{Cov}_{\underline{\theta}}(\underline{\hat{\theta}}) - I(\underline{\theta})^{-1} \text{ is positive semi-definite.}$$

Recall that if $\phi : \mathbb{R}^p \longrightarrow \mathbb{R}$ then

$$\frac{\partial \phi}{\partial \underline{\theta}} = \left[\frac{\partial \phi}{\partial \theta_1} \quad \cdots \quad \frac{\partial \phi}{\partial \theta_p} \right]^T =: \nabla_{\underline{\theta}} \phi$$

← alternate notation

Analogous to the scalar case, it can be shown that

$$I(\underline{\theta}) = E\left\{ \left(\frac{\partial}{\partial \underline{\theta}} \log f(\underline{X}; \theta) \right) \left(\frac{\partial}{\partial \underline{\theta}} \log f(\underline{X}; \underline{\theta}) \right)^T \right\}$$

$$= -E\left\{ \frac{\partial^2}{\partial \underline{\theta} \, \partial \underline{\theta}^T} \log f(\underline{X}; \underline{\theta}) \right\}$$

where

$$\frac{\partial \phi}{\partial \underline{\theta}^T} = \left[\frac{\partial \phi}{\partial \theta_1} \quad \cdots \quad \frac{\partial \phi}{\partial \theta_p} \right] = \left(\frac{\partial \phi}{\partial \underline{\theta}} \right)^T$$

and

$$\frac{\partial^2 \phi}{\partial \underline{\theta} \, \partial \underline{\theta}^T} = \begin{bmatrix} \frac{\partial^2 \phi}{\partial \theta_1^2} & \frac{\partial^2 \phi}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \phi}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \phi}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \phi}{\partial \theta_2^2} & & \vdots \\ \vdots & & \ddots & \\ \frac{\partial^2 \phi}{\partial \theta_p \partial \theta_1} & \cdots & & \frac{\partial^2 \phi}{\partial \theta_p^2} \end{bmatrix} =: \nabla^2_{\underline{\theta}} \phi$$

Theorem (Vector CRLB)

Let $\underline{X} \sim f(\underline{x}; \underline{\theta})$ where $\underline{\theta} \in \Theta \subseteq \mathbb{R}^p$. Assume

1. Θ is an open subset of $\mathbb{R}^p$
2. $f(\underline{x}; \underline{\theta})$ is differentiable in $\underline{\theta}$
3. The following regularity condition holds:

$$E\left\{ \frac{\partial}{\partial \underline{\theta}} \log f(\underline{X}; \underline{\theta}) \right\} = \underline{0} \qquad \forall \underline{\theta} \in \Theta$$

If $\hat{\underline{\theta}}$ is an unbiased estimator of $\underline{\theta}$ then

$$\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \geq I(\underline{\theta})^{-1}$$

with equality iff

$$\frac{\partial}{\partial \underline{\theta}} \log f(\underline{x}; \underline{\theta}) = I(\underline{\theta}) \cdot (\hat{\underline{\theta}}(\underline{x}) - \underline{\theta}) \qquad \forall \underline{x} \in \mathcal{X}.$$

Proof For the most part, the proof generalizes the proof of the scalar case, although some new techniques are necessary. See the book for details.

If A and B are symmetric matrices and $A \geq B$, then $a_{ii} \geq b_{ii}$ $\forall i$. This follows by taking $\underline{z}_i = [0 \cdots 0 \underset{\uparrow i\text{th position}}{1} 0 \cdots 0]^T$ and noting

$$0 \leq \underline{z}_i^T (A - B) \underline{z}$$

$$= a_{ii} - b_{ii}.$$

Therefore we have the following

Corollary | Under the assumptions of the CRLB, if $\underline{\hat{\theta}}$ is unbiased then

$$\text{Var}(\hat{\theta}_i) \geq [I(\underline{\theta})]^{-1}_{ii}$$

Thus, the vector CRLB implies scalar lower bounds on each component of $\underline{\hat{\theta}}$.

Furthermore, if $\text{Cov}_{\underline{\theta}}(\underline{\hat{\theta}}) = I(\underline{\theta})^{-1}$ $\forall \underline{\theta}$, then $\underline{\hat{\theta}}$ is a MVUE because

$$\text{Var}_{\underline{\theta}}(\underline{\hat{\theta}}) = E\left\{(\underline{\hat{\theta}} - \underline{\theta})^T (\underline{\hat{\theta}} - \underline{\theta})\right\}$$

$$= E\left\{\sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2\right\} = \sum_{i=1}^{N} E\left\{(\hat{\theta}_i - \theta_i)^2\right\}$$

$$= \sum_{i=1}^{N} \text{Var}(\hat{\theta}_i)$$

is minimized.

Exercise Suppose $\underline{x} = [x_1 \cdots x_N]^T$ where

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2).$$

Find the CRLB for $\underline{\theta} = [\mu \ \sigma^2]^T$.

Note: $\log f(\underline{x}; \underline{\theta}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$

Solution

$$\frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)$$

$$\frac{\partial \log f(\underline{x};\underline{\theta})}{\partial(\sigma^2)} = \frac{-N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\frac{\partial^2 \log f(\underline{x};\underline{\theta})}{\partial \mu^2} = -\frac{N}{\sigma^2}$$

$$\frac{\partial^2 \log f(\underline{x};\underline{\theta})}{\partial(\sigma^2)^2} = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\frac{\partial^2 \log f(\underline{x};\underline{\theta})}{\partial \mu\, \partial(\sigma^2)} = \frac{-1}{\sigma^4}\sum_{i=1}^{N}(x_i-\mu)$$

$$I(\underline{\theta}) = -E\begin{bmatrix} \frac{-N}{\sigma^2} & \frac{-1}{\sigma^4}\sum_{i=1}^{N}(X_i-\mu) \\ \frac{-1}{\sigma^4}\sum_{i=1}^{N}(X_i-\mu) & \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{N}(X_i-\mu)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$$

$$\Rightarrow \quad \mathrm{Var}(\hat{\mu}) \geq \frac{\sigma^2}{N}$$

$$\mathrm{Var}(\hat{\sigma^2}) \geq \frac{2\sigma^4}{N}$$

Summary

- CRLB = Lower bound on variance of any unbiased estimator
- Bound given by Fisher Information (matrix)
- score function = $\frac{\partial}{\partial \theta} \log f(\underline{x}; \underline{\theta})$
  - determines regularity condition and condition for equality
- Proof: application of Cauchy-Schwarz

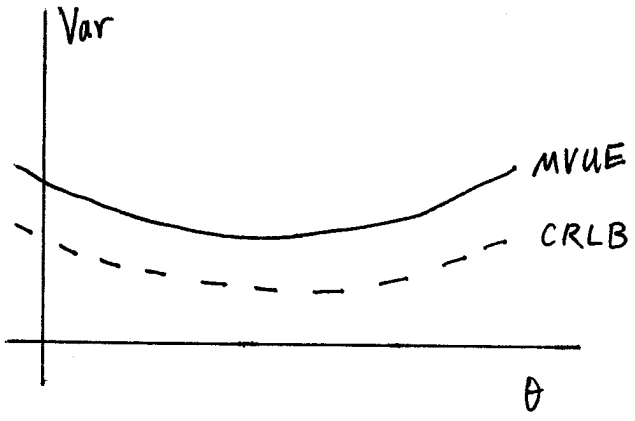
Key

a. $$-E\left\{ \frac{\partial^2 \log f(\underline{X}; \theta)}{\partial \theta^2} \right\}$$

b. $$\int \hat{\theta}(\underline{x}) f(\underline{x}; \theta) \, dx$$

MVUE VIA THE RAO-BLACKWELL THEOREM AND COMPLETE SUFF-ICIENT STATISTICS

The Cramer-Rao lower bound gives a necessary and sufficient condition for the existence of an efficient estimator.



However, MVUE's are not necessarily efficient. What can we do in such cases?

The Rao-Blackwell theorem, when applied in conjunction with a complete suff. stat., gives another way to find MVUE's that applies even when the CRLB is not defined.

Rao-Blackwell Theorem, Version 2

Theorem Let $\underline{Y}, \underline{Z}$ be random variables and define the function

$$g(\underline{z}) = E[\underline{Y} \mid \underline{Z} = \underline{z}].$$

Then

$$E[g(\underline{Z})] = E[\underline{Y}]$$

and

$$Var(g(\underline{Z})) \leq Var(\underline{Y})$$

with equality iff $\underline{Y} = g(\underline{Z})$ almost surely.

Note that this version of R.B. is quite general and has nothing to do with estimation of parameters. However, we can apply it to parameter estimation as follows.

Consider $\underline{X} \sim f_{\underline{\theta}}(\underline{x})$. Let $\hat{\underline{\theta}}_1$ be an unbiased estimator of $\underline{\theta}$, and let $\underline{T} = \tau(\underline{X})$ be a sufficient statistic for $\underline{\theta}$. Apply Rao-Blackwell with

$$\underline{Y} = \hat{\underline{\theta}}_1(\underline{X})$$

$$\underline{Z} = \underline{T} = \tau(\underline{X}).$$

Consider the new estimator

$$\hat{\underline{\theta}}_2(\underline{x}) = g(\tau(\underline{x}))$$

$$= E\left[\hat{\underline{\theta}}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x})\right].$$

Then we may conclude

① $\hat{\underline{\theta}}_2$ is unbiased

② $Var_{\underline{\theta}}(\hat{\underline{\theta}}_2) \leq Var_{\underline{\theta}}(\hat{\underline{\theta}}_1)$

In words, if $\hat{\underline{\theta}}_1$ is any unbiased estimator, then smoothing $\hat{\underline{\theta}}_1$ w.r.t. a sufficient stat decreases the variance while preserving unbiasedness.

Therefore, we can restrict our search for the MVUE to functions of a sufficient statistic.

Proof of Rao-Blackwell, Version 2

First we must show $E[g(\underline{Z})] = E[\underline{Y}]$.

This follows by the law of total expectation:

$$E[g(\underline{Z})] = E[E[\underline{Y} \mid \underline{Z}]] = E[\underline{Y}]$$

Second we must show

$$E[(g(\underline{Z})-\underline{\theta})^T(g(\underline{Z})-\underline{\theta})] \leq E[(\underline{Y}-\underline{\theta})^T(\underline{Y}-\underline{\theta})],$$

where $\underline{\theta} = E[\underline{Y}] = E[g(\underline{Z})]$. To see this, write

$$\begin{aligned}
Var(\underline{Y}) &= E[(\underline{Y}-\underline{\theta})^T(\underline{Y}-\underline{\theta})] \\
&= E[(\underline{Y}-g(\underline{Z})+g(\underline{Z})-\underline{\theta})^T(\underline{Y}-g(\underline{Z})+g(\underline{Z})-\underline{\theta})] \\
&= E[(\underline{Y}-g(\underline{Z}))^T(\underline{Y}-g(\underline{Z}))] + Var(g(\underline{Z})) \\
&\quad + 2E[(\underline{Y}-g(\underline{Z}))^T(g(\underline{Z})-\underline{\theta})]
\end{aligned}$$

Consider the third term:

$$E[(\underline{Y}-g(\underline{Z}))^T(g(\underline{Z})-\underline{\theta})]$$

$$=$$

(a)

Thus, we have shown

$$\begin{aligned} \mathrm{Var}(\underline{Y}) &= \mathrm{Var}(g(\underline{Z})) + E\left[(\underline{Y}-g(\underline{Z}))^T(\underline{Y}-g(\underline{Z}))\right] \\ &= \mathrm{Var}(g(\underline{Z})) + E\left[\|\underline{Y}-g(\underline{Z})\|^2\right] \\ &\geq \mathrm{Var}(g(\underline{Z})) \end{aligned}$$

with equality iff $\underline{Y} = g(\underline{Z})$ with probability one. ∎

Example Consider $\underline{X} \sim N(\theta \cdot \underline{1}, \sigma^2 I_{N\times N})$, σ^2 known. Let $\hat{\theta}_1(\underline{x}) = x_1$. Then

$$E[\hat{\theta}_1] = \theta$$

$$\mathrm{Var}(\hat{\theta}_1) = \sigma^2.$$

Consider the sufficient statistic $T = \sum_{i=1}^{N} X_i$

and define

$$\hat{\theta}_2(\underline{x}) = E\left[\hat{\theta}_1(\underline{X}) \mid T = \Sigma x_i\right]$$

How can we find a formula for $\hat{\theta}_2$?

Observe that X_1, T are jointly Gaussian:

$$\begin{bmatrix} X_1 \\ T \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \end{bmatrix}}_{A} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$$

Then

$$\begin{bmatrix} X_1 \\ T \end{bmatrix} \sim N\left(A \cdot \theta \underline{1} \,,\, A \cdot \sigma^2 I_{N\times N} \cdot A^T \right)$$

$$\sim N\left(\begin{bmatrix} \theta \\ N\theta \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & N \end{bmatrix} \right)$$

Recall the following property of the MVG: If

$$\underline{W} = \begin{bmatrix} \underline{U} \\ \underline{V} \end{bmatrix} \sim N\left(\begin{bmatrix} \underline{\mu}_u \\ \underline{\mu}_v \end{bmatrix}, \begin{bmatrix} R_{uu} & R_{uv} \\ R_{vu} & R_{vv} \end{bmatrix} \right)$$

Then

$$\underline{U} \,|\, \underline{V} = \underline{v} \sim N\Big(\underline{\mu}_u + R_{uv} R_{vv}^{-1} (\underline{v} - \underline{\mu}_v), \; R_{uu} - R_{uv} R_{vv}^{-1} R_{vu} \Big)$$

Applying this to $\begin{bmatrix} X_1 \\ T \end{bmatrix}$ we obtain

$$X_1 \mid T=t \sim N\left(\theta + 1 \cdot \frac{1}{N}(t - N\theta),\ \sigma^2\left(1 - 1 \cdot \frac{1}{N} \cdot 1\right)\right)$$

$$\sim N\left(\frac{t}{N},\ \sigma^2\left(1 - \frac{1}{N}\right)\right)$$

Therefore

$$\hat{\theta}_2(\underline{x}) = E\left[X_1 \mid T = \Sigma x_i\right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i$$

Notice the reduction in variance:

$$Var(\hat{\theta}_2) = \frac{\sigma^2}{N} < \sigma^2 = Var(\hat{\theta}_1).$$

The Rao-Blackwell Theorem tells us how to decrease the variance of an unbiased estimator. But when can we know that we get a MVUE?

The answer: When $\underline{T}$ is a complete suff. stat.

Theorem (Lehmann-Scheffe)

If $\underline{T}$ is complete, there is at most one unbiased estimator that is a function of $\underline{T}$.

Proof Suppose $E[\hat{\underline{\theta}}_1] = E[\hat{\underline{\theta}}_2] = \underline{\theta}$ and

$\hat{\underline{\theta}}_1(\underline{x}) = g_1(\tau(\underline{x}))$, $\hat{\underline{\theta}}_2(\underline{x}) = g_2(\tau(\underline{x}))$.

Define

$$\phi(\underline{t}) = g_1(\underline{t}) - g_2(\underline{t}).$$

Then

$$E\{\phi(\underline{T})\} = E\{g_1(\underline{T}) - g_2(\underline{T})\}$$

$$= E\{\hat{\underline{\theta}}_1(\underline{X}) - \hat{\underline{\theta}}_2(\underline{X})\}$$

$$= \underline{0}.$$

By definition of completeness,

$$P(\phi(I) = 0) = 1 \qquad \forall \underline{\theta}.$$

In other words,

$$\hat{\underline{\theta}}_1 = \hat{\underline{\theta}}_2 \quad \text{with prob. } 1 \qquad \blacksquare$$

This result suggests the following recipe for finding an MVUE:

1.) Find a complete sufficient statistic $\underline{I} = \tau(\underline{X})$

2.a) Find any unbiased estimator $\hat{\underline{\theta}}'$ and set

$$\hat{\underline{\theta}}(x) = E\left\{ \hat{\underline{\theta}}'(\underline{X}) \mid \underline{I} = \tau(\underline{x}) \right\}$$

2.b) Find a function g such that

$$\hat{\underline{\theta}}(\underline{x}) = g(\tau(\underline{x}))$$

is unbiased.

Theorem If $\hat{\underline{\theta}}$ is constructed by the recipe above, then $\hat{\underline{\theta}}$ is the unique MVUE.

Proof | Note that in either construction, $\hat{\underline{\theta}}$ is a function of $\underline{T}$.

Let $\hat{\underline{\theta}}_1$ be any unbiased estimator. We must show

$$Var(\hat{\underline{\theta}}) \leq Var(\hat{\underline{\theta}}_1).$$

Define

$$\hat{\underline{\theta}}_2(\underline{x}) = E\left\{ \hat{\underline{\theta}}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x}) \right\}.$$

By Rao-Blackwell, it suffices to show

$$Var(\hat{\underline{\theta}}) \leq Var(\hat{\underline{\theta}}_2).$$

But $\hat{\underline{\theta}}$ and $\hat{\underline{\theta}}_2$ are both unbiased and functions of a complete S.S.

$$\Longrightarrow \quad \hat{\underline{\theta}} = \hat{\underline{\theta}}_2 \qquad \text{w.p. } 1.$$

To show uniqueness, in the above argument suppose $Var(\hat{\underline{\theta}}_1) = Var(\hat{\underline{\theta}})$. Then the Rao-Blackwell bound holds with equality

$$\Longrightarrow \quad \hat{\underline{\theta}} = \hat{\underline{\theta}}_2 \qquad \text{w.p. } 1$$

$$\Longrightarrow \quad \hat{\underline{\theta}}_1 = \hat{\underline{\theta}} \qquad \text{w.p. } 1$$

because $\hat{\underline{\theta}}_2 = \hat{\underline{\theta}}$ w.p. 1. ∎

We have seen previously that sufficient statistics arising from the exponential family of distributions are complete. Typically, however, MVUE's for the exponential family can be found using the CRLB.

A strength of the Rao-Blackwell approach is that it can produce MVUE's even when CRLB can't.

Example | Suppose $\underline{x} = [x_1 \cdots x_N]^T$ where

$$x_i \overset{iid}{\sim} \text{unif}[0, \theta], \quad i = 1, \ldots, N.$$

Note that the CRLB cannot be applied because

$$\log f(\underline{x}; \theta)$$

is not differentiable w.r.t θ.

What is an unbiased estimator of θ?

$$\hat{\theta}_1 = \frac{2}{N} \sum_{i=1}^{N} X_i$$

is unbiased. However, it is not MVUE.

From

$$f_\theta(\underline{x}) = \prod_{i=1}^{N} \frac{1}{\theta} I_{[0,\theta]}(x_i)$$

$$= \underbrace{\frac{1}{\theta^N} I_{[\max x_i, \infty)}(\theta)}_{g_\theta(\underline{t})} \cdot \underbrace{I_{(-\infty, \min x_i]}(0)}_{h(\underline{x})}$$

we see that

$$T = \max_i X_i$$

is a sufficient statistic. It is left as an exercise to show that T is in fact complete.

Since $\hat{\theta}_1$ is not a function of T, it is not MVUE.

However

$$\hat{\theta}_2(\underline{x}) = E\left\{ \hat{\theta}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x}) \right\}$$

is the MVUE.

It is also left as an exercise to find the precise form of $\hat{\theta}_2$.

Summary

- Rao-Blackwell Thm
  - decreases estimator variance by conditioning on a sufficient statistic
  - filters out the randomness (noise) in the data not captured by suff. stat.

- Complete suff. stat. $\Rightarrow$ Rao-Blackwellization results in the unique MVUE

Key

a. $E[E[(\underline{Y} - g(\underline{Z}))^T (g(\underline{Z}) - \underline{\theta}) \mid \underline{Z}]]$

$= E[E[(\underline{Y} - g(\underline{Z}))^T \mid \underline{Z}] \cdot (g(\underline{Z}) - \underline{\theta})]$

$= E[(g(\underline{Z}) - g(\underline{Z}))^T (g(\underline{Z}) - \underline{\theta})]$

$= 0$

MAXIMUM LIKELIHOOD ESTIMATION

Consider the **usual estimation setup** where we observe

$$\underline{X} \sim f_{\underline{\theta}}(\underline{x}) = f(\underline{x}; \underline{\theta}).$$

Viewing $\underline{x}$ as fixed and $\underline{\theta}$ as the variable, we call

$$l(\underline{\theta}; \underline{x}) := f(\underline{x}; \underline{\theta})$$

the *likelihood* of $\underline{\theta}$ (given $\underline{x}$).

Definition The estimator $\hat{\underline{\theta}}$ is called a *maximum likelihood estimator* if $\forall \underline{x}$

$$l(\hat{\underline{\theta}}(\underline{x}); \underline{x}) = \max_{\underline{\theta} \in (H)} l(\underline{\theta}; \underline{x}).$$

Equivalently, $\hat{\underline{\theta}}$ satisfies

$$\hat{\underline{\theta}}(\underline{x}) = \arg\max_{\underline{\theta} \in (H)} l(\underline{\theta}; \underline{x})$$

Intuitively, the MLE selects the value of $\underline{\theta}$ such that, in retrospect, the observed $\underline{x}$ corresponds to the most probable outcome.

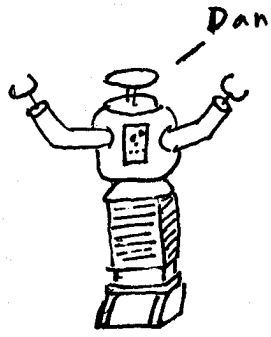
Note: The MLE is not always unique.

Warning

It is tempting to view the likelihood as a density / mass function for $\underline{\theta}$, conditioned on $\underline{X} = \underline{x}$. However, the MLE is a classical estimator that views $\underline{\theta}$ as nonrandom. Furthermore, in some cases

$$\int \ell(\underline{\theta}; \underline{x}) \, d\underline{\theta} = \infty$$

and so the likelihood cannot be normalized.



Likelihood Principle

The information contained in an observation $\underline{x}$ about $\underline{\theta}$ is contained entirely in the likelihood function $\ell(\underline{\theta}; \underline{x})$.

Moreover, if $\underline{x}_1$ and $\underline{x}_2$ are two observations depending on $\underline{\theta}$ (perhaps through different models) such that

$$\ell(\underline{\theta}; \underline{x}_1) = c \cdot \ell(\underline{\theta}; \underline{x}_2) \qquad \forall \underline{\theta}$$

for some constant c, then $\underline{x}_1$ and $\underline{x}_2$ must lead to the same inference about $\underline{\theta}$.

Example Suppose a public health official conducts a survey to estimate $0 \le \theta \le 1$, the percentage of the population eating pizza at least once per week. As a result of the survey, the official found 9 pizza eaters and 3 non-eaters.

If no additional information is available regarding how the survey was implemented, then there are at least two possible probability models

1. $X_1 \sim \text{Bin}(12, \theta)$

$x_1 = 9$ observed

$$p_1(x_1; \theta) = \binom{12}{x_1} \theta^{x_1} (1-\theta)^{12-x_1}$$

2. $X_2 \sim \text{Neg}(3, 1-\theta)$

$x_2 = 12$ observed

$$p_2(x_2; \theta) = \binom{x_2-1}{3-1} (1-\theta)^3 \theta^{x_2-3}$$

In both cases, the likelihood is proportional to

$$l(\theta; x) \propto \theta^9 (1-\theta)^3$$

If we follow the likelihood principle, both models lead to the same inference about θ.

Sufficiency Principle

The MLE also satisfies the sufficiency principle, which states that if $\underline{T} = \tau(\underline{X})$ is sufficient for $\underline{\theta}$ and $\underline{x}_1$ and $\underline{x}_2$ are such that $\tau(\underline{x}_1) = \tau(\underline{x}_2)$, then $\underline{x}_1$ and $\underline{x}_2$ must lead to the same estimate of $\underline{\theta}$.

To see this, note

$$\begin{aligned}\hat{\underline{\theta}}(\underline{x}) &= \arg\max_{\underline{\theta}} f(\underline{x}; \underline{\theta}) \\ &= \arg\max_{\underline{\theta}} g(\tau(\underline{x}); \underline{\theta}) \cdot h(\underline{x}) \\ &= \arg\max_{\underline{\theta}} g(\tau(\underline{x}); \underline{\theta})\end{aligned}$$

which depends on $\underline{x}$ only through $\underline{t} = \tau(\underline{x})$.

Computing the MLE

Since many of the models we work with have an exponential form, it is often convienient to maximize the log-likelihood

$$\log l(\underline{\theta}; \underline{x})$$

If the likelihood function is differentiable, then $\hat{\underline{\theta}}(\underline{x})$ is a solution of

$$\underbrace{\frac{\partial}{\partial \underline{\theta}}}_{\nabla_{\underline{\theta}}} \log l(\underline{\theta}; \underline{x}) = \underline{0}.$$

We also need to verify that such a solution is in fact a local max and not a local min or a saddle point. This can be accomplished by checking to see that

$$\underbrace{\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T}}_{\nabla^2_{\underline{\theta}}} \log f(\underline{\theta}; \underline{x})$$

Hessian

is negative semidefinite at $\hat{\underline{\theta}}(\underline{x})$, or by otherwise arguing that $\hat{\underline{\theta}}$ is a local max. If several local maximums exists, the MLE is the one with largest likelihood.

Example Suppose $\underline{x} = [x_1, \ldots, x_N]^T$ where

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, N.$$

The log-likelihood of μ is

$$\log l(\mu; \underline{x}) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

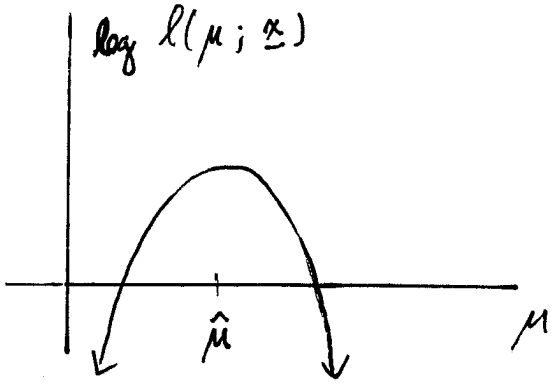
$$\frac{\partial \log l(\mu; \underline{x})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

$$= 0$$

$$\Rightarrow \quad \sum (x_i - \mu) = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Note that $\hat{\mu}$ can't be a local min because $\log l(\mu; \underline{x})$ is concave. Therefore the MLE is the sample mean.



Exercise In the previous example, suppose σ^2 is unknown and find the MLE of $\underline{\theta} = [\mu \ \sigma^2]^T$.

Solution

$$\log \ell(\mu, \sigma^2 ; \underline{x}) = -\frac{N}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\frac{\partial \log \ell(\mu, \sigma^2 ; \underline{x})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

$$\frac{\partial \log \ell(\mu, \sigma^2 ; \underline{x})}{\partial (\sigma^2)} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 .$$

Therefore, the MLE $[\hat{\mu}, \hat{\sigma^2}]^T$ must solve the system of equations

$$\frac{1}{\hat{\sigma^2}} \sum_{i=1}^{N} (x_i - \hat{\mu}) = 0 \qquad (1)$$

$$\frac{-N}{2\hat{\sigma^2}} + \frac{1}{2(\hat{\sigma^2})^2} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 = 0 . \qquad (2)$$

$$(1) \Rightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$(2) \Rightarrow \quad \hat{\sigma^2} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

biased

The solution is a maximum because the Hessian is negative semi-definite at $\underline{\hat{\theta}}(\underline{x})$

This is left as an exercise.

The MLE can be computed in closed form for many common distributions, including many members of the exponential family.

Example A coin has $\text{Prob}\{\text{heads}\} = \theta$. To estimate θ, the following experiment is performed N times: The coin is flipped until 10 heads have been observed, and the total number of flips X is recorded. If the values $x_1, \ldots, x_N$ are observed, find the MLE of θ.

X has a negative binomial distribution:

$$p(x;\theta) = \binom{x-1}{9} \theta^{10} (1-\theta)^{x-10}, \quad x \geq 10$$

Assuming independent experiments,

$$\ell(\theta; \underline{x}) = \prod_{i=1}^{N} p(x_i;\theta)$$

$$= \prod_{i=1}^{N} \binom{x_i-1}{9} \theta^{10} (1-\theta)^{x_i-10}$$

$$= \left[\prod_{i=1}^{N} \binom{x_i-1}{9} \right] \theta^{10N} (1-\theta)^{\sum x_i - 10N}$$

$$\log l(\theta; \underline{x}) = 10N \log \theta + (\Sigma x_i - 10N) \log(1-\theta) + C$$

$$\frac{\partial \log l(\theta; \underline{x})}{\partial \theta} = \frac{10N}{\theta} - \frac{\Sigma x_i - 10N}{1-\theta} = 0$$

$$\Rightarrow (1-\theta) 10N = \theta (\Sigma x_i - 10N)$$

$$\Rightarrow \hat{\theta}(\underline{x}) =$$

If you think about it, this makes good sense.

Asymptotic Properties

Theorem Suppose $\underline{X} \sim f(\underline{x}; \underline{\theta})$. Let $\hat{\underline{\theta}}_N$ be the MLE of $\underline{\theta}$ based on n iid realizations $\underline{X}_1, \ldots, \underline{X}_N$ of $\underline{X}$. Under certain regularity conditions,

$$\sqrt{N}(\hat{\underline{\theta}}_N - \underline{\theta}) \xrightarrow{D} N(\underline{0}, I(\underline{\theta})^{-1})$$

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

Remarks

- The regularity condition amounts to $f(\underline{x}; \underline{\theta})$ having bounded third derivatives w.r.t. $\underline{\theta}$.
- Proof hinges on central limit theorem

- $E\,\hat{\underline{\theta}}_N \longrightarrow \underline{\theta} \implies$ MLE is asymptotically unbiased
- $\mathrm{Cov}\,\hat{\underline{\theta}}_N \longrightarrow I(\underline{\theta})^{-1} \implies$ MLE is asymptotically efficient
- $\sqrt{N}$ characterizes the rate of convergence.

Example Recall the MLE of $\underline{\theta} = [\mu \;\; \sigma^2]^T$ based on

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2), \qquad i = 1, \ldots, N$$

is $\hat{\underline{\theta}} = [\hat{\mu} \;\; \hat{\sigma}^2]^T$ where

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Also recall that

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} (X_i - \overline{X})^2 \sim \chi^2_{N-1}$$

and is independent of $\overline{X} = \hat{\mu}$.

This implies

$$\frac{N}{\sigma^2} \cdot \hat{\sigma}^2 \sim \chi^2_{N-1}$$

$$\Rightarrow \quad E\,\hat{\sigma}^2 = \frac{N-1}{N}\sigma^2$$

$$\text{Var}\,\hat{\sigma}^2 = \frac{2(N-1)}{N^2}\sigma^4$$

Therefore, as $N \to \infty$

$$E\,\hat{\underline{\theta}} = \begin{bmatrix} \mu \\ \frac{N-1}{N}\sigma^2 \end{bmatrix} \to \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \underline{\theta}$$

$$\text{Cov}\,\hat{\underline{\theta}} = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)}{N^2}\sigma^4 \end{bmatrix} \to \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix} = I(\underline{\theta})^{-1}$$

Asymptotic normality can also be verified from the fact

$$\chi^2_r \to \mathcal{N}(r, 2r) \qquad \text{as} \quad r \to \infty$$

which follows from the CLT.

Additional Topics

Nuisance Parameters

Suppose $\underline{X} \sim f(\underline{x}; \underline{\theta})$ and

$\underline{\theta} = \begin{bmatrix} \underline{\theta}_1 \\ \underline{\theta}_2 \end{bmatrix}$, where only $\underline{\theta}_1$ is

of interest. Then the MLE of $\underline{\theta}_1$

is defined to be

$$\tilde{\underline{\theta}}_1(\underline{x}) = \arg\max_{\underline{\theta}_1} \left[\max_{\underline{\theta}_2} \ell(\underline{\theta}_1, \underline{\theta}_2; \underline{x}) \right]$$

Example $X_1, \ldots, X_N \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Invariance

Suppose $\underline{\theta} \in \Theta \subseteq \mathbb{R}^p$ and that our objective is to estimate

$$\underline{\varphi} = g(\underline{\theta})$$

If g is invertible, then we may parametrize $f_1(\underline{x}; \underline{\theta})$ in terms of $\underline{\varphi}$:

$$f_2(\underline{x}; \underline{\varphi}) = f_1(\underline{x}; g^{-1}(\underline{\varphi}))$$

We may then estimate $\underline{\varphi}$ via maximum likelihood:

$$\hat{\underline{\varphi}}(\underline{x}) = \arg\max_{\underline{\varphi}} f_2(\underline{x}; \underline{\varphi})$$

Fortunately the MLE is <u>invariant</u> to such transformations:

<u>Theorem</u> Let $\underline{\varphi} = g(\underline{\theta})$ be invertible and let $\hat{\varphi}$ and $\hat{\theta}$ denote the MLEs. Then

$$\hat{\varphi}(\underline{x}) = g(\hat{\underline{\theta}}(\underline{x})).$$

Proof |

$$\hat{\varphi}(\underline{x}) = \arg\max_{\underline{\varphi}} f_2(\underline{x}; \underline{\varphi})$$

$$= \arg\max_{\underline{\varphi}} f_1(\underline{x}; g^{-1}(\underline{\varphi}))$$

$$= g\left(\arg\max_{\underline{\theta}} f_1(\underline{x}; g^{-1}(g(\underline{\theta})))\right)$$

$$= g\left(\arg\max_{\underline{\theta}} f_1(\underline{x}; \underline{\theta})\right)$$

$$= g(\hat{\underline{\theta}}(\underline{x}))$$

Example | Suppose $\underline{X} \sim N(\mu \cdot \underline{1}, \sigma^2 I)$.

Then the MLE of σ is

(a) $\hat{\sigma} =$

If $\underline{\varphi} = g(\underline{\theta})$ and g is many-to-one (i.e. not invertible), then we cannot parametrize the distribution in terms of $\underline{\varphi}$. In this case we define the MLE of $\underline{\varphi}$ to be

$$\hat{\varphi}(\underline{x}) := g(\hat{\underline{\theta}}(\underline{x})).$$

Now the MLE is invariant by definition.

Exercise] Suppose $X_i \overset{iid}{\sim}$ Bernoulli (θ), $i = 1, \ldots, N$.

Find the MLE of the variance of X.

Solution] The MLE of θ is

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The variance of a Bernoulli trial is

$$\text{Var } X = \theta(1-\theta)$$

Thus the MLE of the variance is

$$\left(\frac{1}{N} \sum_{i=1}^{N} x_i \right) \cdot \left(1 - \frac{1}{N} \sum_{i=1}^{N} x_i \right)$$

MLE of MVG with unknown mean and covariance

Suppose $\underline{X} \sim N(\underline{\mu}, R)$ and a sample $\underline{X}_1, \ldots, \underline{X}_N$ of iid observations is collected.

It can be shown that the MLE of $\underline{\theta} = [\underline{\mu}, R]$ is

$$\hat{\underline{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i$$

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} (\underline{x}_i - \hat{\underline{\mu}})(\underline{x}_i - \hat{\underline{\mu}})^T$$

Note: If $\underline{X}$ is n-dimensional, then the dimension of $\underline{\theta}$ is _____

Proof of this result is more involved than the scalar case ($n=1$). You may find the proof online if you are curious.

Computational Issues

In many cases of interest, the MLE cannot be expressed in closed form. Iterative numerical techniques are then necessary to maximize the likelihood. Examples include

- Newton-Raphson iteration
- The "scoring method" of iteration

$$\underline{\hat{\theta}}_{k+1} = \underline{\hat{\theta}}_k + \left[I(\underline{\theta})^{-1} \frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \underline{\theta}} \right]\Bigg|_{\underline{\theta}=\underline{\theta}_k}$$

- An expectation-maximization (EM) algorithm.

Efficiency and the MLE

In previous examples, we have seen that the MLE is sometimes efficient. There is a precise connection:

Theorem Assume that $f(\underline{x};\underline{\theta}) = \ell(\underline{\theta};\underline{x})$ has ≤ 1 local max. If $\hat{\underline{\theta}}$ is efficient, that is, $E\hat{\underline{\theta}} = \underline{\theta}$ and $\operatorname{Cov}\hat{\underline{\theta}} = I(\underline{\theta})^{-1}$ $\forall \underline{\theta}$, then $\hat{\underline{\theta}}$ is an MLE.

Proof From the CRLB, $\hat{\underline{\theta}}$ is efficient iff

$$\frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \underline{\theta}} = I(\underline{\theta})\left(\hat{\underline{\theta}}(\underline{x}) - \underline{\theta}\right) \quad \forall \underline{x}\ \forall \underline{\theta}.$$

Take $\underline{\theta} = \hat{\underline{\theta}}(\underline{x})$. Then

$$\left.\frac{\partial \log f(\underline{x};\underline{\theta})}{\partial \underline{\theta}}\right|_{\underline{\theta}=\hat{\underline{\theta}}(\underline{x})} = I(\hat{\underline{\theta}}(\underline{x})) \cdot \left(\hat{\underline{\theta}}(\underline{x}) - \hat{\underline{\theta}}(\underline{x})\right)$$

$$= \underline{0}.$$

By the product rule

$$\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(\underline{x}; \underline{\theta}) = -I(\underline{\theta}) + (\hat{\underline{\theta}}(\underline{x}) - \underline{\theta}) \frac{\partial}{\partial \underline{\theta}^T} I(\underline{\theta})$$

Again setting $\underline{\theta} = \hat{\underline{\theta}}(\underline{x})$ we have

$$\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(\underline{x}; \underline{\theta}) \Big|_{\underline{\theta} = \hat{\underline{\theta}}(\underline{x})} = -I(\underline{\theta}) \leq 0$$

$\Longrightarrow$ $\hat{\underline{\theta}}(\underline{x})$ is a local max. Since $\log f(\underline{x}; \underline{\theta})$ has at most one local max, $\hat{\underline{\theta}}$ is the MLE.

Summary

1. MLE is one implementation of the likelihood and sufficiency principles.
2. MLE is asymptotically normal and asymptotically efficient. (under certain conditions)
3. MLE is invariant under reparametrization
4. Efficient estimators are usually MLEs

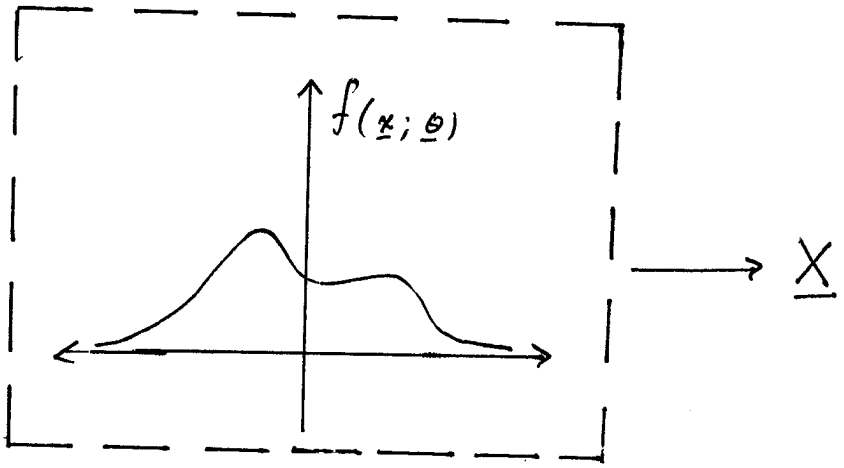
Key

a. $\sqrt{\hat{\sigma^2}} = \sqrt{\frac{1}{N} \sum (x_i - \bar{x})^2}$, b. $n + \frac{n(n+1)}{2}$

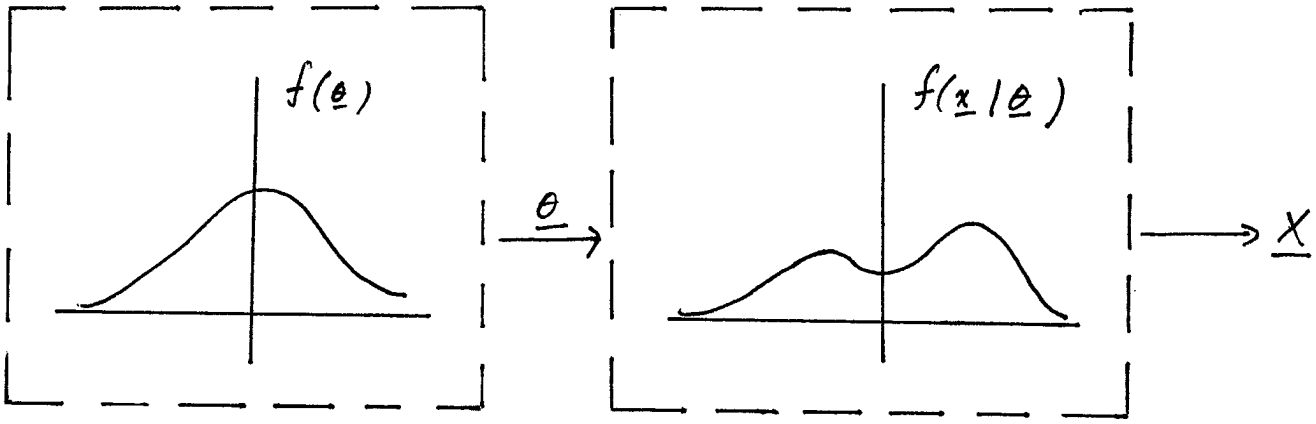
BAYESIAN ESTIMATION

Bayesian Statistical Modeling

In classical estimation, the unknown parameter $\underline{\theta}$ is viewed as nonrandom, and a statistical model is specified entirely by a data model (likelihood) $f(\underline{x}; \underline{\theta})$



In Bayesian estimation, the unknown parameter is viewed as random. A statistical model is specified in terms of a conditional pdf/pmf $f(\underline{x} \mid \underline{\theta})$ and a prior distribution $f(\underline{\theta})$ on $\underline{\theta}$.



The prior $f(\underline{\theta})$ is specified by the investigator and reflects "prior knowledge" about the uncertainty in $\underline{\theta}$.

Note that we now write $f(\underline{x} \mid \underline{\theta})$ instead of $f(\underline{x}; \underline{\theta})$ to reflect that $\underline{\theta}$ is random.

By Bayes' rule, we may express the <u>posterior</u> distribution of $\underline{\theta}$ given $\underline{x}$ as

$$f(\underline{\theta} \mid \underline{x}) = \frac{f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta})}{f(\underline{x})}$$

$$= \frac{f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta})}{\int f(\underline{x} \mid \underline{\theta}') \, f(\underline{\theta}') \, d\underline{\theta}'}$$

Whereas the prior reflects our uncertainty in θ before x is observed, the posterior represents our uncertainty in θ after x is observed.

Example 1 Suppose we have a coin whose probability θ of turning up heads is unknown. We toss the coin N times and observe X heads.

The natural model for X given θ is binomial:

$$p(x|\theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x}$$

One possible prior for θ is the beta distribution:

$$f(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 \le \theta \le 1$$

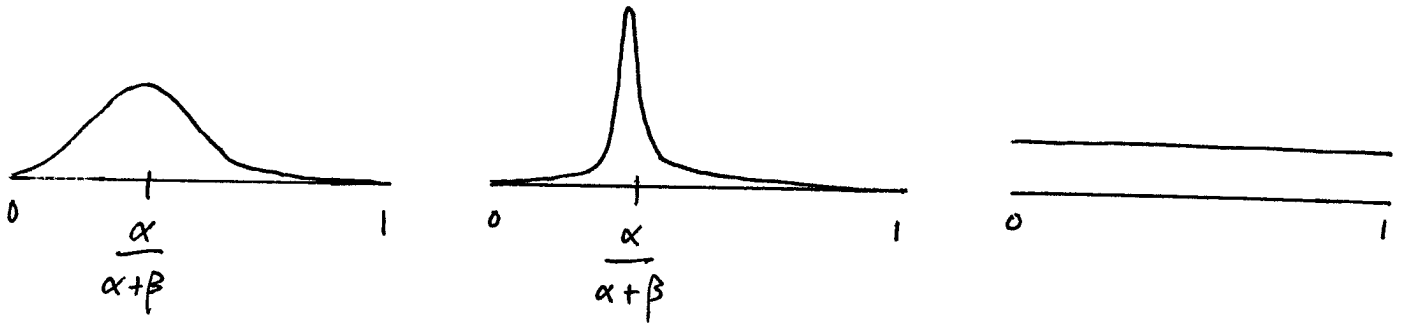
where

$$B(\alpha,\beta) := \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

and $\alpha, \beta \ge 1$.

$$E\{\theta\} = \frac{\alpha}{\alpha+\beta}, \quad Var\{\theta\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



α, β small $\qquad$ α, β large $\qquad$ $\alpha = \beta = 1$ (uniform)

The parameters α, β must be set by the user to reflect prior knowledge

- $\alpha = \beta = 1 \implies \theta$ could be anywhere
- $\alpha = \beta = 2 \implies \theta$ is probably fair or close to fair, but not really sure
- $\alpha = \beta = 10 \implies \theta$ is almost certainly fair

Let's see how our belief about θ changes once x is observed.

Viewing x as a constant, we have

$$f(\theta|x) = \frac{p(x|\theta)\, f(x)}{p(x)}$$

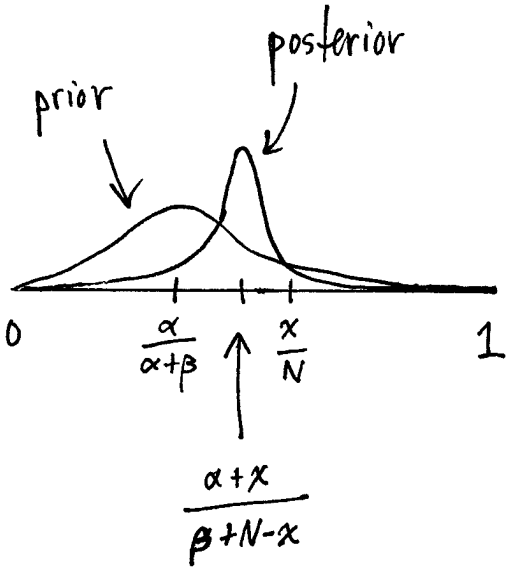
$$\propto p(x|\theta)\, f(x)$$

$$= \binom{N}{x} \theta^x (1-\theta)^{N-x} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

$$\propto \theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}$$

Since $f(\theta|x)$ is a density it must integrate to 1 and we recognize $\theta|x \sim \text{Beta}(\alpha+x, \beta+N-x)$ and

$$f(\theta|x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}}{B(\alpha+x, \beta+N-x)}, \quad 0 \le \theta \le 1$$



The posterior is shifted toward the observed frequency and is more concentrated (reflecting greater certainty).

The beta prior is said to be conjugate to the binomial data model because the prior and posterior belong to the same family.

Confidence Statements

One advantage of Bayesian over classical estimation is that in Bayesian inference, confidence statements are more natural.

For example, if we toss a coin 10 times and observe 10 heads, it makes perfect sense to assert "it is highly probable that the coin is unfair and biased towards heads."

Formally, let

$$\mathcal{H} = \left(\tfrac{1}{2}, 1\right]$$

$$X \sim \binom{N}{x} \theta^x (1-\theta)^{N-x}.$$

What is

$$\text{Prob}\left(\theta \in \mathcal{H} \mid X = x\right)?$$

Asking such a question already suggests that θ is random. In the Bayesian framework, the answer to the question is

$$\text{Prob}\left(\theta \in \textcircled{A} \mid \underline{x}\right) = \int_{\textcircled{A}} f(\theta \mid x)\, d\theta$$

Even though such confidence statements are a normal part of "everyday thinking," they are less natural in the classical setting

Likelihood Principle

All Bayesian methods for estimation are based on the posterior. As a consequence, Bayesian estimation conforms to the likelihood principle. To see this, suppose $\underline{x}_1$ and $\underline{x}_2$ are such that

$$f(\underline{x}_1 \mid \underline{\theta}) = c \cdot f(\underline{x}_2 \mid \underline{\theta}) \qquad \forall \underline{\theta}$$

for some constant c. Then

$$f(\underline{\theta} \mid \underline{x}_1) = \frac{f(\underline{x}_1 \mid \underline{\theta}) \cdot f(\underline{\theta})}{f(\underline{x}_1)}$$

$$= \left[\frac{c f(\underline{x}_2)}{f(\underline{x}_1)}\right] \cdot \frac{f(\underline{x}_2 \mid \underline{\theta}) \cdot f(\underline{\theta})}{f(\underline{x}_2)}$$

$$\propto f(\underline{\theta} \mid \underline{x}_2)$$

$$\Rightarrow f(\underline{\theta} \mid \underline{x}_1) = f(\underline{\theta} \mid \underline{x}_2).$$

Sufficiency Principle

Bayesian inference also obeys the sufficiency principle, which states that if $\tau(\underline{x}_1) = \tau(\underline{x}_2)$, where $\underline{T} = \tau(\underline{X})$ is a sufficient statistic, then $\underline{x}_1$ and $\underline{x}_2$ must lead to the same inference about $\underline{\theta}$.

To see this, note

$$f(\underline{\theta} \mid \underline{x}) = \frac{f(\underline{x} \mid \underline{\theta})\, f(\underline{\theta})}{f(\underline{x})}$$

$$= \frac{f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta})}{\int f(\underline{x} \mid \underline{\theta}')\, f(\underline{\theta}')\, d\underline{\theta}'}$$

$$= \frac{g(\underline{t} \mid \underline{\theta})\, h(\underline{x})\, f(\underline{\theta})}{\int g(\underline{t} \mid \underline{\theta}')\, h(\underline{x})\, f(\underline{\theta}')\, d\underline{\theta}'}$$

$$= \frac{g(\underline{t} \mid \underline{\theta})\, f(\underline{\theta})}{\int g(\underline{t} \mid \underline{\theta}')\, f(\underline{\theta}')\, d\underline{\theta}'}$$

which depends on $\underline{x}$ only through $\underline{t} = \tau(\underline{x})$.

Pros of Bayesian Inference

- allows incorporation of prior information
- leads to better estimates if prior information is accurate
- provides for valid confidence statements
- satisfies the likelihood and sufficiency principles

Cons of Bayesian Inference

- prior knowledge can be difficult to specify
- can lead to worse estimates (relative to classical methods) if prior knowledge is inaccurate
- in practice, the choice of prior can be dictated by tractability considerations (e.g., conjugate priors are usually preferred regardless of how appropriate they are)

Bayesian Estimation

The goal of Bayesian estimation is the same as classical estimation: given an observation $\underline{x}$, estimate a specific value $\underline{\theta}$.

Unfortunately, the convention (which I will adhere to) is to refer to the random parameter *and* its realizations as $\underline{\theta}$.

So, given $\underline{x}$, we want to estimate the specific realization $\underline{\theta}$ that describes the model $f(\underline{x} \mid \underline{\theta})$.

Loss functions and Risk

The quality of an *estimate* is measured by a *loss* (or *cost*) function

$$L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x}))$$

For example, the quadratic loss (squared error) is

$$L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) = (\underline{\theta} - \hat{\underline{\theta}}(\underline{x}))^T (\underline{\theta} - \hat{\underline{\theta}}(\underline{x}))$$

$$= \| \underline{\theta} - \hat{\underline{\theta}}(\underline{x}) \|^2$$

The quality of an estimator is measured by the expected loss, known as the (Bayes) risk:

$$R(\hat{\underline{\theta}}) = E_{\underline{X},\underline{\theta}}\left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{X})) \right\}$$

Note: The expectation is with respect to both $\underline{X}$ and $\underline{\theta}$. For example, if $\underline{X}$ & $\underline{\theta}$ are jointly continuous, then

$$R(\hat{\underline{\theta}}) = \int\int L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) f(\underline{x}, \underline{\theta}) \, d\underline{x} \, d\underline{\theta}$$

$$= \int\int L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) f(\underline{x}|\underline{\theta}) f(\underline{\theta}) \, d\underline{x} \, d\underline{\theta}$$

In general, Bayesian estimation seeks the estimator

$$\hat{\underline{\theta}} = \arg\max_{\underline{\phi}} R(\phi)$$

minimizing the Bayes risk. The optimal estimator will depend on the statistical model and the loss.

In fact, the optimal estimator may be expressed solely in terms of the loss $L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x}))$ and the posterior $f(\underline{\theta} \mid \underline{x})$. To see this, note

$$R(\hat{\underline{\theta}}) = E_{\underline{X},\underline{\theta}} \left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{X})) \right\}$$

$$= E_{\underline{X}} \left\{ E_{\underline{\theta}|\underline{X}} \left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) \mid \underline{X} = \underline{x} \right\} \right\}$$

Thus, to minimize the risk, $\hat{\underline{\theta}}(\underline{x})$ must minimize

$$E_{\underline{\theta}|\underline{X}} \left\{ L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) \mid \underline{X} = \underline{x} \right\}$$

for each $\underline{x}$.

Posterior expected loss: depends only on loss & posterior

Said another way, the optimal estimator is

$$\hat{\underline{\theta}}(\underline{x}) =$$

(a)

Let's apply this result to some specific loss functions.

Minimum Mean Squared Error

In classical estimation, where only $\underline{X}$ is random, the MSE criterion

$$E_{\underline{X}}\left\{ (\underline{\theta} - \hat{\underline{\theta}})^T (\underline{\theta} - \hat{\underline{\theta}}) \right\}$$

did not lead to a practical estimator. In the Bayesian setup, the situation is different.

We define the <u>Bayesian MSE</u> to be the Bayes risk when the loss function is the squared error:

$$\text{BMSE}(\hat{\underline{\theta}}) := E_{\underline{X},\underline{\theta}}\left\{ (\underline{\theta} - \hat{\underline{\theta}})^T (\underline{\theta} - \hat{\underline{\theta}}) \right\}$$

The estimator that minimizes $\text{BMSE}(\hat{\underline{\theta}})$ is called the <u>minimum mean squared error</u> (MMSE) estimator

As we just saw, the MMSE estimator must minimize

$$E_{\underline{\theta}|\underline{X}}\left\{ (\underline{\theta} - \hat{\underline{\theta}}(\underline{x}))^T (\underline{\theta} - \hat{\underline{\theta}}(\underline{x})) \,\middle|\, \underline{X} = \underline{x} \right\}$$

for each $\underline{x}$.

Observe

$$E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-\hat{\underline{\theta}}(\underline{x}))^T(\underline{\theta}-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

$$= E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}]+E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))^T(\underline{\theta}-E[\underline{\theta}|\underline{x}]+E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

$$= E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}])^T(\underline{\theta}-E[\underline{\theta}|\underline{x}])\,|\,\underline{x}\right\}$$

$$+ 2E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}])^T(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

$$+ E_{\underline{\theta}|\underline{x}}\left\{(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))^T(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))\,|\,\underline{x}\right\}$$

First term: independent of $\hat{\underline{\theta}}(\underline{x})$

Second term: $2(E[\underline{\theta}|\underline{x}]-\hat{\underline{\theta}}(\underline{x}))^T \cdot \underbrace{E_{\underline{\theta}|\underline{x}}\left\{(\underline{\theta}-E[\underline{\theta}|\underline{x}])\,|\,\underline{x}\right\}}_{=\underline{0}}$

Third term: minimized by taking

$$\hat{\underline{\theta}}(\underline{x}) = E[\underline{\theta}|\underline{x}]$$

$$= \int \underline{\theta}\, f(\underline{\theta}|\underline{x})\, d\underline{\theta}$$

(b)

$$=$$

Exercise | Suppose $X \sim \text{Bin}(N, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. Find the MMSE estimator of θ. Prove the formula for the mean of a Beta random variable.

Solution | Earlier we saw

$$f(\theta|x) = \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}}{B(\alpha+x, \beta+N-x)}$$

$$\Rightarrow \hat{\theta}(x) = E[\underline{\theta}|\underline{x}] = \int \theta\, f(\theta|x)\, d\theta$$

$$= \frac{1}{B(\alpha+x,\beta+N-x)} \int \theta^{\alpha+x}(1-\theta)^{\beta+N-x-1}\, d\theta$$

$$= \frac{B(\alpha+x+1, \beta+N-x)}{B(\alpha+x, \beta+N-x)}$$

$$= \frac{\Gamma(\alpha+x+1)\,\Gamma(\beta+N-x)}{\Gamma(\alpha+\beta+N+1)} \cdot \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+x)\cdot\Gamma(\beta+N-x)}$$

$$= \frac{\alpha+x}{\alpha+\beta+N}$$

$$\boxed{\frac{\Gamma(z+1)}{\Gamma(z)} = z}$$

Minimum Mean Absolute Error

For a scalar parameter θ define the absolute error loss

$$L(\theta, \hat{\theta}(\underline{x})) = |\theta - \hat{\theta}(\underline{x})|.$$

The posterior expected loss may be written

$$E_{\theta|\underline{x}}\left[L(\theta, \hat{\theta}(\underline{x})) \mid \underline{x} \right]$$

$$= \int_{-\infty}^{\infty} |\theta - \hat{\theta}(\underline{x})| \, f(\theta \mid \underline{x}) \, d\theta$$

integration by parts

$$= \int_{-\infty}^{\hat{\theta}(\underline{x})} (\hat{\theta}(\underline{x}) - \theta) f(\theta|\underline{x}) d\theta + \int_{\hat{\theta}(\underline{x})}^{\infty} (\theta - \hat{\theta}(\underline{x})) f(\theta|\underline{x}) d\theta$$

$$= \int_{-\infty}^{\hat{\theta}(\underline{x})} F(\theta \mid \underline{x}) \, d\theta + \int_{\hat{\theta}(\underline{x})}^{\infty} (1 - F(\theta \mid \underline{x})) \, d\theta$$

where $F(\theta \mid \underline{x})$ is the posterior cumulative distribution function of θ given $\underline{x}$.

To minimize this expression let's take the derivative w.r.t $\hat{\theta}(\underline{x})$:

$$F(\hat{\theta}(\underline{x}) \mid \underline{x}) - (1 - F(\hat{\theta}(\underline{x}) \mid \underline{x})) = 0$$

$$\Rightarrow F(\hat{\theta}(\underline{x}) \mid \underline{x}) = \frac{1}{2}$$

$\Rightarrow \hat{\theta}(\underline{x})$ is the _____

Minimum Mean Uniform Error Estimation

The <u>uniform error loss</u> is defined to be

$$L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) = I_{\{\|\underline{\theta} - \hat{\underline{\theta}}(\underline{x})\| > \epsilon\}}$$

$$= \begin{cases} 1 & \text{if } \|\underline{\theta} - \hat{\underline{\theta}}(\underline{x})\| > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

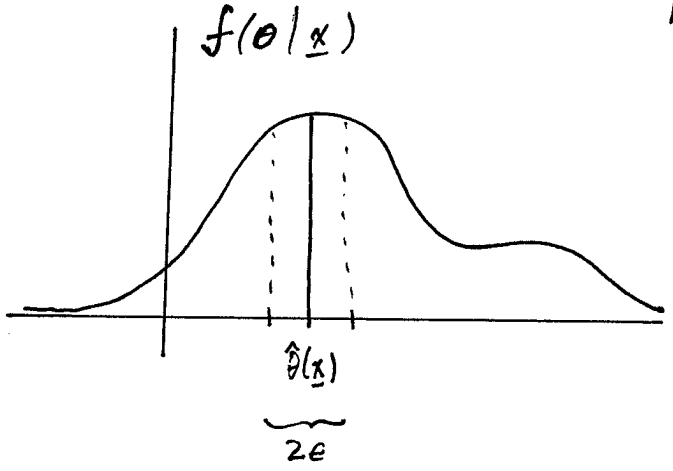
where $\epsilon > 0$ is small.

The posterior expected loss is

$$E\{L(\underline{\theta}, \hat{\underline{\theta}}(\underline{x})) \mid \underline{x}\} = P(\|\underline{\theta} - \hat{\underline{\theta}}(\underline{x})\| > \epsilon \mid \underline{x}),$$

the posterior probability that $\underline{\theta}$ deviates from $\hat{\underline{\theta}}(\underline{x})$ by more than ϵ. This probability is minimized, for ϵ sufficiently small, when $\hat{\underline{\theta}}(\underline{x})$ is near the _____.

(d)



Taking the limit as $\epsilon \to 0$ gives rise to the

maximum a posteriori (MAP) estimator:

$$\begin{aligned}
\hat{\underline{\theta}}(\underline{x}) &= \arg\max_{\underline{\theta}} \; f(\underline{\theta} \mid \underline{x}) \\
&= \arg\max_{\underline{\theta}} \; \frac{f(\underline{x} \mid \underline{\theta})\, f(\underline{\theta})}{f(\underline{x})} \\
&= \arg\max_{\underline{\theta}} \; f(\underline{x} \mid \underline{\theta}) \cdot f(\underline{\theta}).
\end{aligned}$$

This last expression is often easiest to compute: It avoids having to determine $f(\underline{x})$ or $f(\underline{\theta} \mid \underline{x})$.

Exercise | Suppose $X \sim \text{Bin}(N, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$. Find the MAP estimator of θ.

Solution | Recall

$$f(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}$$

$$\Rightarrow \log f(\theta|x) = (\alpha+x-1)\log\theta + (\beta+N-x-1)\log(1-\theta) + C$$

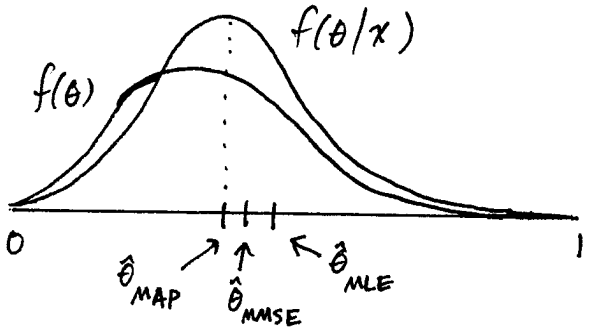
$$\Rightarrow \frac{\partial \log f(\theta|x)}{\partial\theta} = \frac{\alpha+x-1}{\theta} - \frac{\beta+N-x-1}{1-\theta} = 0$$

$$\Rightarrow \hat{\theta}_{MAP}(x) = \frac{\alpha+x-1}{\alpha+\beta+N-2}$$

Compare:

$$\hat{\theta}_{MMSE}(x) = \frac{\alpha+x}{\alpha+\beta+N}$$

$$\hat{\theta}_{MLE}(x) = \frac{x}{N}$$

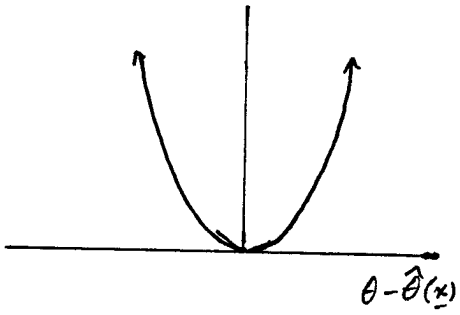


Note that both Bayesian estimators tend to the classical estimator as $N \to \infty$; a flood of data can overwhelm any amount of prior knowledge.

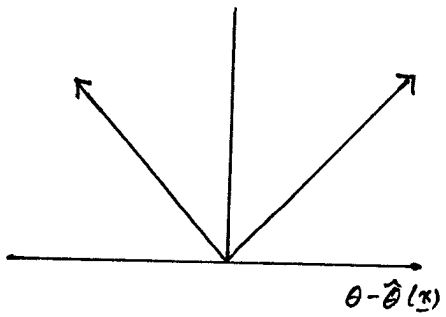
Note: the posterior median requires a closed form expression for the posterior CDF which is not available.

Discussion and Summary

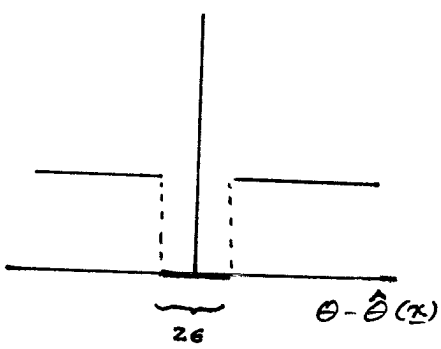
squared error



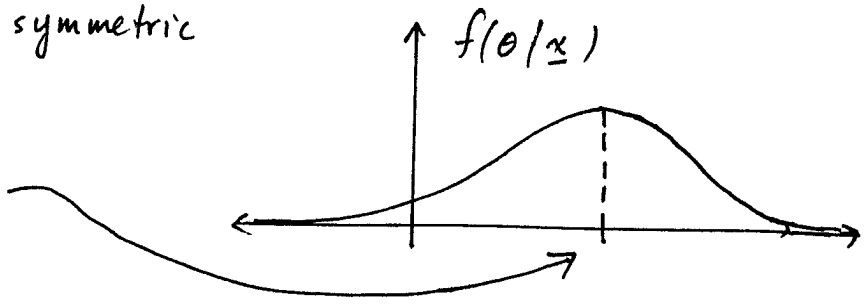
absolute error



uniform error



- These are the primary three loss functions used
- Estimate depends on $\underline{x}$ through the posterior:
  posterior mean, posterior median, and posterior mode
- If the posterior is symmetric and unimodal, then
  $\hat{\theta}_{MMSE} = \hat{\theta}_{MMAE} = \hat{\theta}_{MAP}$



- Limiting cases:
  - $N \to \infty \implies$ Bayesian $\to$ classical
  - prior $\to$ least informative $\not\Rightarrow$ Bayesian $\to$ classical
- $\hat{\theta}_{MMAE}$ does not generalize well to vector $\underline{\theta}$.
- Both $\hat{\theta}_{MMSE}$ and $\hat{\theta}_{MMAE}$ require integrating w.r.t. $f(\underline{\theta}|\underline{x})$. Often this calculation will be intractable. How can we approximate these estimators numerically?

If we can simulate $\underline{\theta}_1, \ldots, \underline{\theta}_M$ from $f(\underline{\theta} | \underline{x})$, then we can apply the following Monte Carlo estimates:

$$\hat{\theta}_{MMSE}(\underline{x}) \approx \frac{1}{M} \sum_{i=1}^{M} \theta_i$$

$$\hat{\theta}_{MMAE}(\underline{x}) \approx \text{median}\{\theta_1, \ldots, \theta_M\}$$

- If the posterior mode cannot be determined analytically, then many of the numerical approaches for MLE can be applied.
- Which of the three loss functions to use is often dictated by computational considerations

Key

a. $\hat{\underline{\theta}}(\underline{x}) = \arg\min_{\underline{\phi}} E_{\underline{\theta}|\underline{X}}\{L(\underline{\theta}, \underline{\phi}) \mid \underline{X} = \underline{x}\}$

b. posterior mean

c. posterior median

d. posterior mode

BAYESIAN ESTIMATION: THE GAUSSIAN LINEAR MODEL

Consider the Bayesian statistical model

$$\underline{X} = H \cdot \underline{\theta} + \underline{W}$$

where

$\underline{\theta}$ is unknown, $p \times 1$

H is known, $N \times p$

$\underline{\theta} \sim \mathcal{N}(\underline{\mu}_\theta, R_\theta)$

$\underline{W} \sim \mathcal{N}(\underline{0}, R_w)$

$\underline{\theta}$ and $\underline{W}$ are independent

$R_\theta, R_w, \underline{\mu}_\theta$ are known.

This model amounts to a signal subspace with a Gaussian prior on $\underline{\theta}$ and a Gaussian conditional distribution of $\underline{X}$ given $\underline{\theta}$.

This formulation is quite general and encompasses many interesting and important examples.

Example Suppose $\underline{X} = \underline{S} + \underline{W}$ where

$$s(n) = \cos(2\pi f n + \phi), \quad n = 0, 1, \ldots, N-1$$

and $-\frac{L}{N} \leq f \leq \frac{L}{N}$. On the homework we have seen that it is possible to approximate $\underline{S} = H\underline{\theta}$ where the dimension of $\underline{\theta}$ is $p = 2L+1$, and $\underline{\theta}$ follows a Gaussian distribution.

Result The posterior distribution of $\underline{\theta}/\underline{x}$ is

$$\underline{\theta}/\underline{x} \sim \mathcal{N}(\underline{\mu}_{\theta/x}, R_{\theta/x})$$

where

$$\underline{\mu}_{\theta/x} = \underline{\mu}_\theta + R_\theta H^T (H R_\theta H^T + R_w)^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

$$R_{\theta/x} = R_\theta - R_\theta H^T (H R_\theta H^T + R_w)^{-1} H R_\theta$$

Proof $\underline{X}$ and $\underline{\theta}$ are jointly Gaussian:

$$\begin{bmatrix} \underline{X} \\ \underline{\theta} \end{bmatrix} = \begin{bmatrix} H & I_N \\ I_p & 0 \end{bmatrix} \begin{bmatrix} \underline{\theta} \\ \underline{w} \end{bmatrix}$$

where

$$\begin{bmatrix} \underline{\theta} \\ \underline{w} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \underline{\mu}_\theta \\ \underline{0} \end{bmatrix}, \begin{bmatrix} R_\theta & 0 \\ 0 & R_w \end{bmatrix} \right)$$

$$\Rightarrow \begin{bmatrix} \underline{X} \\ \underline{\theta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} H\underline{\mu}_\theta \\ \underline{\mu}_\theta \end{bmatrix}, \begin{bmatrix} H R_\theta H^T + R_w & H R_\theta \\ R_\theta H^T & R_\theta \end{bmatrix} \right)$$

Now apply the Gaussian conditioning principle.

It can be shown using the matrix inversion lemma that

$$\underline{\mu}_{\theta|x} = \underline{\mu}_{\theta} + R_{\theta} H^T (H R_{\theta} H^T + R_w)^{-1} (\underline{x} - H\underline{\mu}_{\theta})$$

$$= \underline{\mu}_{\theta} + (H^T R_w^{-1} H + R_{\theta}^{-1})^{-1} H^T R_w^{-1} (\underline{x} - H\underline{\mu}_{\theta})$$

and

$$R_{\theta|x} = R_{\theta} - R_{\theta} H^T (H R_{\theta} H^T + R_w)^{-1} H R_{\theta}$$

$$= (H^T R_w^{-1} H + R_{\theta}^{-1})^{-1}$$

These alternative formulas are sometimes more convenient to work with.

To verify these formulas is a tedious but manageable exercise

Estimation

The posterior distribution is Gaussian, which is symmetric and unimodal. Therefore, the optimal estimator (minimizing the Bayes risk) is

$$\hat{\underline{\theta}}(\underline{x}) = \underline{\mu}_{\theta|x} = \underline{\mu}_\theta + R_\theta H^T (H R_\theta H^T + R_\omega)^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

$$= \underline{\mu}_\theta + (H^T R_\omega^{-1} H + R_\theta^{-1})^{-1} H^T R_\omega^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

regardless of the loss function.

Observations

1. $\hat{\underline{\theta}}(\underline{x})$ is an affine function of $\underline{x}$.
2. $\hat{\underline{\theta}}(\underline{X})$ is again multivariate Gaussian.
3. Consider the case where $R_\theta = \sigma^2 I_p$ and $\sigma^2 \to \infty$. This can be thought of as a "non committal" prior. Then $R_\theta^{-1} \to 0_p$ and

$$\hat{\underline{\theta}}(\underline{x}) = \underline{\mu}_\theta + (H^T R_\omega^{-1} H)^{-1} H^T R_\omega^{-1} (\underline{x} - H\underline{\mu}_\theta)$$

$$= (H^T R_\omega^{-1} H)^{-1} H^T R_\omega^{-1} \underline{x}$$

$$= \text{MLE} / \text{MVUE}$$

Exercise Suppose we observe

$$X_i = A + W_i \quad , \quad i = 1, \ldots, N$$

where A is an unknown scalar and

$$\left.\begin{aligned} A &\sim N(\mu_A, \sigma_A^2) \\ W_i &\overset{iid}{\sim} N(0, \sigma_w^2) \end{aligned}\right\} \text{independent}$$

with $\mu_A, \sigma_A^2, \sigma_w^2$ known. Find the Bayesian estimate $\hat{A}$. Interpret your result. Analyze limiting cases.

Solution / The problem falls within the linear model with

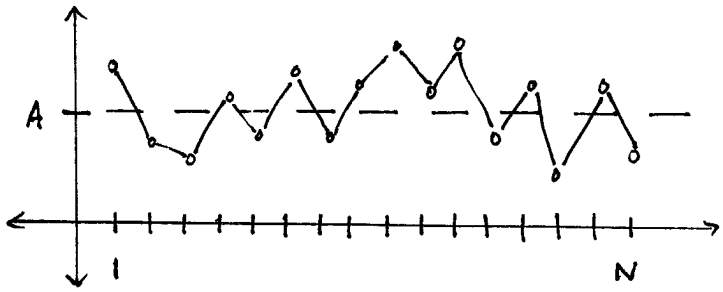
$$H = \underline{1} \quad (N \times 1)$$

$$\theta = A \quad (1 \times 1)$$

$$\mu_\theta = \mu_A \quad (1 \times 1)$$

$$R_\theta = \sigma_A^2 \quad (1 \times 1)$$

$$R_w = \sigma^2 I_N \quad (N \times N)$$



Using the second formula for $\mu_{A/x}$ (the one that comes from the matrix inversion lemma) we obtain

$$\hat{A}(\underline{x}) = \mu_{A/x} = \mu_A + \left(\underline{1}^T \cdot \underline{1} \cdot \frac{1}{\sigma_w^2} + \frac{1}{\sigma_A^2}\right)^{-1} \underline{1}^T \cdot \frac{1}{\sigma_w^2} (\underline{x} - \underline{1}\mu_A)$$

$$= \mu_A + \left(\frac{N}{\sigma_w^2} + \frac{1}{\sigma_A^2}\right)^{-1} \frac{1}{\sigma_w^2} \left(\sum x_i - N\mu_A\right)$$

$$= \mu_A + \frac{1}{\frac{N}{\sigma_w^2} + \frac{1}{\sigma_A^2}} \cdot \frac{N}{\sigma_w^2} \left(\bar{x} - \mu_A\right)$$

$$= \mu_A + \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}} \left(\bar{x} - \mu_A\right)$$

Thus

$$\hat{A}(\underline{x}) = (1-\alpha)\mu_A + \alpha \cdot \bar{x}$$

where

$$\alpha = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}}$$

controls the tradeoff between prior knowledge and data.

Limiting cases:

(a)

$N \to \infty \Rightarrow \alpha \to \quad \Rightarrow \hat{A} \to$

$N = 0 \Rightarrow \alpha = \quad \Rightarrow \hat{A} =$

$\sigma_A^2 \to \infty \Rightarrow \alpha \to \quad \Rightarrow \hat{A} \to$

$\sigma_A^2 \to 0 \Rightarrow \alpha \to \quad \Rightarrow \hat{A} \to$

It suffices to focus on the case $\underline{\mu}_\theta = \underline{0}$. Then the Bayesian estimator is

$$\hat{\underline{\theta}}(\underline{x}) = \underline{\mu}_{\theta|x} = R_\theta H^T \left(H R_\theta H^T + R_w \right)^{-1} \underline{x}$$

$$= \left(H^T R_w^{-1} H + R_\theta^{-1} \right)^{-1} H^T R_\theta^{-1} \underline{x}$$

If ever $\underline{\mu}_\theta \neq \underline{0}$, we may apply the above estimator to $\underline{x} - H\underline{\mu}_\theta$ and add $\underline{\mu}_\theta$ to the result.

Simultaneously Diagonalizable Covariance Matrices

Consider the problem of estimating a signal in additive Gaussian noise

$$\underline{x} = \underline{s} + \underline{w}$$

where

$\underline{x}$ = observed noisy signal

$\underline{s}$ = clean signal

$\underline{w}$ = noise

This can be modeled using the general linear model with

$$\underline{\theta} = \underline{s}$$

$$H = I_N$$

and adopting a Gaussian prior for $\underline{s}$:

$$\underline{s} \sim N(\underline{0}, R_{ss}).$$

The Bayesian estimate for $\underline{s}$ is

$$\hat{\underline{s}} =$$

Now suppose that R_{ss} and R_{ww} are simultaneously diagonalizable, meaning $\exists$ an orthogonal matrix U such that

$$R_{ss} = U \Lambda_s U^T$$

and

$$R_{ww} = U \Lambda_w U^T$$

with Λ_s, Λ_w diagonal.

Example $R_{ww} = \sigma^2 I_N$ and R_{ss} is arbitrary

Then the estimator becomes

$$\hat{\underline{s}} = R_{ss}\left(R_{ss} + R_{ww}\right)^{-1}\underline{x}$$

$$= U\Lambda_s U^T\left(U\Lambda_s U^T + U\Lambda_w U^T\right)^{-1}\underline{x}$$

$$= U\Lambda_s U^T\left(U\left[\Lambda_s + \Lambda_w\right]U^T\right)^{-1}\underline{x}$$

$$= U\cdot\underbrace{\left[\Lambda_s\left(\Lambda_s+\Lambda_w\right)^{-1}\right]}_{\Lambda}U^T\underline{x}$$

where

$$\Lambda = \begin{bmatrix} \frac{\lambda_1^s}{\lambda_1^s+\lambda_1^w} & & \\ & \frac{\lambda_2^s}{\lambda_2^s+\lambda_2^w} & \\ & & \ddots \\ & & & \frac{\lambda_N^s}{\lambda_N^s+\lambda_N^w} \end{bmatrix}$$

Interpretation:

U = change of basis matrix

$\underline{y} = U^T x$: coefficients of $\underline{x}$ in new basis

$\underline{z} = \Lambda \underline{y}$: coordinate-wise rescaling of $\underline{y}$

$\hat{\underline{s}} = U\underline{z}$: reconstruction of $\underline{s}$ from $\underline{z}$

How should we interpret the weights

$$\lambda_i = \frac{\lambda_i^s}{\lambda_i^s + \lambda_i^w} ?$$

Notice that $U^T \underline{x} = U^T \underline{s} + U^T \underline{w}$ and

$$U^T \underline{s} \sim N(\underline{0}, U^T R_{ss} U) = N(\underline{0}, \Lambda_s)$$

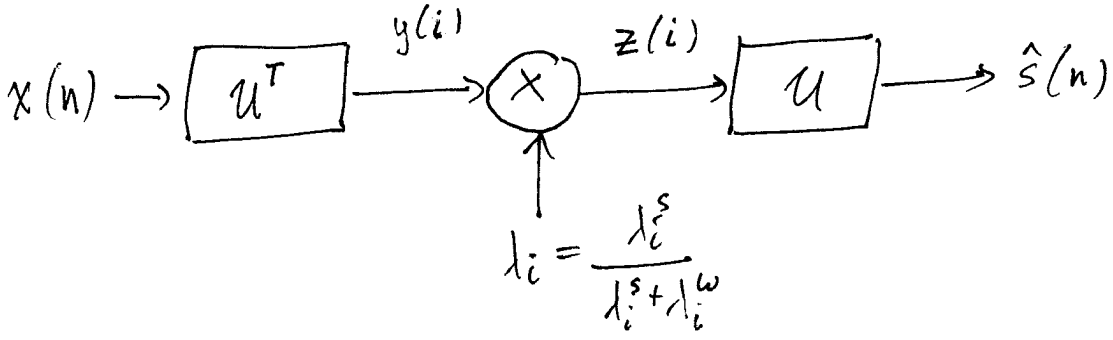
$$U^T \underline{w} \sim N(\underline{0}, U^T R_{ww} U) = N(\underline{0}, \Lambda_w)$$

Writing $U = [\underline{u}_1 \ \cdots \ \underline{u}_N]$ we have

$$\underline{u}_i^T \underline{s} \sim N(0, \lambda_i^s)$$

$$\underline{u}_i^T \underline{w} \sim N(0, \lambda_i^w)$$

Thus, λ_i reflects the proportion of the projection onto $\underline{u}_i$ that is due to the signal.



"analysis → processing → synthesis"

Application: Bandpass Filtering

Suppose we observe

$$\underline{x} = \underline{s} + \underline{w}$$

and we know a priori that the signal of interest occupies a certain passband.

In other words, $|\underline{u}_k^H \underline{X}|$ is large on average for certain DFT basis vectors $\underline{u}_k$, and small for others.

How can we incorporate this prior knowledge into the prior for $\underline{s}$? In other words, what should we take for R_{ss}?

Let us assume we can specify

$$\sigma_k^2 = E\left\{ |\underline{u}_k^H \underline{S}|^2 \right\},$$

the average signal energy at frequency k/N.

Let's also assume that signal content at different frequencies are independent.

In essence we are taking

$\mathcal{U}$ = DFT matrix

$$\mathcal{U}^H \underline{S} \sim N(\underline{0}, \Sigma)$$

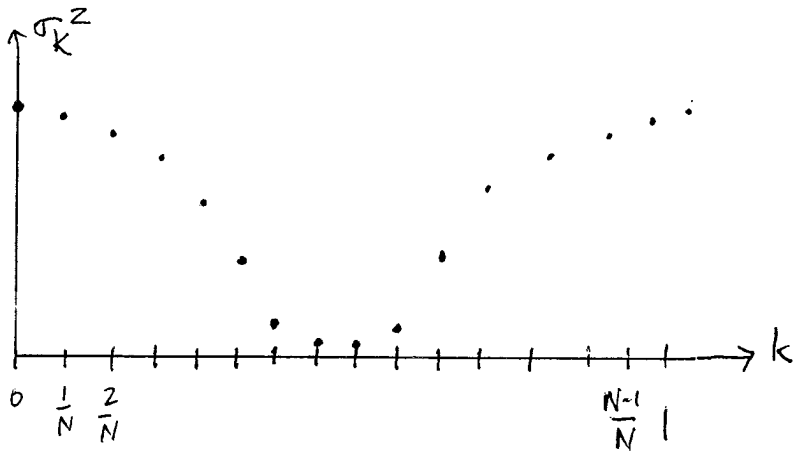
as a prior, where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_N^2 \end{pmatrix}.$$

Equivalently, the prior on $\underline{S}$ is

$$\underline{S} \sim N(\underline{0}, \underbrace{\mathcal{U} \Sigma \mathcal{U}^H}_{R_{SS}})$$

For example



Note: $\sigma_k^2 = \sigma_{N-k}^2$ by conjugate symmetry

is a lowpass model.

Notice that the energy of $\underline{s}$ is

$$E\{\underline{s}^T\underline{s}\} = E\{(\mathcal{U}^H\underline{s})^H(\mathcal{U}^H\underline{s})\}$$

$$= \sum_{k=0}^{N-1} \sigma_k^2$$

So to specify the σ_k^2 it suffices to know the signal energy and the <u>shape</u> of the frequency response.

Assume the noise is IID:

$$R_{ww} = \sigma^2 I_N,$$

σ^2 known. Then the MMSE estimator is

$$\hat{\underline{s}} = R_{ss}(R_{ss} + R_{ww})^{-1}\underline{x}$$

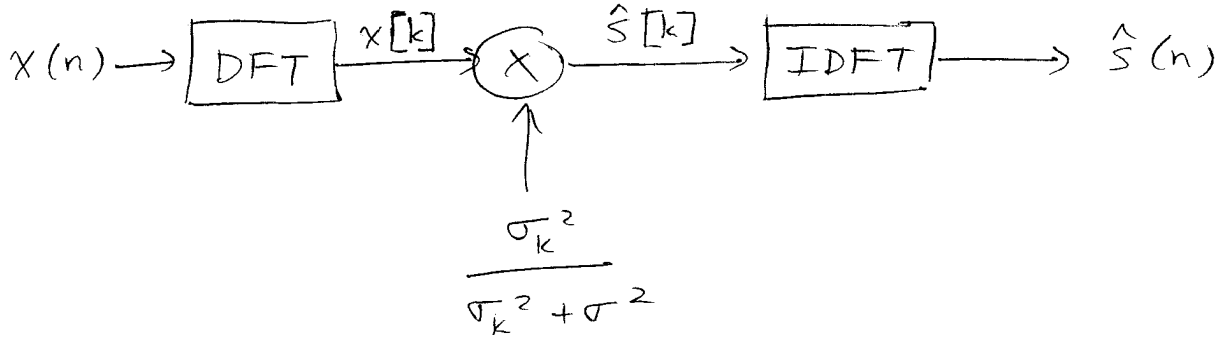
$$= \mathcal{U}\Sigma\mathcal{U}^H(\mathcal{U}[\Sigma + \sigma^2 I]\mathcal{U}^H)^{-1}\underline{x}$$

$$= \mathcal{U}[\Sigma(\Sigma + \sigma^2 I)^{-1}]\mathcal{U}^H\underline{x}$$

Note that

$$\Sigma(\Sigma+\sigma^2 I)^{-1} = \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\sigma_0^2} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2+\sigma_0^2} & & \\ & & \ddots & \\ & & & \frac{\sigma_N^2}{\sigma_N^2+\sigma^2} \end{bmatrix}$$

Therefore, the estimator is a bandpass filter



Interpretation:

- $\sigma_k^2 \gg \sigma^2 \Longrightarrow$ keep most of signal
- $\sigma_k^2 \ll \sigma^2 \Longrightarrow$ kill most of signal
- $\sigma_k^2 \approx \sigma^2 \Longrightarrow$ keep some of signal

Summary

- Extension of signal subspace model to Bayesian setting
- When subspace coefficients (prior) and observation noise (likelihood) are jointly Gaussian, posterior is also Gaussian (conjugate prior)
- Posterior mean (mode) is a linear / affine function.
- Classical estimators fall out in limiting cases.
- When R_θ, R_w are simultaneously diagonalizable
  $\Rightarrow$ transform domain "shrinkage"
  e.g., bandpass filtering.

Key

a. $1, \bar{x}$ b. $R_{ss}(R_{ss} + R_{ww})^{-1}\underline{x}$

$0, \mu_A$

$1, \bar{x}$

$0, \mu_A$

APPLICATION: WAVELET DENOISING

The Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a linear map

$$W^T: \mathbb{R}^N \longrightarrow \mathbb{R}^N, \quad \underline{x} \longmapsto \underline{y} = W^T \underline{x}$$

satisfying certain special properties.

Although a thorough and rigorous definition and treatment of the DWT is beyond our scope, we can understand it through analogy with the discrete Fourier transform (DFT).

DWT vs. DFT

- Both can be represented by orthogonal matrices
- Both have efficient implementations

  DFT: $O(N \log N)$

  DWT: $O(N)$

- Both are discretizations of continuous transforms

continuous and infinite	→	discrete and infinite	→	discrete and finite
CFT		DTFT		DFT
CWT		DTWT		DWT

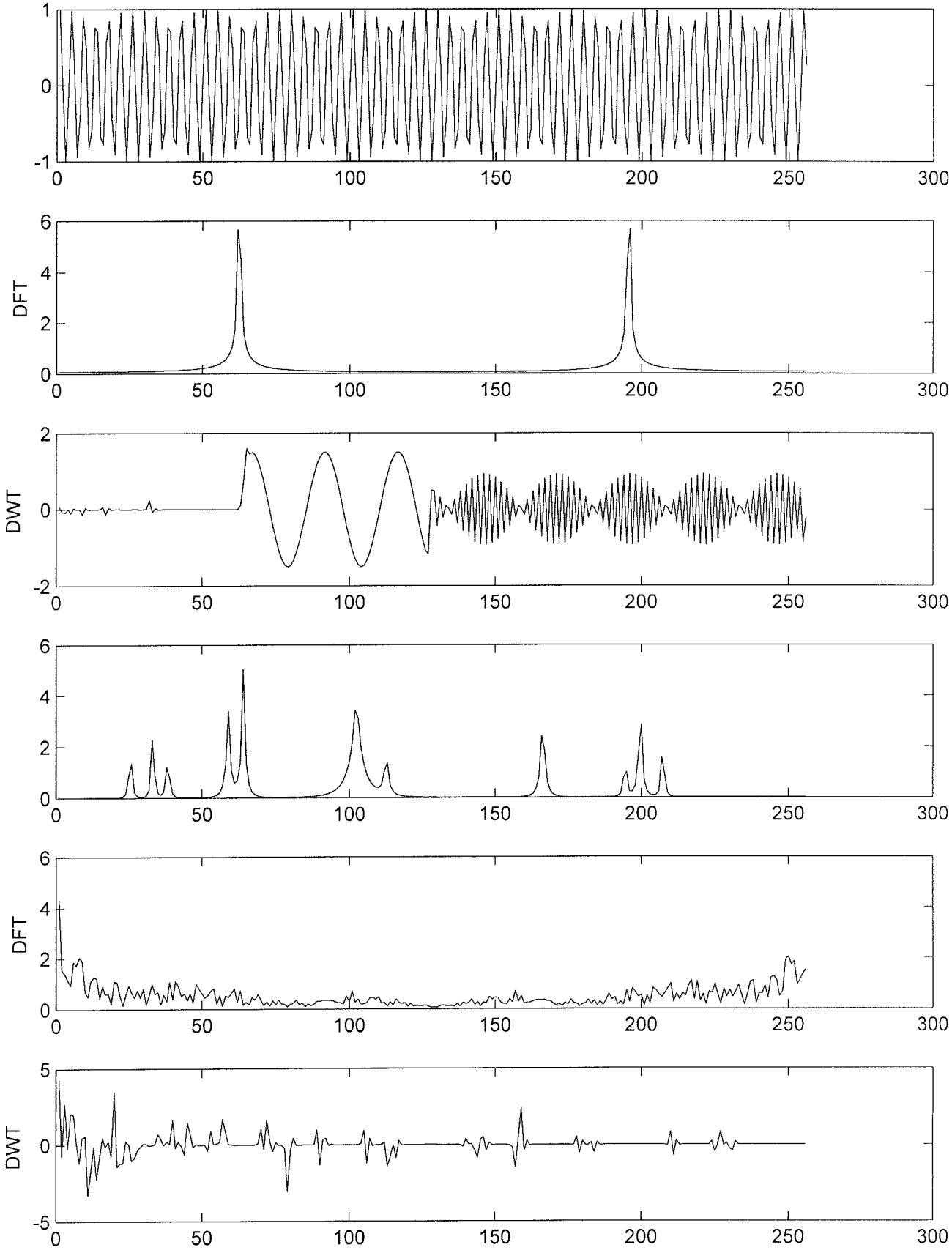
- Both are "change of basis" operators that compute the expansion coefficients of the signal in a different basis

DFT ⟹ Fourier basis

sparse representation of sinusoidal signals

DWT ⟹ Wavelet basis (columns of W)

sparse representation of piecewise polynomial signals


1
0
-1
0
50
100
150
200
250
300
DFT
6
4
2
0
DWT
2
-2
-5
5

Haar Wavelet Transform

While there's only one DFT, there are in fact many different DWTs. The simplest is the Haar wavelet transform.

Consider a signal of length $N = 2^L$.

$$\underline{x} = [x(1) \quad x(2) \quad \cdots \quad x(N)]^T.$$

Define

$$\underline{y}_1 = [c_1(1) \; c_1(2) \cdots c_1(\tfrac{N}{2}) \mid d_1(1) \cdots d_1(\tfrac{N}{2})]^T$$

where

$$c_1(1) = \frac{x(1)+x(2)}{\sqrt{2}} \qquad d_1(1) = \frac{x(1)-x(2)}{\sqrt{2}}$$

$$c_1(2) = \frac{x(3)+x(4)}{\sqrt{2}} \qquad d_1(2) = \frac{x(3)-x(4)}{\sqrt{2}}$$

$$\vdots$$

Observe

- This transformation is invertible: we can recover $\underline{x}$ from $\underline{y}_1$
- The transformation is an orthogonal linear map

$$\begin{bmatrix} c_1(1) \\ c_1(2) \\ c_1(3) \\ \vdots \\ d_1(1) \\ d_1(2) \\ d_1(3) \\ \vdots \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & & & & & \\ & & 1 & 1 & & & \\ & & & & 1 & 1 & \\ & & & & & & \ddots \\ 1 & -1 & & & & & \\ & & 1 & -1 & & & \\ & & & & 1 & -1 & \\ & & & & & & \ddots \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ \\ \\ \\ x(N) \end{bmatrix} = W^T \underline{x}$$

- The coefficients $c_1(n)$ are local *averages* and represent *coarse* information about the signal
- The coefficients $d_1(n)$ are local *differences* and represent *detailed* information

This operation is called the *level-1 Haar wavelet transform*. The idea behind the general Haar wavelet transform is to **recursively** apply this operation to the **coarse** coefficients.

$$\underline{y}_1 = \left[c_1(1) \quad \cdots \quad c_1\left(\tfrac{N}{2}\right) \,\middle|\, d_1(1) \quad \cdots \quad d_1\left(\tfrac{N}{2}\right) \right]^T$$

$$\underline{y}_2 = \left[c_2(1) \quad \cdots \quad c_2\left(\tfrac{N}{4}\right) \,\middle|\, d_2(1) \quad \cdots \quad d_2\left(\tfrac{N}{4}\right) \,\middle|\, d_1(1) \quad \cdots \quad d_1\left(\tfrac{N}{2}\right) \right]^T$$

$$\underline{y}_3 = \left[c_3(1) \cdots c_3\left(\tfrac{N}{8}\right) \,\middle|\, d_3(1) \cdots d_3\left(\tfrac{N}{8}\right) \,\middle|\, d_2(1) \quad \cdots \quad d_2\left(\tfrac{N}{4}\right) \,\middle|\, d_1(1) \quad \cdots \quad d_1\left(\tfrac{N}{2}\right) \right]^T$$

$$\vdots$$

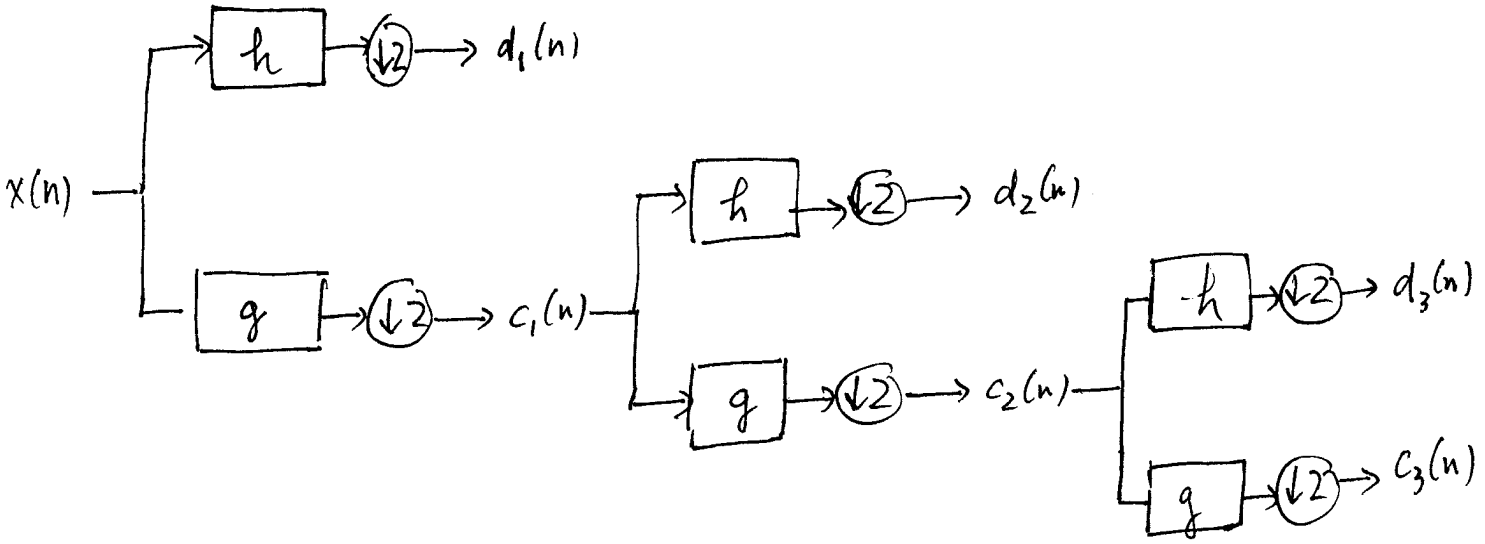
$$\underline{y}_L = \left[c_L(1) \,\middle|\, d_L(1) \,\middle|\, d_{L-1}(1) \;\; d_{L-1}(2) \,\middle|\, d_{L-2}(1) \quad \cdots \quad d_{L-2}(4) \,\middle|\, \cdots \right]^T$$

We call $\underline{y}_\ell$ the <u>level-ℓ Haar wavelet transform</u>

<u>Filter bank implementation</u>:

$$h = \left[\tfrac{1}{\sqrt{2}} \quad -\tfrac{1}{\sqrt{2}} \right] \quad \text{(high pass)}$$

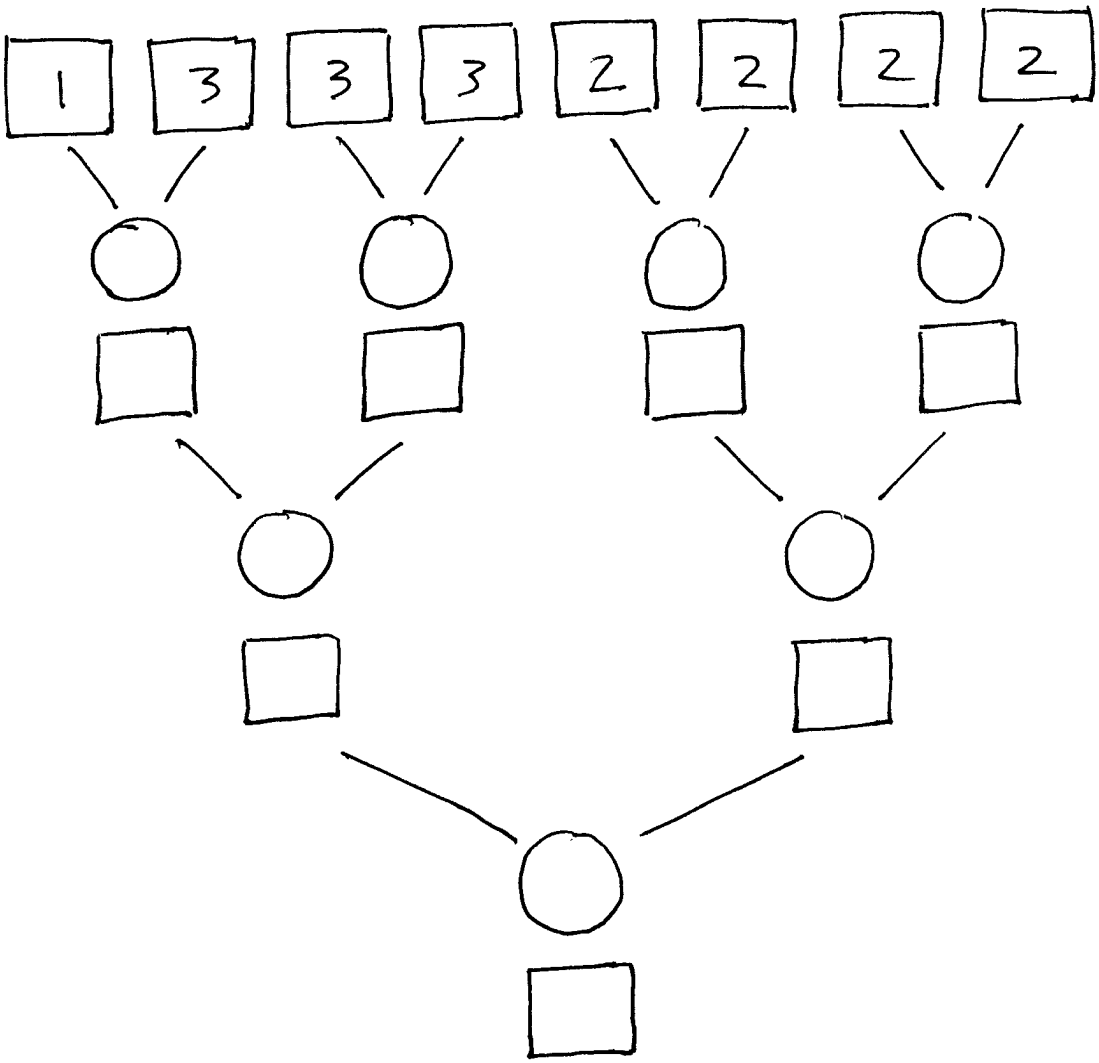
$$g = \left[\tfrac{1}{\sqrt{2}} \quad \tfrac{1}{\sqrt{2}} \right] \quad \text{(low pass)}$$



$\underline{y}_1$ $\underline{y}_2$ $\underline{y}_3$

Example

$\underline{x}$ =



$\underline{y}_1 = [\quad]^T$

$\underline{y}_2 = [\quad]^T$

$\underline{y}_3 = [\quad]^T$

Important things to notice:

- the detail coefficients are zero where the signal is constant. In particular, if $x(n)$ is constant on the interval

$$\left[k\cdot 2^{\ell}+1, k\cdot 2^{\ell}+2, \dots, (k+1)2^{\ell}\right], \text{ then } d_{\ell}(k) = 0$$

$\Rightarrow$ **sparsity**

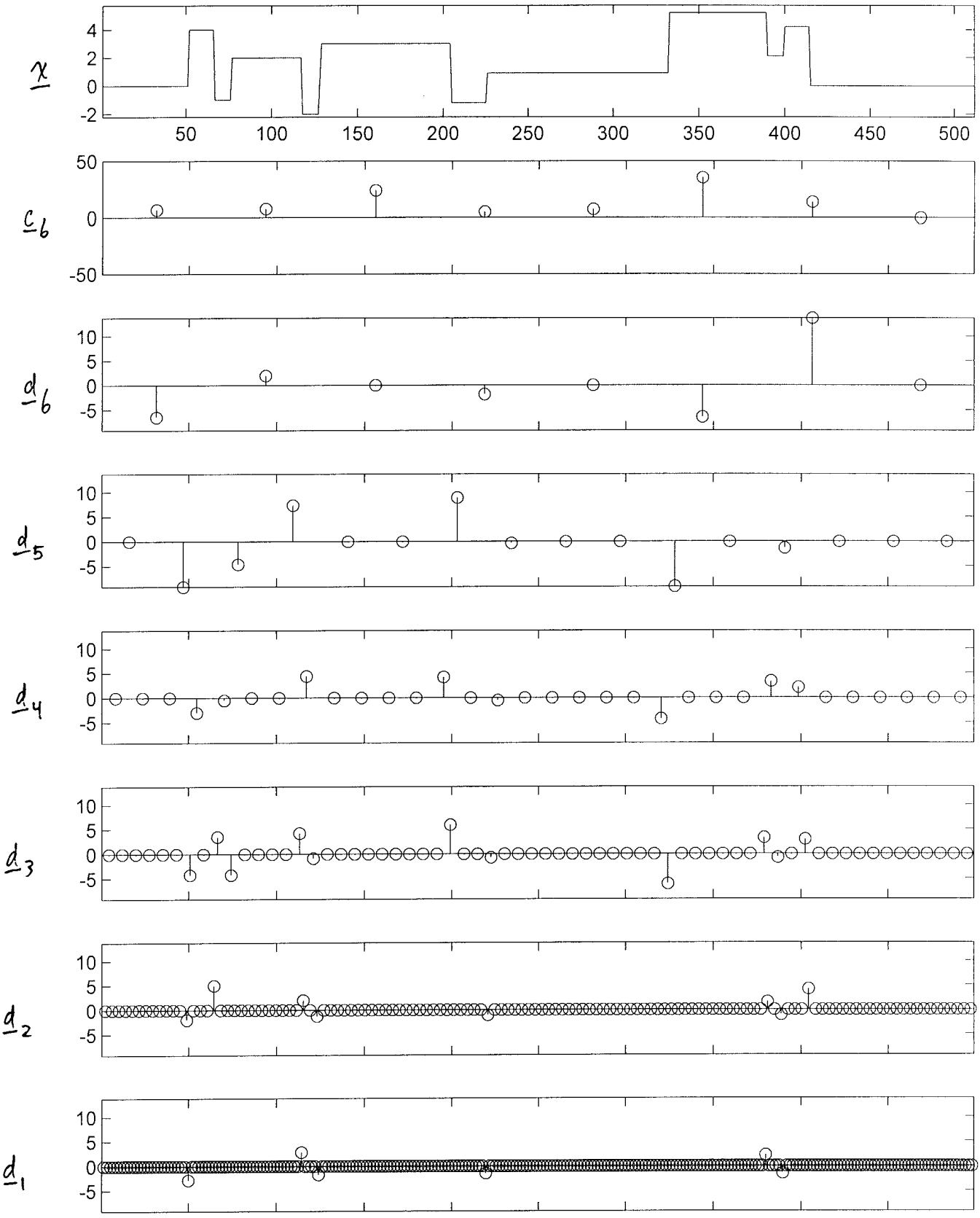
- the detail coefficients have a natural hierarchical (or tree-structured) arrangement; we can say that $d_{\ell}(k)$ is the <u>parent</u> of $d_{\ell-1}(2k-1)$ and $d_{\ell-1}(2k)$, who are its <u>children</u>
- $c_{\ell}(n)$ is a low resolution approximation to $x(n)$; it is the result of averaging and downsampling $x(n)$ ℓ times
- different levels capture different <u>resolutions</u> of detail:

$\{d_1(k)\}_{k=1}^{2^{L-1}}$ — high frequency detail

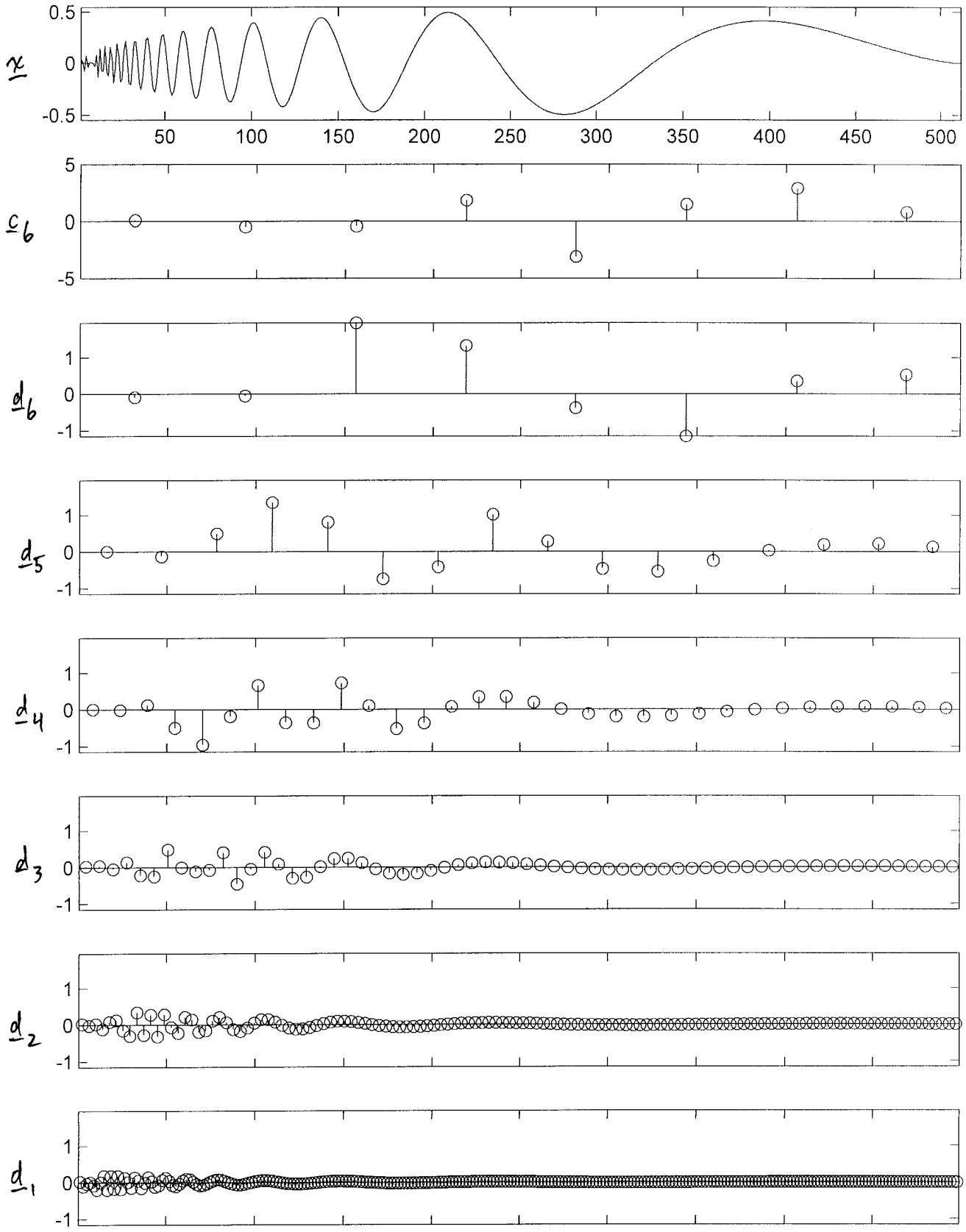
$\Downarrow$ $\Longleftrightarrow$ $\Downarrow$

$\{d_{L-1}(k)\}_{k=1}^{2}$

$\{d_L(k)\}_{k=1}^{1}$ — low frequency detail



$\Rightarrow$ sparse representation


x
0.5
0
-0.5
50
100
150
200
250
300
350
400
450
500
c6
5
0
-5
d6
1
0
-1
d5
1
0
-1
d4
1
0
-1
d3
1
0
-1
d2
1
0
-1
d1
1
0
-1

Other Wavelet Transforms

The Haar wavelet transform can be generalized by using different high-pass and low-pass filters h & g. These filters must satisfy certain properties for the resulting transform to be orthogonal (and qualify as a DWT).

The most important generalization are the Daubechies DWT. They are based on certain filters h_p, g_p such that

- the length of h_p and g_p is $2p$.
- if a signal behaves locally like a $(p-1)^{th}$ order polynomial, the corresponding detail coefficients are zero

Examples

$p = 1 \Longrightarrow$ Haar

$p = 2 \Rightarrow g = [.4830 \quad .8365 \quad .2241 \quad -.1294]$

$h = [.1294 \quad .2241 \quad -.8365 \quad .4830]$

Wavelet Denoising

Suppose we measure a noisy signal

$$\underline{x} = \underline{s} + \underline{v} \qquad (*)$$

and assume

- $\underline{s} = [s_1 \ \dots \ s_N]^T$ has a sparse representation in a certain wavelet basis, e.g., $\underline{s}$ is piecewise constant / Haar basis
- $\underline{v} \sim \mathcal{N}(\underline{0}, \sigma^2 I)$

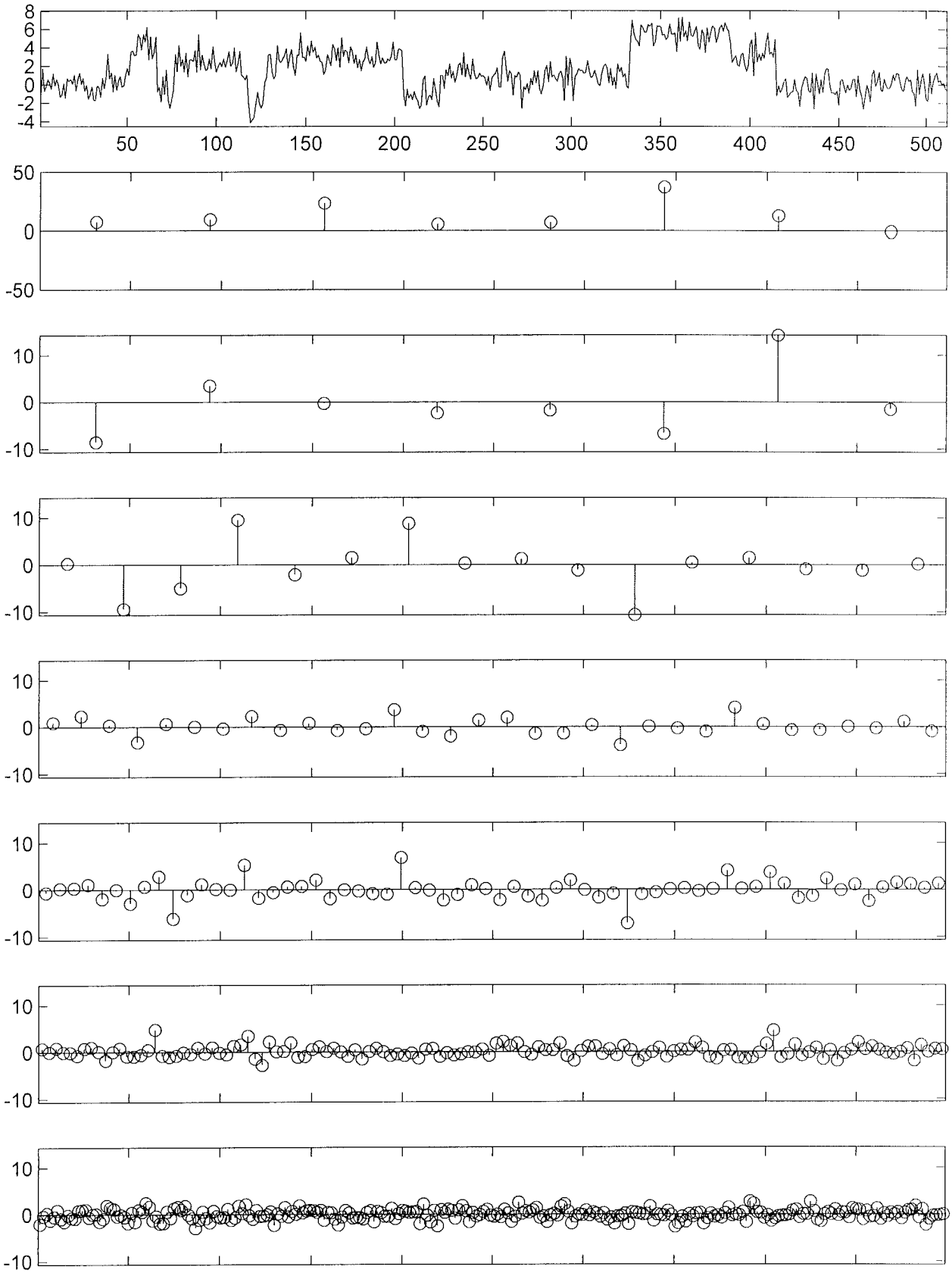
Now take the wavelet transform of (*):

$$\underline{y} = \underline{\theta} + \underline{z}$$

$$\underline{y} = W^T \underline{x}$$
$$\underline{\theta} = W^T \underline{s}$$
$$\underline{z} = W^T \underline{v}$$

Then we know

- most elements in $\underline{\theta} = [\theta_1, \dots, \theta_N]^T$ are zero or very close to zero
- $\underline{z} \sim \mathcal{N}(\underline{0}, \sigma^2 I)$ because $W^T W = I$


8
6
4
2
0
-2
-4
50
100
150
200
250
300
350
400
450
500
50
0
-50
10
0
-10
10
0
-10
10
0
-10
10
0
-10
10
0
-10
10
0
-10

Since W is orthogonal, the estimation problem amounts to recovery of a signal in iid Gaussian noise, whether we treat the problem in the "time" domain or the "wavelet" domain

From a Bayesian perspective, we should tackle the problem in the domain for which it is easiest to specify a prior.

On one hand, the fact that $\underline{\theta}$ is sparse suggest a subspace model. Unfortunately we don't know a priori <u>which</u> detail coefficients will be zero.

<u>Q</u>: How can we take advantage of the prior knowledge that $\underline{\theta}$ is sparse? What statistical model captures this information?

<u>A</u>: One solution is to employ a _____

Mixture Modeling

View the detail coefficients $\theta_2, \ldots, \theta_N$ as realizations of a single random variable θ.

We know

- Most θ_i are small (sparsity assumption)
- Some θ_i are large (W is orthogonal, so energy must be preserved)
- θ is zero mean, since θ_i are local differences
- θ_i are "approximately" independent, since the θ_i are local differences

This suggests the following prior:

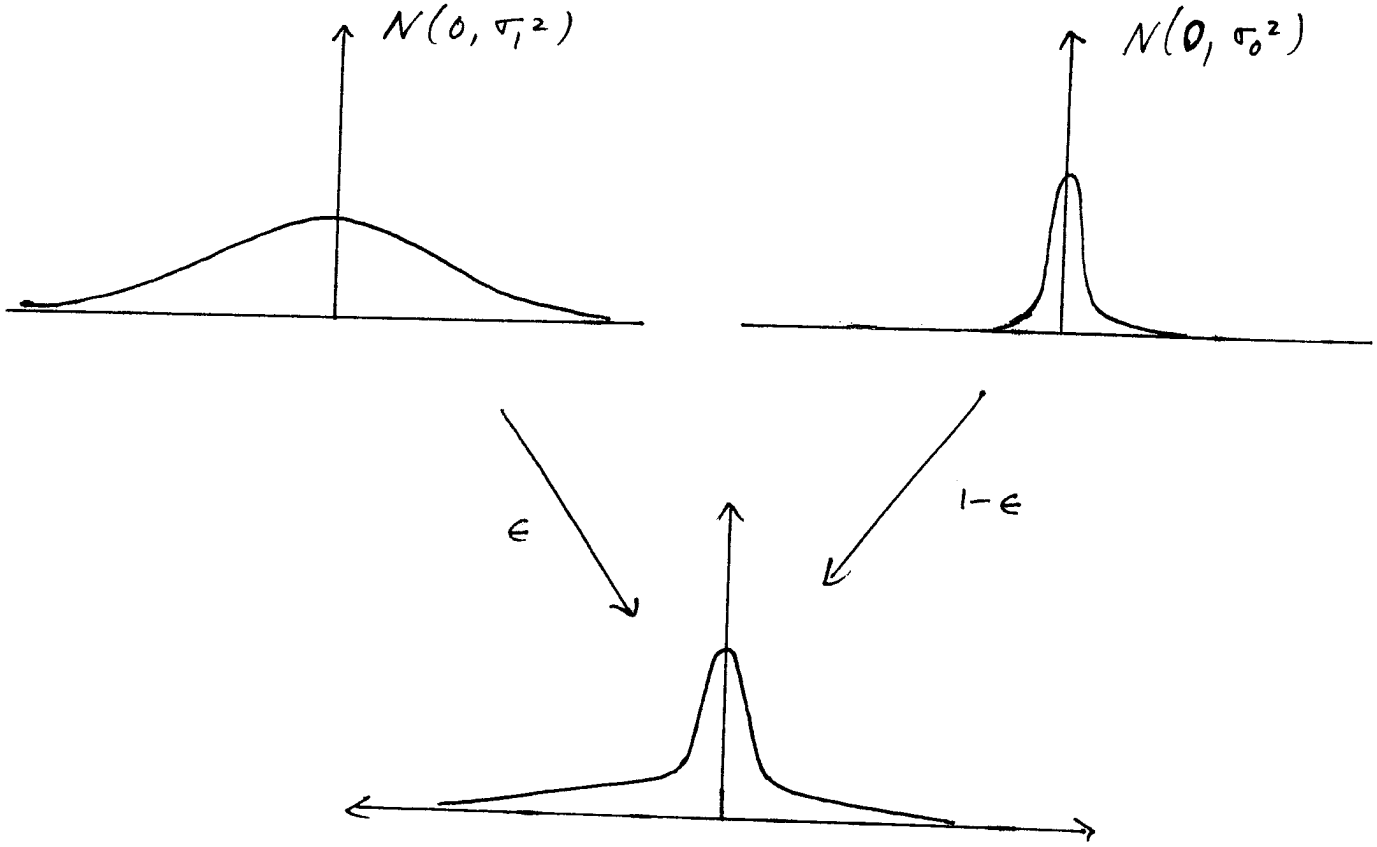
$$\theta_i \overset{iid}{\sim} \epsilon \mathcal{N}(0, \sigma_1^2) + (1-\epsilon)\mathcal{N}(0, \sigma_0^2).$$

where $\sigma_0^2 \ll \sigma_1^2$ and

ϵ = proportion of "significant" coefficients

σ_1^2 = variance of significant "

σ_0^2 = " " insignificant "



According to this prior, a detail coefficient θ is generated according to the following algorithm:

1. Flip an "ϵ-coin"
2. If heads,

$$\theta \sim N(0, \sigma_1^2)$$

Else

$$\theta \sim N(0, \sigma_0^2)$$

The Big Picture

- We observe $\underline{x} = \underline{s} + \underline{v}$, $\underline{v} \sim N(0, \sigma^2 I)$
- Compute $\underline{y} = \underline{\theta} + \underline{w}$ by taking wavelet transform
- View

$$y_i = \theta_i + w_i \quad , \quad w_i \sim N(0, \sigma^2)$$

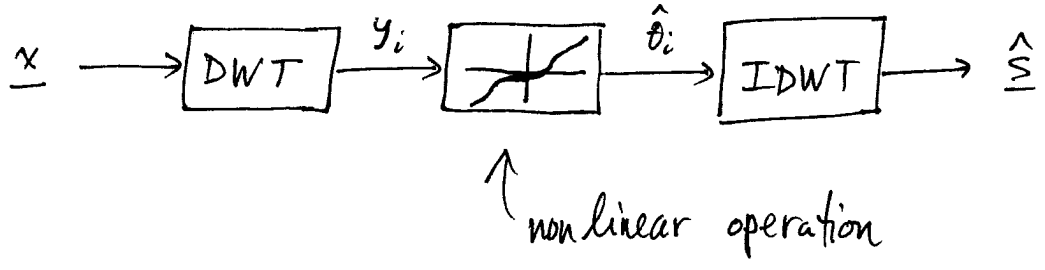
as independent estimation problems.

- Leave "coarse" coefficients unprocessed: noise will be "averaged out" since these are local averages
- Assume a mixture model prior on the detail coefficients and estimate

$$\hat{\theta}_i = E[\theta_i / y_i]$$

- Apply inverse wavelet transform to obtain

$$\underline{\hat{s}} = W \underline{\hat{\theta}}$$



Setting Parameters

Before this method is practical, we need ways to to set σ^2, ϵ, τ_1^2, and τ_0^2.

Donoho and Johnstone suggested the estimate

$$\hat{\sigma} = \frac{\underset{i>N/2}{MAD}(y_i)}{.6745},$$

which takes the "median absolute deviation" of the wavelet coefficients at the "finest" level of detail, and .6745 makes the estimate unbiased *if* all of the θ_i are in fact 0.

The mixture model parameters ϵ, τ_1^2, τ_0^2 may be estimated via maximum likelihood:

$$(\hat{\epsilon}, \hat{\tau}_1^2, \hat{\tau}_0^2) = \underset{(\epsilon, \tau_1^2, \tau_0^2)}{\arg\max} \; l(\epsilon, \tau_1^2, \tau_0^2 ; \underline{y})$$

Exercise | Determine a formula for the likelihood of $\epsilon, \sigma_1^2, \sigma_0^2$ given the detail coefficients $\underline{y}_{\text{detail}}$ $= [y_2, \ldots, y_N]^T$ (assuming a max-level wavelet transform)

Solution | Denote

$$\phi(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

Since

$$y_i = \theta_i + \omega_i$$

independent

where

$$f(\theta_i) = \epsilon\,\phi(\theta_i;0,\sigma_1^2) + (1-\epsilon)\,\phi(\theta_i;0,\sigma_0^2)$$

$$f(\omega_i) = \phi(\omega_i;0,\sigma^2)$$

it follows that

$$f(y_i) = f(\theta_i) \star f(\omega_i)$$

$$= \epsilon\,\phi(y_i;0,\sigma^2+\sigma_1^2) + (1-\epsilon)\,\phi(y_i;0,\sigma^2+\sigma_0^2)$$

Hence the likelihood of $\epsilon, \sigma_1^2, \sigma_0^2$ is

$$l(\epsilon,\sigma_1^2,\sigma_0^2;\underline{y}_{\text{detail}}) = \prod_{i=2}^{N} f(y_i;\epsilon,\sigma_1^2,\sigma_0^2)$$

$$= \prod_{i=2}^{N}\left[\epsilon\,\phi(y_i;0,\sigma^2+\sigma_1^2) + (1-\epsilon)\,\phi(y_i;0,\sigma^2+\sigma_0^2)\right]$$

Typically one uses an iterative algorithm such as an EM algorithm to fit mixture models. However, because there are only 3 unknowns, we could also maximize the likelihood by an exhaustive grid search.

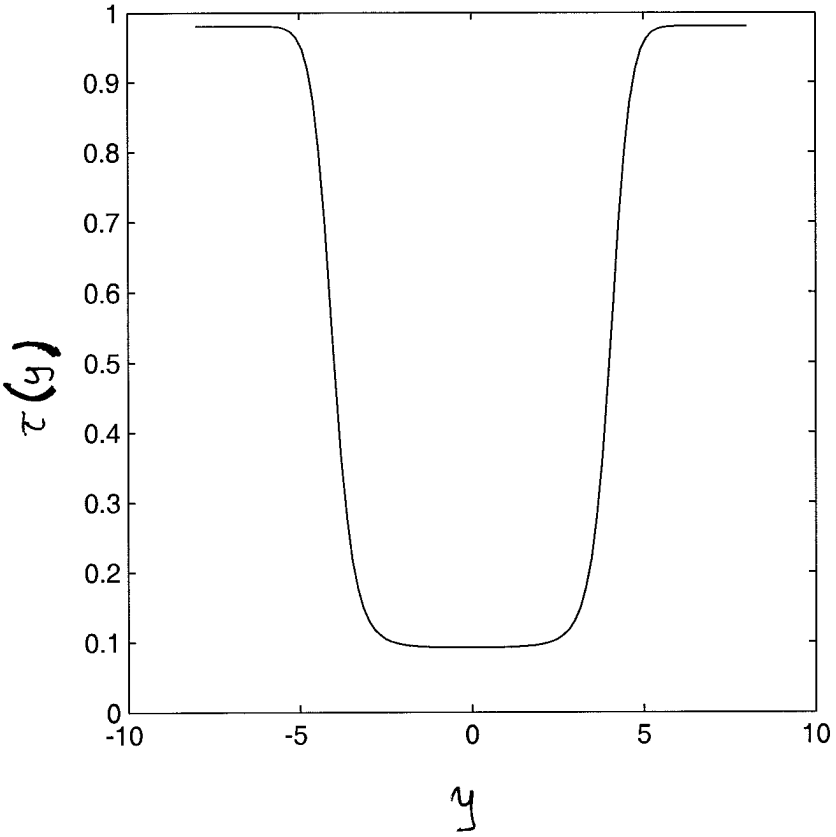
Note that we are using the data to determine our prior – hence we are not strictly adhering to the Bayesian philosophy. This kind of procedure is called an <u>empirical Bayesian</u> method.

You will show on the homework that

$$\hat{\theta} = E[\theta \mid y] = \tau(y) \cdot y$$

where $0 < \tau(y) < 1$ is called the <u>shrinkage</u> factor.

the shrinkage factor



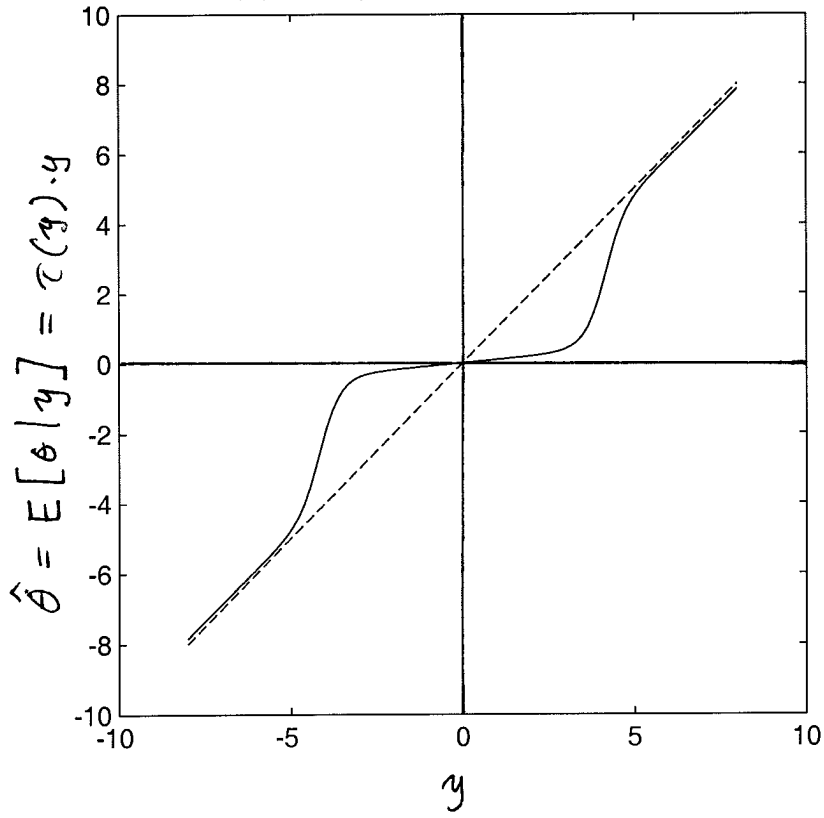
$\epsilon = 0.05$

$\sigma^2 = 1$

$\sigma_0^2 = .1$

$\sigma_1^2 = 50$

estimated wavelet coefficient



The effect of the shrinkage factor is that

- small coefficients are set nearly to zero
- large coefficients are virtually unaltered.

This property is consistent with our understanding

- small coefficients are mostly noise
- large coefficients contain actual signal

Extensions

- Image denoising
- More sophisticated priors

Summary

Gaussian noise
Mixture of Gaussians prior
} $\Rightarrow$ Wavelet shrinkage

LINEAR ESTIMATION

Linear estimators are an important class of estimators

- simplicity
- dependence on first and second order moments only
- ease of implementation

This last point is especially important for filtering problems where we must process data real-time.

Our discussion of filtering will rely on Bayesian linear estimators, but for completeness we begin with classical linear estimators.

We make the following distinctions:

constant: $\hat{\underline{\theta}}(\underline{x}) = \underline{b}$

linear: $\hat{\underline{\theta}}(\underline{x}) = H\underline{x}$

affine: $\hat{\underline{\theta}}(\underline{x}) = H\underline{x} + \underline{b}$

where $H \in \mathbb{R}^{p \times N}$, $\underline{b} \in \mathbb{R}^p$. We'll study all three cases.

Best Linear Unbiased Estimation

The MVUE is often not computable.

- CRLB or Rao-Blackwell not applicable
- intractable mathematical model
- randomness is only known up to first and second order moments.

In such cases, we must be content with a suboptimal estimator. One approach is to compute

Definition The best linear unbiased estimator (BLUE) is the linear estimator

$$\hat{\underline{\theta}}(\underline{x}) = H\underline{x} \qquad , \quad H \in \mathbb{R}^{p \times N}$$

with smallest variance among all linear, unbiased estimators.

Note that for $\hat{\underline{\theta}}$ to be unbiased we must have

$$\underline{\theta} = E\{\hat{\underline{\theta}}\} = E\{H\underline{X}\} = HE\{\underline{X}\}$$

Therefore the mean of the data must obey a linear relationship with the true parameter.

This relationship will not always be true, so even the suboptimal BLUE isn't always feasible.

However, there is an important class of problems where it does hold.

Linear Model

Suppose $\underline{X}$ and $\underline{\theta}$ are related through

$$\underline{X} = A\underline{\theta} + \underline{W}$$

$\underline{\theta}$ fixed but unknown

where

A is $N \times p$ and known, full rank

$\underline{W}$ is random

$E[\underline{W}] = \underline{0}$

$E[\underline{W}\underline{W}^T] =: R$ is known, positive definite

Note: $\underline{W}$ is not necessarily Gaussian

Question: Can you give an H such that

$H E[\underline{X}] = \underline{\theta}$?

Taking $H = (A^T A)^{-1} A^T$ we find

$$H E[\underline{X}] = \underbrace{(A^T A)^{-1} A^T A}_{I}\,\underline{\theta} + (A^T A)^{-1} A^T \underbrace{E[\underline{W}]}_{\underline{0}}$$

$$= \underline{\theta}.$$

So the pseudo inverse is <u>an</u> unbiased estimator.

But is its variance minimal?

<u>Theorem</u> (Gauss-Markov Theorem)

In the linear model described above, the BLUE is

$$\hat{\underline{\theta}}(\underline{x}) = (A^T R^{-1} A)^{-1} A^T R^{-1} \underline{x}$$

and its covariance matrix is

$$R_{\hat{\theta}} = E[(\hat{\underline{\theta}} - \underline{\theta})(\hat{\underline{\theta}} - \underline{\theta})^T] = (A^T R^{-1} A)^{-1}$$

Proof 1 $\hat{\underline{\theta}}$ is unbiased $\iff E\{H\underline{X}\} = \underline{\theta} \quad \forall \underline{\theta}$.

Now $E\{H\underline{X}\} = HE\{X\} = HE\{A\theta + \underline{W}\} = HA\underline{\theta}$.

Thus $\hat{\underline{\theta}}$ is unbiased $\iff HA\underline{\theta} = \underline{\theta} \quad \forall \underline{\theta} \iff HA = I_{p\times p}$.

Now assume $\hat{\underline{\theta}}$ is unbiased and let's compute

$$\begin{aligned}
\text{Var}(\hat{\underline{\theta}}) &= E\{(\hat{\underline{\theta}} - \underline{\theta})^T(\hat{\underline{\theta}} - \underline{\theta})\} \\
&= E\{(H\underline{X} - \underline{\theta})^T(H\underline{X} - \underline{\theta})\} \\
&= E\{(HA\underline{\theta} + H\underline{W} - \underline{\theta})^T(HA\underline{\theta} + H\underline{W} - \underline{\theta})\} \\
&= E\{(H\underline{W})^T(H\underline{W})\}.
\end{aligned}$$

Denote

$$H = [\underline{h}_1 \cdots \underline{h}_p]^T = \begin{bmatrix} \underline{h}_1^T \\ \vdots \\ \underline{h}_p^T \end{bmatrix} \qquad (p \times N)$$

Then

$$\begin{aligned}
\text{Var}(\hat{\underline{\theta}}) &= E\left\{\sum_{i=1}^{p} (\underline{h}_i^T \underline{W})^2\right\} \\
&= \sum_{i=1}^{p} E\{(\underline{h}_i^T \underline{W})^2\} \\
&= \sum_{i=1}^{p} E\{(\underline{h}_i^T \underline{W})(\underline{W}^T \underline{h}_i)\} \\
&= \sum_{i=1}^{p} \underline{h}_i^T R \underline{h}_i
\end{aligned}$$

Thus, we need to solve the constrained optimization problem

$$\min_{H} \sum_{i=1}^{p} \underline{h}_i^T R \underline{h}_i$$

$$s.t \quad H \cdot A = I$$

By the theory of Lagrange multipliers, it suffices to solve the unconstrained problem

$$\min_{H, \underline{\lambda}} L(H, \underline{\lambda})$$

where L is the Lagrangian and $\underline{\lambda}$ is a vector of real numbers called Lagrange multipliers, one for each equality constraint.

The Lagrangian is

$$L = \sum_{i=1}^{P} \underline{h}_i^T R \underline{h}_i + \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_j^{(i)} (\underline{h}_i^T \underline{q}_j - \delta_{ij})$$

Taking derivatives

$$\frac{\partial L}{\partial \underline{h}_i} = 2R\underline{h}_i + \sum_{j=1}^{P} \lambda_j^{(i)} \underline{q}_j$$

$$= 2R\underline{h}_i + A\underline{\lambda}^{(i)}$$

where $\underline{\lambda}^{(i)} = [\lambda_1^{(i)} \cdots \lambda_P^{(i)}]^T$.

$$\Rightarrow \hat{\underline{h}}_i = -\frac{1}{2} R^{-1} A \underline{\lambda}^{(i)}$$

From the constraint we know

$$A^T \hat{\underline{h}}_i = \underline{e}_i = [0 \cdots \underset{\text{ith position}}{1} \cdots 0]$$

$$\Rightarrow \underline{e}_i = -\frac{1}{2} A^T R^{-1} A \underline{\lambda}^{(i)}$$

$$\Rightarrow \underline{\lambda}^{(i)} = -2 (A^T R^{-1} A)^{-1} \underline{e}_i$$

$$\Rightarrow \hat{\underline{h}}_i = R^{-1} A (A^T R^{-1} A) \underline{e}_i$$

Assembling these scalar results, we have

$$\hat{H} = \left[\hat{\underline{h}}_1 \cdots \hat{\underline{h}}_p \right]^T$$

$$= \left(R^{-1} A \left(A^T R^{-1} A \right)^{-1} \cdot I_{p \times p} \right)^T$$

$$= \left(A^T R^{-1} A \right)^{-1} A^T R$$

The covariance matrix of $\hat{\underline{\theta}}$ is

$$\text{Cov}(\hat{\underline{\theta}}) =$$

Proof 2 From proof 1, we know that the BLUE depends only on the first and second order moments of $\underline{x}$ (or equivalently, of $\underline{w}$).

Therefore, we may assume

$$\underline{w} \sim \mathcal{N}(\underline{0}, R).$$

We have previously seen that

$$\hat{\theta}(\underline{x}) = (A^T R^{-1} A)^{-1} A^T R^{-1} \underline{x}$$

is MVUE. Since this estimator is already linear, it is also the BLUE ∎

Remark As the above discussion notes, when $\underline{w}$ is Gaussian, the BLUE is the MVUE for the linear model, i.e. the BLUE is optimal.

Linear Minimum Mean Squared Error Estimation

Let us now turn to a Bayesian setting. There are some important similarities/differences w.r.t. classical linear estimation:

- the optimal linear estimator depends only on first and second order moments, but now for $\underline{X}$ and $\underline{\theta}$
- the optimal linear estimator always exists in the Bayesian setting
- Bayesian linear estimation has important geometric interpretations in terms of orthogonal projections in Hilbert space

Definition $\hat{\underline{\theta}}(\underline{x}) = \hat{H}\underline{x}$ is the LMMSE estimator if $\hat{H}$ minimizes

$$\text{BMSE}(H) := E_{X,\theta}\left[(\underline{\theta} - H\underline{X})^T(\underline{\theta} - H\underline{X})\right]$$

Introduce the notation

$$R_{\theta\theta} = E\left\{(\underline{\theta} - E\underline{\theta})(\underline{\theta} - E\underline{\theta})^T\right\}$$

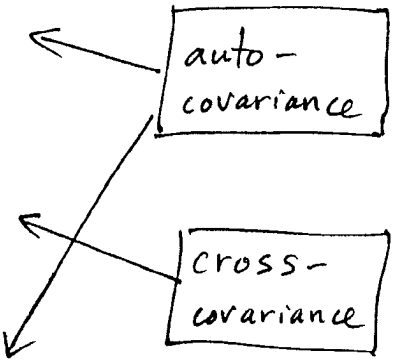
$(p \times p)$

$$R_{\theta x} = E\left\{(\underline{\theta} - E\underline{\theta})(\underline{x} - E\underline{x})^T\right\}$$

$(p \times N)$

$$R_{xx} = E\left\{(\underline{x} - E\underline{x})(\underline{x} - E\underline{x})^T\right\}$$

$(N \times N)$



Theorem If $E\underline{\theta} = \underline{0}$ and $E\underline{x} = \underline{0}$

then the LMMSE is

$$\hat{\underline{\theta}}(\underline{x}) = R_{\theta x} R_{xx}^{-1} \underline{x}$$

provided R_{xx} is positive definite.

Proof |

$$BMSE(H) = E\left[(H\underline{X} - \underline{\theta})^T (H\underline{X} - \underline{\theta})\right]$$

$$= E\left[\sum_{i=1}^{p} (\underline{h}_i^T \underline{X} - \theta_i)^2\right]$$

$$= \sum_{i=1}^{p} E\left[(\underline{h}_i^T \underline{X} - \theta_i)^2\right]$$

To minimize this term w.r.t $H = [\underline{h}_1, \dots \underline{h}_p]^T$, it suffices to optimize each $\underline{h}_i$ independently. That is, the vector linear estimation problem reduces to the scalar linear estimation problem.

Thus, let's drop the subscript i and focus on minimizing $E\left[(\underline{h}^T \underline{X} - \theta)\right]$ w.r.t. $\underline{h}$.

Now

$R_{\theta\theta} = Var(\theta)$ is 1×1

$R_{\theta x}$ is $1 \times N$

Now

$$\mathrm{BMSE}(\underline{h}) = E_{\underline{x},\theta}\left[(\theta - \underline{h}^T\underline{x})^2\right]$$

$$= E[\theta^2] - 2E[\underline{h}^T\underline{x}\cdot\theta] + E[(\underline{h}^T\underline{x})^2]$$

$$= E[\theta^2] - 2\underline{h}^T E[\underline{x}\cdot\theta] + E[\underline{h}^T\underline{x}\cdot\underline{x}^T\cdot\underline{h}]$$

$$= R_{\theta\theta} - 2\underline{h}^T R_{x\theta} + \underline{h}^T R_{xx}\underline{h} \qquad (*)$$

By completing the square, we have

$$\mathrm{BMSE}(\underline{h}) = \left(\underline{h} - R_{xx}^{-1}R_{x\theta}\right)^T R_{xx}\left(\underline{h} - R_{xx}^{-1}R_{x\theta}\right)$$

$$\underbrace{- R_{x\theta}^T R_{xx}^{-1} R_{x\theta} + R_{\theta\theta}}_{\text{independent of } \underline{h}}$$

Since R_{xx} is positive definite, the unique minimizer is

$$\hat{\underline{h}} = R_{xx}^{-1}R_{x\theta}$$

$$\Updownarrow$$

$$\hat{\underline{h}}^T = R_{\theta x}R_{xx}^{-1}.$$

Alternatively, you could apply vector calculus to minimize $(*)$.

Now the vector LMMSE estimator is obtained by stacking these scalar estimators. ▨

Nonzero Means and Affine Estimators

When $E\underline{\theta} \neq \underline{0}$ or $E\underline{x} \neq \underline{0}$, the LMMSE is generalized by the affine MMSE estimator (AMMSE), which is defined to be the estimator $\hat{\underline{\theta}}(\underline{x}) = \hat{H}\underline{x} + \hat{\underline{b}}$

minimizing

$$BMSE(H, \underline{b}) = E\left[(\underline{\theta} - H\underline{x} - \underline{b})^T (\underline{\theta} - H\underline{x} - \underline{b})\right].$$

Theorem The AMMSE is

$$\hat{\underline{\theta}}(\underline{x}) = E\underline{\theta} + R_{\theta x} \cdot R_{xx}^{-1} (\underline{x} - E\underline{x})$$

i.e.

$$\hat{H} = R_{\theta x} R_{xx}^{-1}$$

$$\hat{\underline{b}} = E\underline{\theta} - \hat{H} E\underline{x}$$

Proof Introduce the variables

$$\underline{\theta}' = \underline{\theta} - E\underline{\theta}$$

$$\underline{x}' = \underline{x} - E\underline{x}$$

Then $BMSE(H, \underline{b})$

$$= E\left[(\underline{\theta}' - H\underline{x}' - (\underline{b} - E\underline{\theta} + HE\underline{x}))^T \cdot (\underline{\theta}' - H\underline{x}' - (\underline{b} - E\underline{\theta} + HE\underline{x}))\right]$$

$$= E\left[(\underline{\theta}' - H\underline{X}')^T(\underline{\theta}' - H\underline{X}')\right]$$

$$-2E\left[(\underline{\theta}' - H\underline{X}')^T(\underline{b} - E\underline{\theta} + HE\underline{X})\right]$$

$$+ E\left[(\underline{b} - E\underline{\theta} + HE\underline{X})^T(\underline{b} - E\underline{\theta} + HE\underline{X})\right]$$

The second term is zero since $\underline{b} - E\underline{\theta} + HE\underline{X}$ is constant and $\underline{\theta}' - H\underline{X}'$ is zero mean.

For fixed H, the optimal $\underline{b}$ must minimize the third term, which can be made zero by taking

$$\underline{b} = E\underline{\theta} - HE\underline{X}. \qquad (\star)$$

Therefore, H is obtained by minimizing the first term. Since $\underline{\theta}'$, $\underline{X}'$ are zero mean, $\hat{H}$ is the LMMSE estimator for $\underline{\theta}'$, $\underline{X}'$, i.e.

$$\hat{H} = R_{\theta'X'} R_{X'X'}^{-1} = R_{\theta X} R_{XX}^{-1} \qquad \square$$

Remark] From the above argument and $(\star)$, we see that the optimal constant estimator ($H = 0$) is

$$\hat{\underline{b}} = E\underline{\theta},$$

the prior mean.

Connection to the Jointly Gaussian Case

The LMMSE and AMMSE estimators look exactly like the MMSE estimators for the case where $\underline{\theta}, \underline{x}$ are jointly Gaussian. In fact, we can use the Gaussian case to give an alternate derivation of the LMMSE/AMMSE.

Let's consider the AMMSE. From the definition of the Bayesian MSE

$$BMSE(H,b) = E\left[(\underline{\theta} - H\underline{x} - \underline{b})^T(\underline{\theta} - H\underline{x} - \underline{b})\right]$$

is can be seen that this criterion depends only on the first and second order moments of $\underline{\theta}$ and $\underline{x}$. Therefore, we can assume the higher order moments are whatever we want. So let's assume $\underline{\theta}$ & $\underline{x}$ are jointly Gaussian with the given means and covariances:

$$\begin{bmatrix} \underline{x} \\ \underline{\theta} \end{bmatrix} \sim N\left(\begin{bmatrix} E\underline{x} \\ E\underline{\theta} \end{bmatrix}, \begin{bmatrix} R_{xx} & R_{x\theta} \\ R_{\theta x} & R_{\theta\theta} \end{bmatrix} \right)$$

We know the MMSE estimator is the posterior mean, which is

$$E\underline{\theta} + R_{\theta x} R_{xx}^{-1} (\underline{x} - E\underline{x})$$

by the Gaussian conditioning principle.

Since this estimator is affine, it is the AMMSE ∎

Theorem | Bayesian Gauss-Markov Theorem

Assume

$$\underline{X} = A\underline{\theta} + \underline{W}$$

where

A is full rank, $N \times p$, known

$E\underline{\theta}$, $R_{\theta\theta}$ known

$E\underline{W} = \underline{0}$, R_{ww} known

$R_{\theta w} = 0_{p \times N}$ ($\underline{\theta}$ & $\underline{W}$ uncorrelated)

Then the affine MMSE estimator is

$$\hat{\underline{\theta}}(\underline{x}) = E\underline{\theta} + R_{\theta\theta} A^T (A R_{\theta\theta} A^T + R_{ww})^{-1} (\underline{x} - A E\underline{\theta})$$

$$= E\underline{\theta} + (R_{\theta\theta}^{-1} + A^T R_{ww}^{-1} A)^{-1} A^T R_{ww}^{-1} (\underline{x} - A E\underline{\theta})$$

How would you prove this?

Proofs

1. Show $R_{\theta x} =$

(a)

$$R_{xx} =$$

and apply the AMMSE result.

2. Argue that the BMSE depends only on the first and second order moments of $\underline{\theta}$ and $\underline{w}$. Then assume $\underline{\theta}, \underline{w}$ are jointly Gaussian. The MMSE estimator has the stated form, and since it is affine it is the AMMSE.

Exercise Suppose

$$X_i = A + W_i, \qquad i = 1, \ldots, N$$

where

$$A \sim \text{unif}(-A_0, A_0)$$
$$W_i \overset{iid}{\sim} N(0, \sigma_W^2)$$

} independent

1. Find the posterior $f(A/\underline{x})$
2. Can you compute $E[A/\underline{x}]$?
3. Determine the LMMSE estimator of A.

Solution

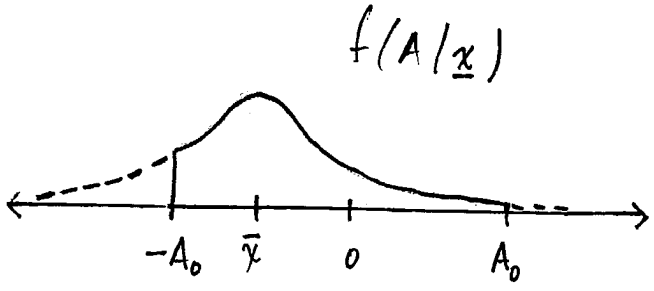
1. $f(A|\underline{x}) \propto f(\underline{x}|A) \cdot f(A)$

$$= \prod_{i=1}^{N} \phi(x_i | A, \sigma^2) \cdot \frac{1}{2A_0} I_{[-A_0, A_0]}(A)$$

$$\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - A)^2 \right\}$$

$$= \exp\left\{ -\frac{1}{2\sigma^2} \left[\Sigma (x_i - \bar{x})^2 + N(\bar{x} - A)^2 \right] \right\}$$

$$\Rightarrow f(A|x) \propto \exp\left\{ \frac{(A - \bar{x})^2}{2(\sigma^2/N)} \right\} \cdot I_{[-A_0, A_0]}(A)$$



Truncated normal.

2. $E[A|\underline{x}] = \int_{-\infty}^{\infty} A f(A|\underline{x})\, dA$

$$= \frac{\int_{-A_0}^{A_0} A \phi(A | \bar{x}, \frac{\sigma^2}{N})\, dA}{\int_{-A_0}^{A_0} \phi(A | \bar{x}, \frac{\sigma^2}{N})\, dA}$$

← Normalization Constant

Numerator : closed form solution

Denominator : no closed form solution

Note: $E[A|\underline{x}]$ is nonlinear.

3. $\underline{X} = \underline{1} \cdot A + \underline{W}, \quad \underline{W} \sim N(\underline{0}, \sigma_w^2 I)$

$$\sigma_A^2 = \int_{-A_0}^{A_0} \frac{x^2}{2A_0}\, dx = \frac{1}{6A_0} x^3 \Big|_{-A_0}^{A_0} = \frac{A_0^2}{3}$$

$$\Rightarrow \hat{A} = \left(\left(\sigma_A^2\right)^{-1} + \left(\underline{1}^T \sigma_w^2 I \underline{1} \right)^{-1} \right)^{-1} (\sigma_w^2 I)^{-1} \underline{1}^T \underline{x}$$

$$= \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_w^2}{N}} \bar{x}$$

$$= \frac{A_0^2/3}{A_0^2/3 + \sigma^2/N} \bar{x}$$

Remarks

a) To compute the LMMSE estimator of A, we didn't need to know it was uniform, just its mean and variance.

b) Estimator is the same as the MMSE estimator for the prior

$$A \sim N\left(0, \frac{A_0^2}{3}\right)$$

The Orthogonality Principle

The LMMSE estimator satisfies

$$R_{\theta X} = \hat{\underline{h}}^T R_{XX}$$

or equivalently

$$\underline{0} = R_{\theta X} - \hat{\underline{h}}^T R_{XX}$$

$$= E\left[\theta \underline{X}^T - \hat{\underline{h}}^T \underline{X}\,\underline{X}^T\right]$$

$$= E\left[(\theta - \hat{\underline{h}}^T \underline{X})\,\underline{X}^T\right]$$

$$= E\left[(\theta - \hat{\theta})\,\underline{X}^T\right]$$

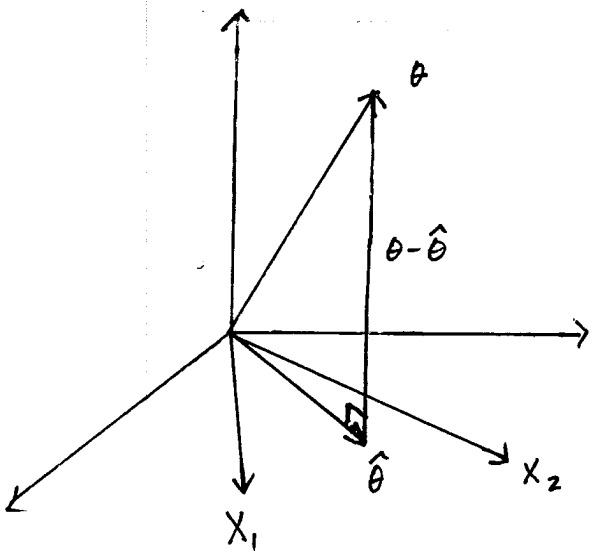
In words, this says that the prediction error is orthogonal to every measurement variable,

$$\theta - \hat{\theta} \perp X_i \qquad \forall i$$

where orthogonality is with respect to the inner product given by expectation.

As a consequence,

$$\theta - \hat{\theta} \perp \underline{h}^T \underline{X} \qquad \forall \underline{h}.$$



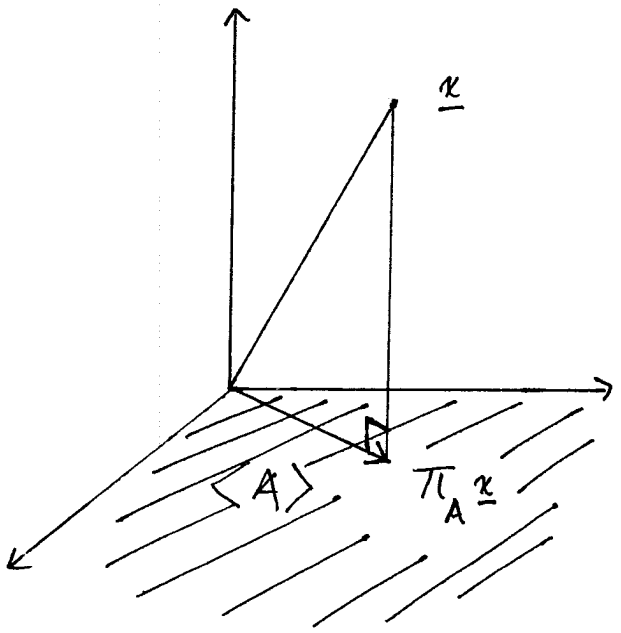
Therefore, the prediction error is orthogonal to the entire subspace of linear estimators.

This is one manifestation of the orthogonality principle.

The orthogonality principle is also manifested in nonstatistical projections

In particular, if $\underline{x} \in \mathbb{R}^N$ and $\langle A \rangle \subseteq \mathbb{R}^N$, then

$$\underline{x} - \Pi_A \underline{x} \perp \underline{u} \qquad \forall \underline{u} \in \langle A \rangle$$



Here the inner product is the standard dot product.

To prove this fact, suppose $\underline{u} \in \langle A \rangle$.
Then $\underline{u} = A\underline{\phi}$ for some $\underline{\phi}$,

$$\langle \underline{x} - \Pi_A \underline{x}, \underline{u} \rangle = \underline{u}^T (\underline{x} - \Pi_A \underline{x})$$

$$= \underline{\phi}^T A^T \left(\underline{x} - A(A^TA)^{-1}A^T \underline{x} \right)$$

$$= \underline{\phi}^T A^T \underline{x} - \underline{\phi}^T A^T \underline{x} = 0$$

In full generality...

Definition Let $\mathcal{H}$ be an inner product space, and let $S \subseteq \mathcal{H}$ be a linear subspace. The *orthogonal complement* of S is

$$S^{\perp} = \{ \underline{v} \in \mathcal{H} \mid \underline{v} \perp \underline{u} \quad \forall \underline{u} \in S \}$$

Theorem Let $\mathcal{H}$ be a Hilbert space and $S \subseteq \mathcal{H}$ a closed linear subspace.

1. *Projection Theorem*: $\mathcal{H} = S \oplus S^{\perp}$, i.e. $\forall \underline{w} \in \mathcal{H}$, $\exists$ unique $\underline{u} \in S$, $\underline{v} \in S^{\perp}$ s.t. $\underline{w} = \underline{u} + \underline{v}$. Thus we can define the (orthogonal) projection onto S,

   $\Pi_S(\underline{w}) = \underline{u}$

2. *Closest point property*: $\Pi_S \underline{w}$ is the unique solution of $\min_{\underline{u} \in S} \| \underline{w} - \underline{u} \|$ where $\| \cdot \|$ is the norm induced by $\langle \cdot, \cdot \rangle$.

3. *Orthogonality principle*: For all $\underline{w} \in \mathcal{H}$,

   $$\underline{w} - \Pi_S \underline{w} \perp \underline{u} \qquad \forall \underline{u} \in S$$

   i.e., $\underline{w} - \Pi_S \underline{w} \in S^{\perp}$.

Note 2 & 3 follow easily from 1. See Moon and Stirling for details.

Application to LMMSE Estimation

Consider the space $\mathcal{H}$ of all scalar random variables with zero mean and finite variance. It can be shown that $\mathcal{H}$ is a Hilbert space with inner product

$$\langle V_1, V_2 \rangle = E\{V_1 \cdot V_2\}$$

Let $X_1, \ldots, X_N$ be random measurements and define the subspace

$$S = \{ \underline{h}^T \underline{X} \mid \underline{h} \in \mathbb{R}^N \}$$

where $\underline{X} = [X_1 \ \cdots \ X_N]^T$.

If $\theta \in \mathcal{H}$ is a scalar parameter of interest, the LMMSE estimator is the projection

$$\begin{aligned} \hat{\theta} &= \Pi_S \, \theta \\ &= \underset{\hat{\theta} \in S}{\operatorname{argmin}} \ \| \theta - \hat{\theta} \|^2 \\ &= \underset{\underline{h} \in \mathbb{R}^N}{\operatorname{argmin}} \ E\left[(\theta - \underline{h}^T \underline{X})^2 \right] \end{aligned}$$

By the orthogonality principle,

$$\theta - \hat{\theta} \perp u \qquad \forall u \in S.$$

Writing $\hat{\theta} = \underline{\hat{h}}^T \underline{X}$ and taking $u = X_i$, $i = 1, \ldots, N$ we have

$$\theta - \hat{\theta} \perp X_i \iff \theta - \underline{\hat{h}}^T \underline{X} \perp X_i$$

$$\iff E\left[(\theta - \underline{\hat{h}}^T \underline{X}) X_i\right] = 0$$

Applying this for $i = 1, \ldots, N$ we have

$$E\left[(\theta - \underline{\hat{h}}^T \underline{X}) \underline{X}^T\right] = [0 \cdots 0]$$

$$\Updownarrow$$

$$R_{\theta X} = \underline{\hat{h}}^T R_{XX}$$

$$\Updownarrow$$

$$\underline{\hat{h}}^T = R_{\theta X} R_{XX}^{-1}$$

Conclusion: The orthogonality principle gives us another proof of the form of the LMMSE estimator.

We will apply the orthogonality principle often in our study of filtering. Furthermore, the orthogonality principle applies when there are an infinite number of equations.

Terminology

The equations

$$R_{xx}\,\underline{\hat{h}} = R_{x\theta}$$

are called the Wiener-Hopf or normal equations.

The optimal $\underline{\hat{h}}$ is also called a Wiener estimator, after Norbert Wiener, especially in the context of estimating a signal in additive noise.

Summary

- Classical and Bayesian linear estimation
  - only depends on first and second order moments
  - optimal in Gaussian case
  - proof: show that solution determined by 1st & 2nd order moments, assume Gaussianity, and invoke Gaussian results
- classical BLUE
  - solution of constrained quadratic minimization, doesn't always exist
- Bayesian LMMSE
  - solution of unconstrained quadratic minimization, always exists
  - projection in Hilbert space, obeys orthogonality principle

Key a. $R_{\theta x} = E[(\underline{\theta} - E\underline{\theta})(\underline{x} - E\underline{x})^T]$

$= E[(\underline{\theta} - E\underline{\theta})(A\underline{\theta} + \underline{w} - AE\underline{\theta})^T]$

$= E[(\underline{\theta} - E\underline{\theta})\underline{w}^T] + E[(\underline{\theta} - E\underline{\theta})(\underline{\theta} - E\underline{\theta})^T A^T]$

$= 0 + R_{\theta\theta} A^T = R_{\theta\theta} A^T$

$$R_{xx} = E\left[(\underline{X} - E\underline{X})(\underline{X} - E\underline{X})^T\right]$$

$$= E\left[(A\underline{\theta} + \underline{w} + AE\underline{\theta})(A\underline{\theta} + \underline{w} + AE\underline{\theta})^T\right]$$

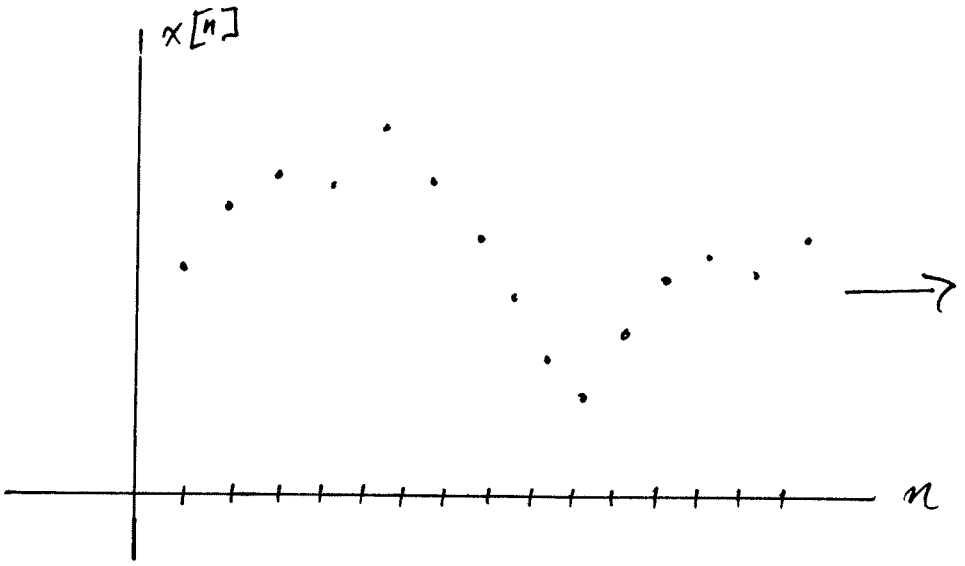
$$= A R_{\theta\theta} A^T + R_{ww}$$

where we use the assumption that $\underline{\theta}, \underline{w}$ are uncorrelated, i.e. $E[(\underline{\theta} - E\underline{\theta})\underline{w}^T] = 0$.

FILTERING

Thus far in our study of estimation the data have been static: All the data are available at once, and the number of measurements N is not so large that R_{xx} is difficult to invert.

Now we turn our attention to the situation where data are dynamic, that is, we observe a "stream" $\{x[n]\}$ of measurements one value at a time.



$$x[n] = s[n] + w[n]$$

The following problems are of interest:

1. Filtering: Given $x[n], x[n-1], \cdots,$

   estimate $s[n]$.

2. Smoothing: Given $x[n], x[n-1], \ldots$

   estimate $s[m]$, $m < n$.

3. Signal prediction: Given $x[n], x[n-1], \ldots$

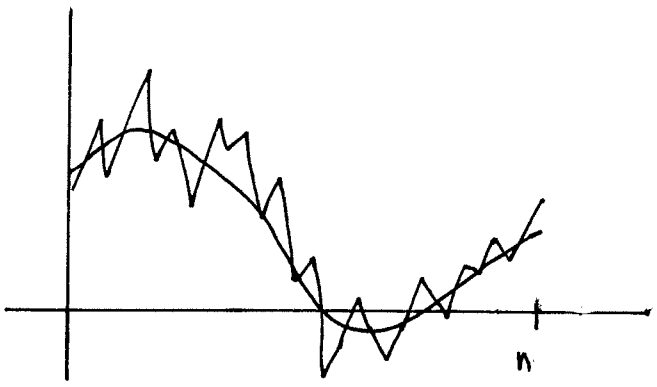
   estimate $s[m]$, $m > n$.

4. Measurement prediction: Given $x[n], x[n-1], \ldots$

   estimate $x[m]$, $m > n$

5. Interpolation: Given $\ldots x[n-2], x[n-1], x[n+1], x[n+2] \ldots$

   estimate a missing value $x[n]$.



To address these problems we need computationally efficient estimators that can be efficiently updated as new data arrive.

Our approach includes

- linear estimators
- signal models that lead to "structured" covariance matrices whose inverses are easier to compute and update than typical matrices.

Example | If $\{x[n]\}$ is wide-sense stationary (WSS) with auto-correlation function (assuming zero mean)

$$r_{xx}[k] = E\{x[n]\,x[n+k]\},$$

and we observe $\underline{x} = [x[n-1] \cdots x[n-p]]^T$ then the data autocovariance matrix is

$$R_{xx} = \begin{bmatrix} r_{xx}[0] & r_{xx}[1] & & \\ r_{xx}[1] & r_{xx}[0] & r_{xx}[1] & \\ & r_{xx}[1] & \ddots & \ddots \\ & & \ddots & \end{bmatrix}$$

is a Toeplitz matrix.

The inverse of a Toeplitz matrix can be computed in $O(p^2)$ operations as opposed to the usual $O(p^3)$. More on this later.

Note] We will adhere to the convention of always using lowercase letters to denote random processes, regardless of whether we mean a random variable or a realization.

WSS Random Processes

Since one of our major assumptions (for linear prediction and Wiener filtering) will be wide-sense stationarity, lets look at some examples:

1: White noise: A white noise process $w[n]$ satisfies three properties:

(a) values at different times are uncorrelated

(b) $E\{w[n]\} = 0 \quad \forall n \in \mathbb{Z}$

(c) $\text{Var}\{w[n]\} = \sigma_w^2 \quad \forall n \in \mathbb{Z}$

An important special case is Gaussian white noise,

$$w[n] \overset{iid}{\sim} N(0, \sigma_w^2).$$

Why do you think "white" is used to describe such a process?

Notation: $wn(\sigma_w^2)$

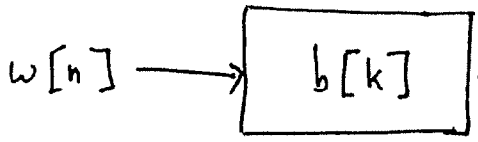
2. <u>Moving average process</u>: a linear combination of terms in a white noise process,

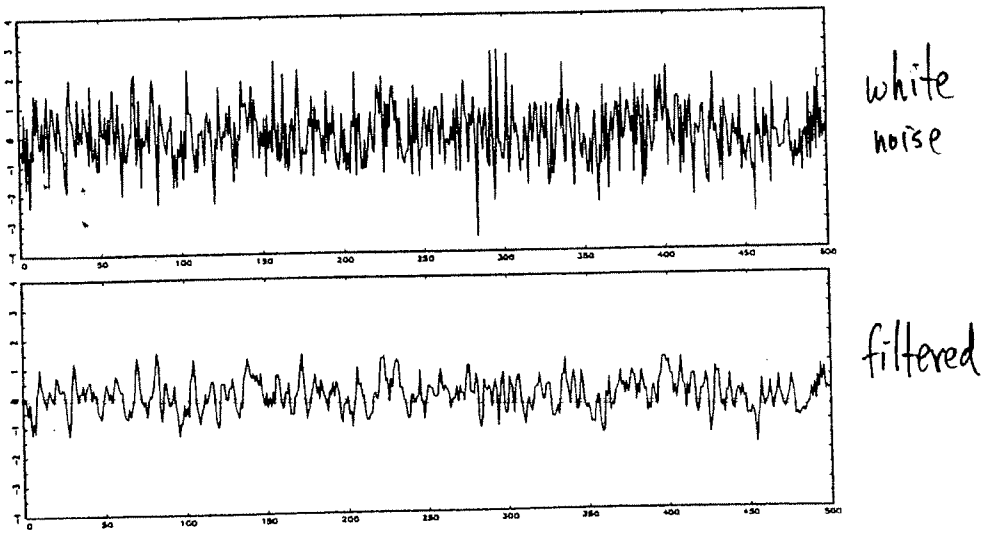
$$x[n] = \sum_{k=0}^{q} b[k]\, w[n-k]$$

such as

$$x[n] = \frac{1}{3} w[n] + \frac{1}{3} w[n-1] + \frac{1}{3} w[n-2].$$

In general, an MA process is obtained by passing white noise through a finite impulse response (FIR) filter





Notation: MA(q)

3. Autoregressive process

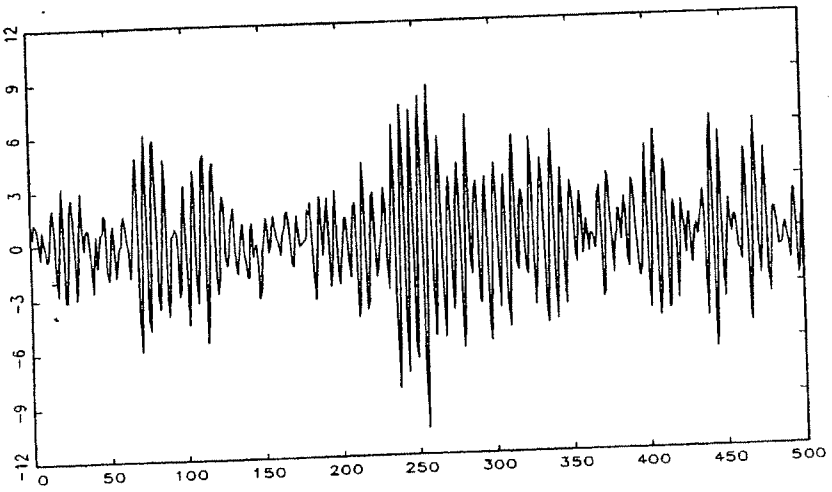
$$x[n] = -\sum_{k=1}^{p} a[k]\, x[n-k] + w[n]$$

where $w[n]$ is white noise. An AR process is obtained by passing white noise through an IIR filter

$$w[n] \longrightarrow \boxed{\text{IIR}} \longrightarrow x[n].$$

Simple example:

$$x[n] = x[n-1] - .9\, x[n-2] + w[n].$$



Notation: AR(p)

4. **ARMA process** An auto-regressive moving average process of orders p and q obeys

$$x[n] = -\sum_{k=1}^{p} a[k]\, x[n-k] + \sum_{k=0}^{q} b[k]\, w[n-k]$$

where $w[n]$ is white noise.

Notation: ARMA(p, q)

Fact: Essentially any discrete-time WSS RP can be approximated arbitrarily well by a

- MA(q) model, $q \to \infty$
- AR(p) model, $p \to \infty$
- ARMA(p,q) model, p and/or $q \to \infty$

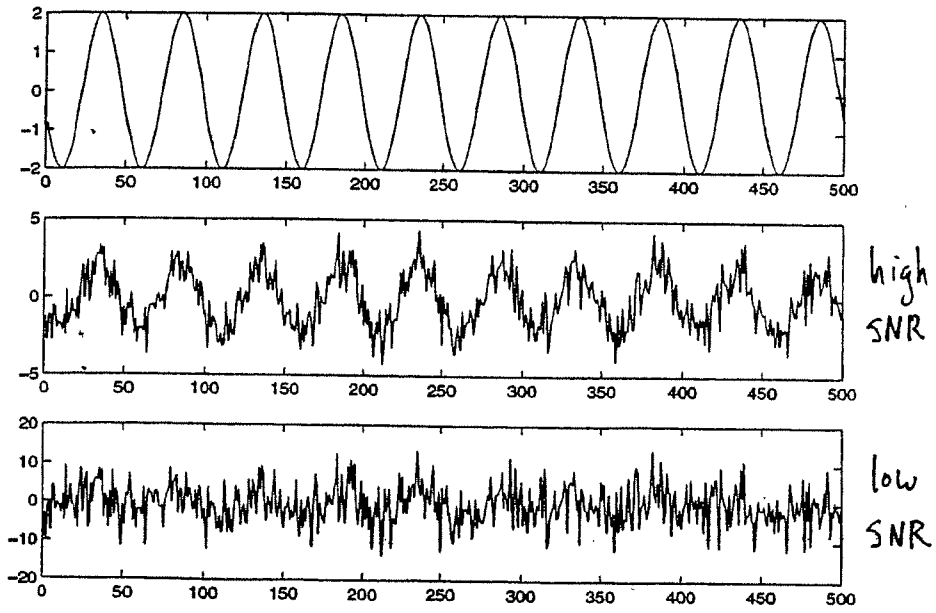
5. Sinusoid in white noise, uniform phase

$$x[n] = s[n] + w[n]$$

where

$$s[n] = A\cos(2\pi f n + \phi)$$

$$\phi \sim \text{unif}[0, 2\pi)$$



ACF / Spectrum Estimation

Throughout our discussion of filtering we will assume knowledge of first and second order moments. In practice, however, these may also need to be estimated.

There is an extensive literature on how to estimate an ACF or its Fourier transform, the power spectral density (PSD).

These are important topics that are beyond the scope of this discussion.

LINEAR PREDICTION

Linear prediction is an application of LMMSE estimation theory that is widely used in speech processing, spectral estimation, and elsewhere.

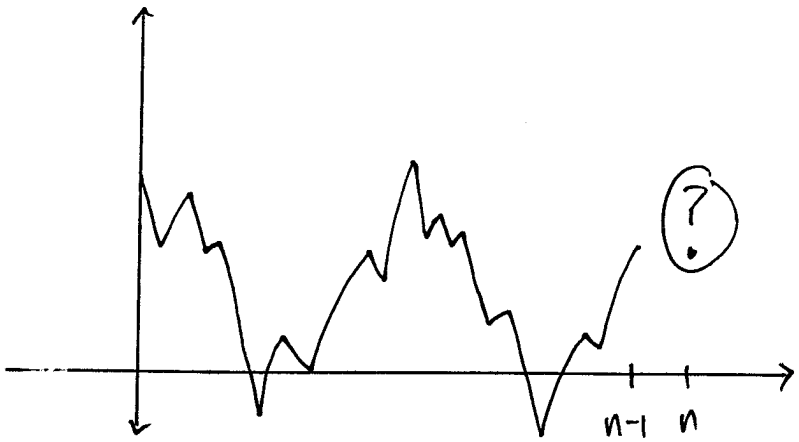
Let $\{x[n]\}$ be a zero-mean, WSS random process with ACF

$$r_{xx}[k] = E\{x[n]\, x[n+k]\}.$$

We observe

$$\underline{x} = [x[n-1] \cdots x[n-p]]^T$$

and our task is to predict the next value $x[n]$.



Let's view $x[n]$ as an unknown parameter to be estimated:

$$\theta = x[n]$$

A linear estimator for $x[n]$ has the form

$$\hat{\theta} = \hat{x}[n] = \sum_{k=1}^{p} h_p[k]\, x[n-k].$$

We know that the optimal predictor coefficients must satisfy the Wiener-Hopf equations

$$R_{xx}\, \underline{h}_p = R_{x\theta}$$

where $\underline{h}_p = [h_p[1] \cdots h_p[p]]^T$.

To simplify notation, denote

$$r_k = r_{xx}[k].$$

Then

$$R_{xx} = E\{\underline{x}\,\underline{x}^T\} = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots & r_{p-2} \\ r_2 & r_1 & r_0 & & \vdots \\ \vdots & \vdots & & \ddots & \\ r_{p-1} & r_{p-2} & \cdots & & r_0 \end{bmatrix} =: R_p$$

Toeplitz ↗

and

$$R_{x\theta} = E\left\{ \underline{x} \cdot x[n] \right\} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} =: \underline{r}_p$$

Then the optimal linear predictor satisfies

$$R_p \cdot \underline{h}_p = \underline{r}_p$$

$$\Updownarrow$$

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} h_p[1] \\ h_p[2] \\ \vdots \\ h_p[p] \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix}$$

Mean-Squared Prediction Error

$$E\left\{ \left(x[n] - \hat{x}[n] \right)^2 \right\}$$

$$=$$

(a)

Base of recursion

The WH equations for $p = 1$ are

$$[r_0] \cdot [h_1[1]] = [r_1]$$

$$\Rightarrow \quad h_1[1] = \frac{r_1}{r_0} .$$

This will initialize the recursive algorithm

General recursion

We wish to express

$$\underline{h}_p = \begin{bmatrix} \underline{h}_{p-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \underline{d}_{p-1} \\ k_p \end{bmatrix}$$

The goal is to find $\underline{d}_{p-1}$, k_p.

Our approach is to exploit the recursive structure of the WH equations.

According to the WH equations

$$R_p \cdot \underline{h}_p = \underline{r}_p .$$

$$\Updownarrow$$

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{p-1} \\ r_1 & r_0 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & \\ r_{p-1} & r_{p-2} & & r_0 \end{bmatrix} \begin{bmatrix} h_p(1) \\ h_p(2) \\ \vdots \\ h_p(p) \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix}$$

Now observe

$$R_p = \begin{bmatrix} R_{p-1} & \tilde{\underline{r}}_{p-1} \\ \tilde{\underline{r}}_{p-1}^T & r_0 \end{bmatrix} , \quad \underline{r}_p = \begin{bmatrix} \underline{r}_{p-1} \\ r_p \end{bmatrix}$$

where $\tilde{\underline{r}}_p :=$ "$\underline{r}_p$ upside-down". Plugging in the recursive formula for $\underline{h}_p$ we get

$$\begin{bmatrix} R_{p-1} & \tilde{\underline{r}}_{p-1} \\ \tilde{\underline{r}}_{p-1}^T & r_0 \end{bmatrix} \cdot \left\{ \begin{bmatrix} \underline{h}_{p-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \underline{d}_{p-1} \\ k_p \end{bmatrix} \right\} = \begin{bmatrix} \underline{r}_{p-1} \\ r_p \end{bmatrix}$$

This gives us a system of equations

$$R_{p-1}\,\underline{h}_{p-1} + R_{p-1}\,\underline{d}_{p-1} + \tilde{\underline{r}}_{p-1}\,k_p = \underline{r}_{p-1} \qquad (1)$$

$$\tilde{\underline{r}}_{p-1}^{T}\cdot\underline{h}_{p-1} + \tilde{\underline{r}}_{p-1}^{T}\cdot\underline{d}_{p-1} + r_0\,k_p = r_p \qquad (2)$$

Recall: we're assuming $\underline{h}_{p-1}$ known and we need to solve for $\underline{d}_{p-1}$, k_p.

How can we simplify the first equation?

$$R_{p-1}\,\underline{h}_{p-1} = \underline{r}_{p-1} \quad \Longrightarrow \quad \underline{d}_{p-1} = -k_p\, R_{p-1}^{-1}\,\tilde{\underline{r}}_{p-1}$$

Exercise | Show that $R_{p-1}^{-1}\,\tilde{\underline{r}}_{p-1} = \tilde{\underline{h}}_{p-1}$, or equivalently, $R_{p-1}\,\tilde{\underline{h}}_{p-1} = \tilde{\underline{r}}_{p-1}$. For concreteness, you may wish to consider $p = 3$ or 4.

Solution | Suppose

$$A \cdot \underline{b} = \underline{c}$$

where A is $m \times n$, $\underline{b}$ is $n \times 1$, $\underline{c}$ is $m \times 1$.

When is $A\tilde{\underline{b}} = \tilde{\underline{c}}$?

If $A = (a_{ij})$, then define

$$\tilde{A} = (\tilde{a}_{ij})$$

where $\tilde{a}_{ij} = a_{m-i,\, n-j}$. Then

$$A\underline{b} = \underline{c} \iff \sum_{j=1}^{n} a_{ij} b_j = c_i \quad \forall i$$

$$\iff \sum_{j=1}^{n} a_{i,n-j} b_{n-j} = c_i \quad \forall i$$

$$\iff \sum_{j=1}^{n} a_{m-i,n-j} b_{n-j} = c_{m-i} \quad \forall i$$

$$\iff \sum_{j=1}^{n} \tilde{a}_{ij} \tilde{b}_j = \tilde{c}_i$$

$$\iff \tilde{A} \cdot \tilde{\underline{b}} = \tilde{\underline{c}}$$

So we need $A = \tilde{A}$. This is true when A is symmetric and Toeplitz, as is the case for R_{p-1}.

Conclusion: $\boxed{\underline{d}_{p-1} = -k_p \, \underline{\tilde{h}}_{p-1}}$

Thus far we have shown

$$\underline{h}_p = \begin{bmatrix} \underline{h}_{p-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \underline{d}_{p-1} \\ k_p \end{bmatrix} = \begin{bmatrix} \underline{h}_{p-1} \\ 0 \end{bmatrix} + k_p \begin{bmatrix} -\underline{h}_{p-1} \\ 1 \end{bmatrix}.$$

It remains to find k_p.

Recall equation (2):

$$\underline{\tilde{r}}_{p-1}^T \, \underline{h}_{p-1} + \underline{\tilde{r}}_{p-1}^T \, \underline{d}_{p-1} + r_0 k_p = r_p.$$

Plugging in $\underline{d}_{p-1} = -k_p \underline{\tilde{h}}_{p-1}$ we get

$$\underline{\tilde{r}}_{p-1}^T \, \underline{h}_{p-1} + k_p \left(-\underline{\tilde{r}}_{p-1}^T \, \underline{\tilde{h}}_{p-1} + r_0 \right) = r_p$$

or

$$\boxed{k_p = \frac{r_p - \underline{\tilde{r}}_{p-1}^T \, \underline{h}_{p-1}}{-\underline{r}_{p-1}^T \, \underline{h}_{p-1} + r_0}}$$

The two boxed formulas define the general Levinson-Durbin recursion.

Extensions

The L-D algorithm may be extended to other settings including

1. ℓ-step prediction: given $x[n-1], \ldots, x[n-p]$, predict $x[n+\ell]$
2. multiple predictions: given $x[n-1], \ldots, x[n-p]$, predict $x[n], \ldots, x[n+1]$.

Application: Linear Predictive Coding

Given: A signal $x[0], x[1], \ldots, x[N]$

Task: Compress signal (store with as few bits as possible while still providing an accurate representation.

Idea: Store the first value as is, and encode the prediction errors of the remaining values:

$$e[n] = x[n] - \hat{x}[n]$$

where

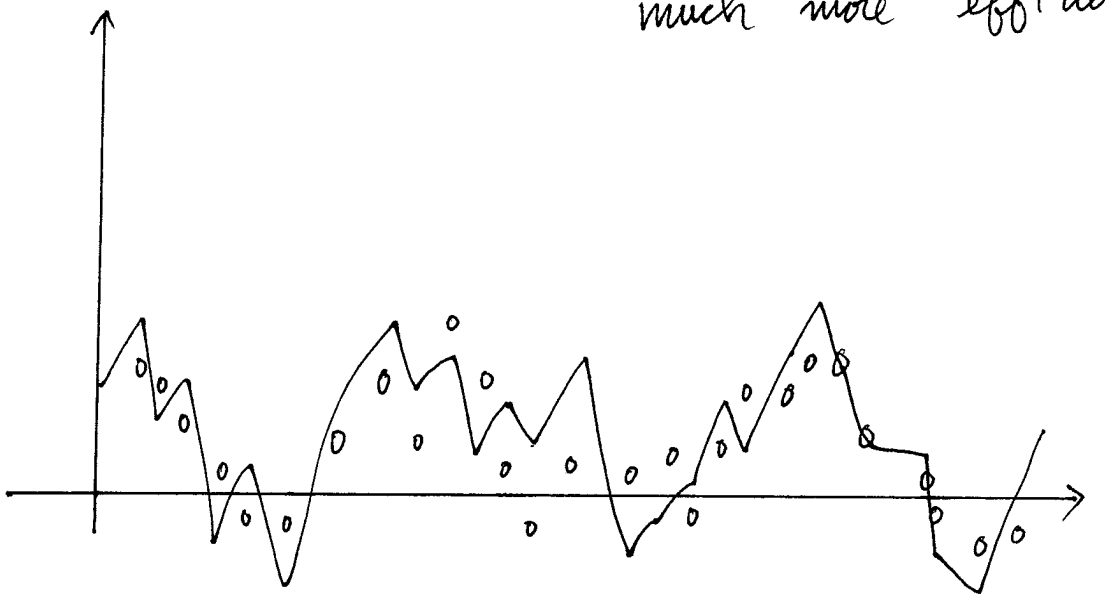
$$\hat{x}[n] = \sum_{k=1}^{n} h_n[k]\, x[n-k]$$

In other words, there are two equivalent representations

$$x[0],\ x[1],\ \ldots,\ x[N]$$

$$x[0],\ e[1],\ \ldots,\ e[N]$$

but the prediction errors have much smaller variance and can therefore be encoded much more efficiently.



LPC is a key ingredient in modern speech processing applications such as compression and synthesis. Since speech is not a stationary process, speech signals are encoded in short blocks and signal characteristics are updated for each block (otherwise the prediction errors would get really big).

Summary

- Linear prediction: application of LMMSE theory
- WSS process $\Rightarrow$ Toeplitz auto-covariance matrix
- Toeplitz ACV matrix $\Rightarrow$ Levinson-Durbin algorithm for fast matrix inversion and predictor coefficient updates.
- Important application: LPC, widely used in speech modeling.

Key a.

$$E\{(x[n]-\hat{x}[n])^2\}$$

$$= E\{(x[n]-\underline{h}_p^T\underline{x})(x[n]-\underline{h}_p^T\underline{x})\}$$

$$= E\{(x[n]-\underline{h}_p^T\underline{x})\cdot x[n]\}$$

$$= E\{x[n]^2\} - \underline{r}_p^T R_p^{-1} E\{\underline{x}\,x[n]\}$$

$$= r_0 - \underline{r}_p^T R_p^{-1}\underline{r}_p$$

by orthogonality principle

WIENER FILTERING

Wiener filtering is the application of LMMSE estimation to recovery of a signal in additive noise under wide sense stationarity assumptions.

Problem Statement

$$x[n] = s[n] + w[n]$$

$x[n]$: observation; $s[n]$: signal of interest; $w[n]$: noise

We observe $x[n], x[n-1], \ldots, x[n-p+1]$ and would like to estimate

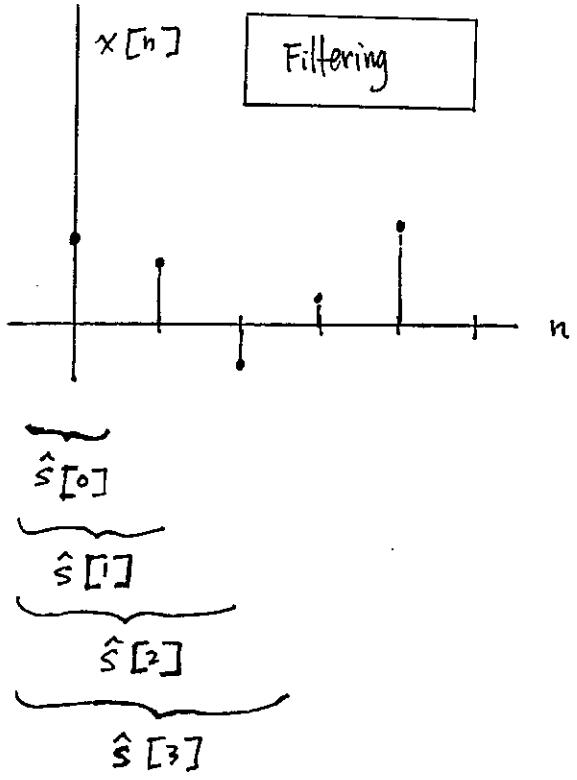
$$\theta = s[n+D]$$

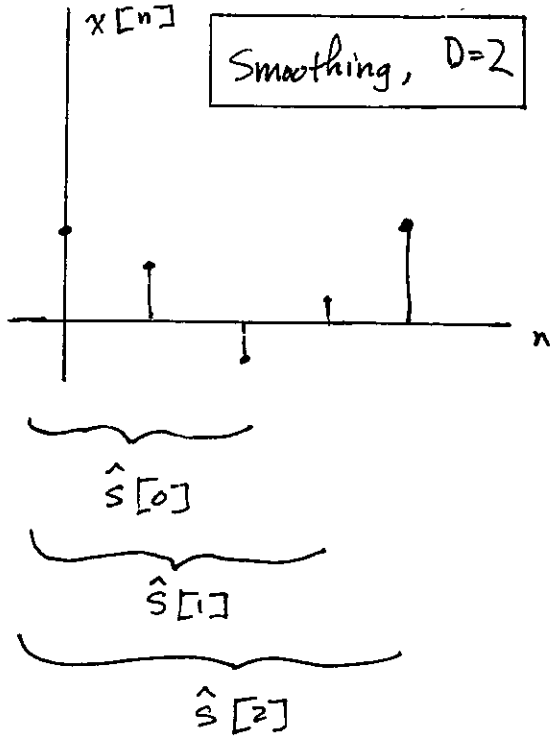
where D is an integer, using a linear estimator

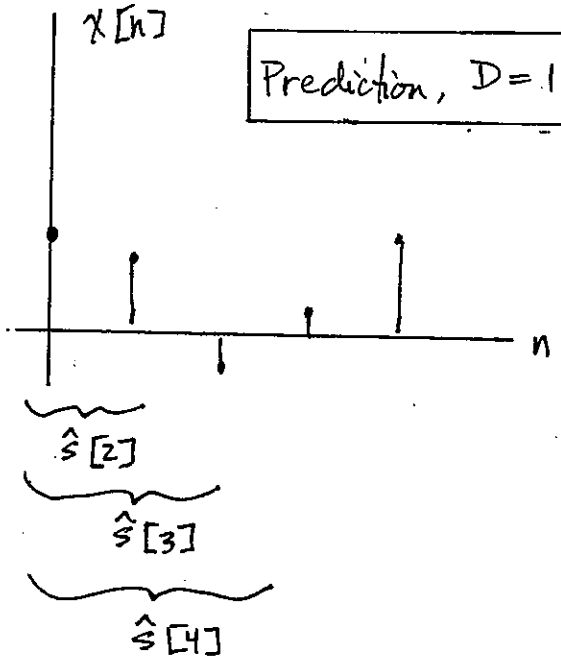
$$\hat{\theta} = \hat{s}[n+D] = \sum_{k=0}^{p-1} h_p[k]\, x[n-k]$$

Three Cases

1. $D = 0$ — filtering
2. $D > 0$ — signal prediction
3. $D < 0$ — smoothing



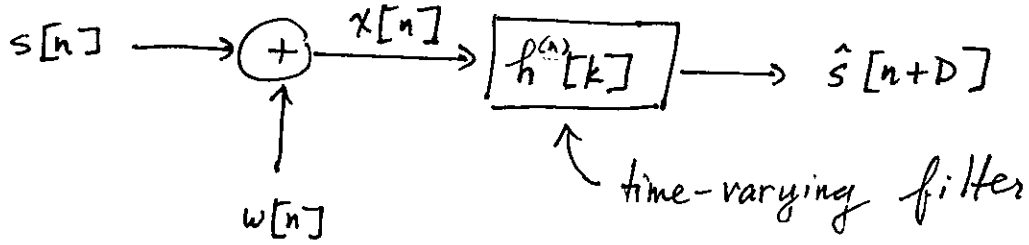




Note] Signal prediction is different from the "measurement prediction" problem discussed previously:

$$\hat{s}[n+D] \neq \hat{x}[n+D].$$

Filtering Interpretation



Smoothing requires a <u>noncausal</u> filter in the sense that, to estimate a present signal value, you need data from the future.

Filtering and prediction, on the other hand, are <u>causal</u> operations.

Assumptions

We assume all first and second order moments are known, as required for LMMSE estimation. Furthermore, we assume

1. $s[n]$ and $w[n]$ are zero mean
2. $x[n]$ is wide-sense stationary (WSS) with autocorrelation
$$r_{xx}[k] = E\{x[n]\, x[n+k]\}$$
3. $x[n]$ and $s[n]$ are jointly WSS with cross-correlation
$$r_{xs}[k] = E\{x[n]\, s[n+k]\}$$

Example

These conditions hold when $s[n]$ and $w[n]$ are zero-mean, WSS, and uncorrelated.

Wiener – Hopf Equations

Let's focus on the filtering problem ($D = 0$)

From LMMSE estimation theory we know the optimal filter satisfies the Wiener–Hopf equations:

$$\boxed{R_{xx}\,\underline{h}_p = R_{x\theta}}$$

where R_{xx} and $R_{x\theta}$ are given in terms of $r_{xx}[k]$ and $r_{xs}[k]$.

So in theory, we can compute the Wiener filter. In practice, however, we want a fast, online <u>algorithm</u> for computing and updating $\underline{h}_p$ as data streams in.

Generalized Levinson-Durbin Algorithm

The Wiener-Hopf equations are

$$\begin{bmatrix} r_{xx}[0] & r_{xx}[1] & \cdots & r_{xx}[p-1] \\ r_{xx}[1] & r_{xx}[0] & \cdots & r_{xx}[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}[p-1] & r_{xx}[p-2] & \cdots & r_{xx}[0] \end{bmatrix} \begin{bmatrix} h_p[0] \\ h_p[1] \\ \vdots \\ h_p[p-1] \end{bmatrix} = \begin{bmatrix} r_{xs}[0] \\ r_{xs}[1] \\ \vdots \\ r_{xs}[p-1] \end{bmatrix}$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$R_p \qquad\qquad \underline{h}_p \qquad\qquad \underline{b}_p$$

This system of equations is very similar to the WH equation for linear (measurement) prediction. The difference is that $\underline{b}_p$ is not related to the rows/columns of R_p.

In linear prediction we had

$$R_p \cdot \underline{h}_p = \underline{r}_p, \qquad \underline{r}_p = \begin{bmatrix} r_{xx}[1] \\ r_{xx}[2] \\ \vdots \\ r_{xx}[p] \end{bmatrix}$$

Let's try to update $\underline{h}_p$ from $\underline{h}_{p-1}$:

$$\underline{h}_p = \begin{bmatrix} \underline{h}_{p-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \underline{d}_{p-1} \\ k_p \end{bmatrix}$$

Recall

$$R_p = \begin{bmatrix} R_{p-1} & \underline{\tilde{r}}_{p-1} \\ \underline{\tilde{r}}_{p-1}^T & r_{xx}[0] \end{bmatrix}$$

$\underline{\tilde{r}}_p = \underline{r}_p$ upside-down

and also notice

$$\underline{b}_p = \begin{bmatrix} \underline{b}_{p-1} \\ r_{xs}[p-1] \end{bmatrix}$$

Thus we may write the WH equations

$$\begin{bmatrix} R_{p-1} & \underline{\tilde{r}}_{p-1} \\ \underline{\tilde{r}}_{p-1}^T & r_{xx}[0] \end{bmatrix} \cdot \left\{ \begin{bmatrix} \underline{h}_{p-1} \\ 0 \end{bmatrix} + \begin{bmatrix} \underline{d}_{p-1} \\ k_p \end{bmatrix} \right\} = \begin{bmatrix} \underline{b}_{p-1} \\ r_{xs}[p-1] \end{bmatrix}$$

$$\Updownarrow$$

$$R_{p-1}\,\underline{h}_{p-1} + R_{p-1}\,\underline{d}_{p-1} + k_p\,\underline{\tilde{r}}_{p-1} = \underline{b}_{p-1} \qquad (1)$$

$$\underline{\tilde{r}}_{p-1}^T\,\underline{h}_{p-1} + \underline{\tilde{r}}_{p-1}^T\,\underline{d}_{p-1} + k_p\,r_{xx}[0] = r_{xs}[p-1] \qquad (2)$$

To simplify (1), observe

$$R_{p-1}\, \underline{h}_{p-1} = \underline{b}_{p-1}$$

which implies

$$\underline{d}_{p-1} =$$

How can this be simplified?

Recall the WH equations for linear prediction:

$$R_p \, \underline{g}_p = \underline{r}_p$$

where $\hat{x}[n] = \sum_{k=1}^{p} g_p[k] \, x[n-k]$ is the LMMSE predictor of $x[n]$.

Previously we used the fact that R_p is symmetric and Toeplitz to show

$$R_{p-1} \, \underline{\tilde{g}}_{p-1} = \underline{\tilde{r}}_{p-1}$$

$$\Rightarrow \quad R_{p-1}^{-1} \, \underline{\tilde{r}}_{p-1} = \underline{\tilde{g}}_{p-1}$$

$$\Rightarrow \quad \boxed{\underline{d}_{p-1} = -k_p \cdot \underline{\tilde{g}}_{p-1}}$$

Exercise | Determine k_p

Solution: From equation (2), and plugging in $\underline{d}_{p-1} = -k_p \underline{\tilde{g}}_{p-1}$, we obtain

$$\underline{\tilde{r}}_{p-1}^T \underline{h}_{p-1} + k_p \left(r_{xx}[0] - \underline{\tilde{r}}_{p-1}^T \underline{\tilde{g}}_{p-1} \right) = r_{xs}[p-1]$$

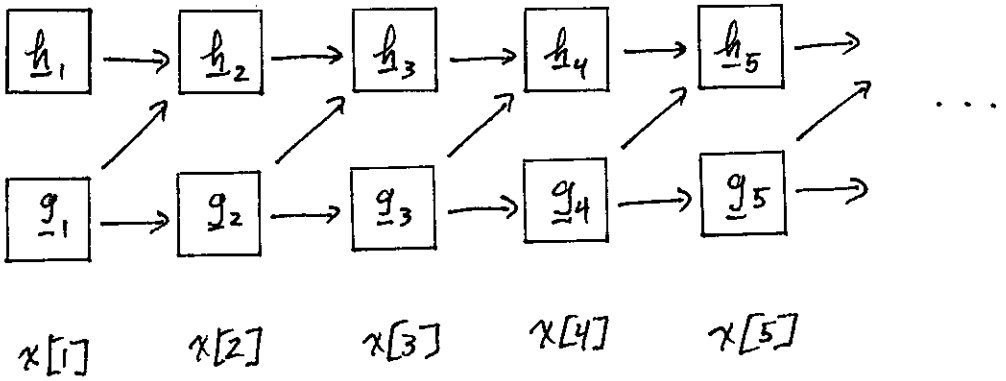
$$\Rightarrow \boxed{k_p = \frac{r_{xs}[p-1] - \underline{\tilde{r}}_{p-1}^T \underline{h}_{p-1}}{r_{xx}[0] - \underline{r}_{p-1}^T \underline{g}_{p-1}}}$$

This leads to the Generalized Levinson-Durbin algorithm:

(a) Initialize $\underline{h}_1 =$ $\underline{g}_1 =$

Iterate

1. Update $\underline{g}_p$ from $\underline{g}_{p-1}$ using Levinson-Durbin recursion
2. Update $\underline{h}_p$ from $\underline{h}_{p-1}$ and $\underline{g}_{p-1}$



For prediction and smoothing, a similar algorithm can be derived. The WH equations are

$$R_p \underline{h}_p = \underline{b}_p$$

where

$$\underline{b}_p = \begin{bmatrix} r_{xs}[D] \\ r_{xs}[D+1] \\ \vdots \\ r_{xs}[D+p-1] \end{bmatrix}$$

The GLD recursion changes only slightly.

In general, the GLD algorithm requires $O(p^2)$ operations to compute $\{\underline{h}_1, \ldots, \underline{h}_p\}$.

IIR Wiener Filtering

So far we have discussed "FIR Wiener Filtering," because $\hat{\theta} = \hat{s}[n+D]$ only depends on a finite number of observations $x[n], x[n-1], \ldots, x[n-p+1]$, which implies $\underline{h}_p$ has finitely many nonzero taps.

We will consider two IIR problems

- The causal, IIR Wiener filter

$$\hat{s}[n] = \sum_{k=0}^{\infty} h[k]\, x[n-k]$$

- The noncausal, IIR Wiener smoother

$$\hat{s}[n] = \sum_{k=-\infty}^{\infty} h[k]\, x[n-k]$$

Infinite Wiener Smoother

Given $\{x[n]\}_{n=-\infty}^{\infty}$ we seek the filter $\{h[k]\}_{k=-\infty}^{\infty}$ such that

$$E\left\{\left(s[n] - \sum_{k=-\infty}^{\infty} h[k]\, x[n-k]\right)^2\right\}$$

is minimized.

Observe that $\hat{s}[n]$ is the projection of $s[n]$ onto the closed linear span of $\{x[n]\}_{n=-\infty}^{\infty}$.

By the orthogonality principle, we know

$$E\left\{(s[n] - \hat{s}[n]) \cdot x[n-l]\right\} = 0 \quad \forall l$$

$$\Updownarrow$$

$$E\left\{ \left(s[n] - \sum_{k=-\infty}^{\infty} h[k]\, x[n-k] \right) x[n-\ell] \right\} = 0 \qquad \forall \ell$$

$$\Updownarrow$$

$$E\left\{ x[n-\ell]\, s[n] \right\} = \sum_{k=-\infty}^{\infty} h[k]\, E\left\{ x[n-\ell]\, x[n-k] \right\}$$

$$\Updownarrow$$

$$\boxed{\sum_{k=-\infty}^{\infty} h[k]\, r_{xx}[\ell-k] = r_{xs}[\ell] \qquad \forall \ell \in \mathbb{Z}}$$

These are the Wiener-Hopf equations for the infinite Wiener smoother.

How can we solve for $\{h[k]\}_{k=-\infty}^{\infty}$?

Take the DTFT of both sides:

$$\text{DTFT}\{ h * r_{xx} \} = \text{DTFT}\{ r_{xs} \}$$

$$\| \qquad\qquad \|$$

$$H(f) \cdot P_{xx}(f) \qquad P_{xs}(f)$$

$P_{xx}(f)$: spectral density

$P_{xs}(f)$: cross spectral density

$$\Rightarrow \boxed{H(f) = \frac{P_{xs}(f)}{P_{xx}(f)}}$$

When signal and noise are uncorrelated

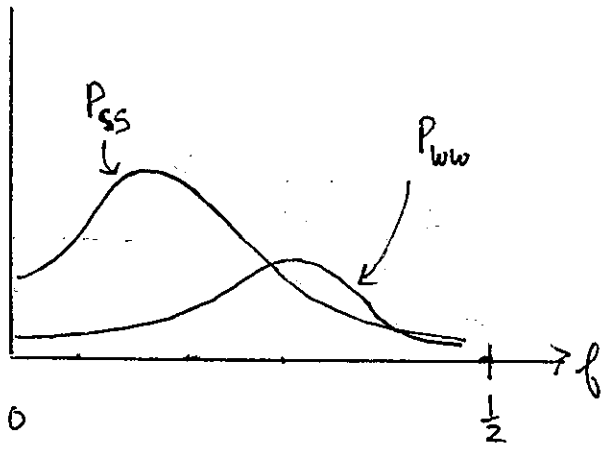
$P_{xs}(f) =$

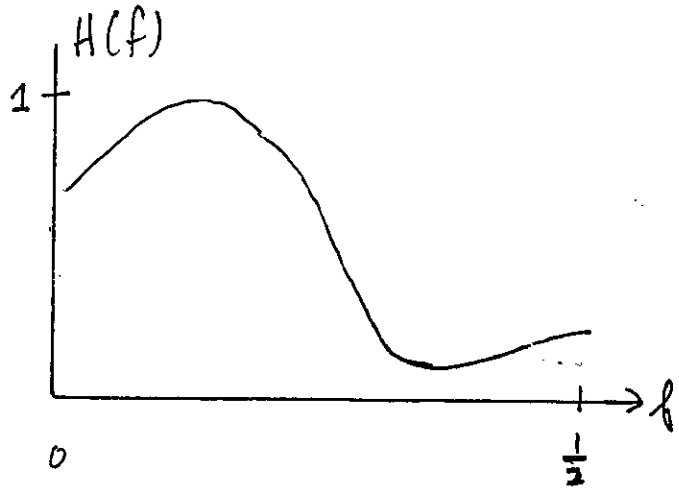
$P_{xx}(f) =$

and therefore

$$\boxed{H(f) =}$$

Interpretation:





Infinite Wiener Filter

Now let's try to estimate $s[n]$ based on data from the present and infinite past.

$$\hat{s}[n] = \sum_{k=0}^{\infty} h[k]\, x[n-k]$$

As before, we may apply the orthogonality principle to arrive at the Wiener-Hopf equations:

$$\sum_{k=0}^{\infty} h[k]\, r_{xx}[l-k] = r_{xs}[l] \qquad \forall l \geq 0$$

However, since these equations only hold for $l \geq 0$ (as opposed to all $l \in \mathbb{Z}$), it is not true that $h * r_{xx} = r_{xs}$.

Therefore, we can not solve for $h[k]$ or $H(f)$ by simply taking the DTFT.

It can be shown that

$$H(z) = \frac{1}{G(z)} \cdot \left[\frac{P_{xs}(z)}{G(z^{-1})} \right]_+$$

where

- $P_{xx}(z) := \sum_{k=-\infty}^{\infty} r_{xx}[k] z^{-k}$

  $= G(z) G(z^{-1})$

  spectral factorization

  ↖ minimum phase, causal

- $[Y(z)]_+ := \sum_{k=0}^{\infty} y[k] z^{-k}$,

  the z-transform of the causal part of $\{y[k]\}_{k=-\infty}^{\infty}$

These IIR Wiener estimators are not just theoretical curiosities. For large p, $\underline{h}_p$ is well-approximated by its IIR counterpart. Therefore, when sufficient data are available, it is convenient to use the IIR estimators which have convenient frequency domain implementations.

Summary

- Wiener filtering $\Leftrightarrow$ LMMSE recovery of signal in additive noise, assuming WSS
- Three basic problems: prediction, filtering, smoothing
- Generalized Levinson-Durbin $\Rightarrow$ efficient algorithm for updating FIR filter for streaming apps.
- IIR estimators obtained by orthogonality principle and frequency/z-transform domain techniques.

Key

a. $\underline{h}_1 = \dfrac{r_{xs}[0]}{r_{xx}[0]}$, $\underline{g}_1 = \dfrac{r_{xx}[1]}{r_{xx}[0]}$

b.

$$P_{ss}(f) = \text{DTFT}\{r_{xs}\} = \text{DTFT}\{r_{ss}\} = P_{ss}(f)$$

$$R_{xx}(f) = P_{ss}(f) + P_{ww}(f)$$

$$H(f) = \frac{P_{ss}(f)}{P_{ss}(f) + P_{ww}(f)}$$

KALMAN FILTERING

The Kalman filter is an important generalization of the Wiener filter.

Wiener filter	Kalman filter
• WSS process	• Gauss-Markov process
• Data stream goes back into the infinite past	• Data stream starts at a specific point in time
• scalar signals	• vector signals
• Nonadaptive	• Adaptive: model may evolve over time

Problem Statement

The goal is to estimate a time-varying state vector $\underline{s}(n) \in \mathbb{R}^p$ based on observations

$$\underline{x}(0), \underline{x}(1), \ldots, \underline{x}(n) \in \mathbb{R}^M$$

Linear Dynamical Model

We assume the following model:

$$\underline{s}(n) = A\,\underline{s}(n-1) + B\,\underline{u}(n)$$
$$\underline{x}(n) = H\,\underline{s}(n) + \underline{w}(n) \qquad n \geq 0$$

where

A: $p \times p$ state transition matrix

$\underline{u}(n)$: $r \times 1$ zero-mean excitation or driving noise with covariance Q, uncorrelated for different n

B: $p \times r$ state noise matrix

H: $M \times 1$ observation matrix

$\underline{w}(n)$: $M \times 1$ observation noise with diagonal covariance matrix $R_w(n)$, uncorrelated for different n

↖ time varying

We assume

- $A, B, R_w(n), H$ and Q are known
- A, B, H, Q may be time-varying also

The Kalman filter is the LMMSE estimator of $\underline{s}(n)$ based on $\underline{X}(n) := \{\underline{x}(0), \ldots, \underline{x}(n)\}$

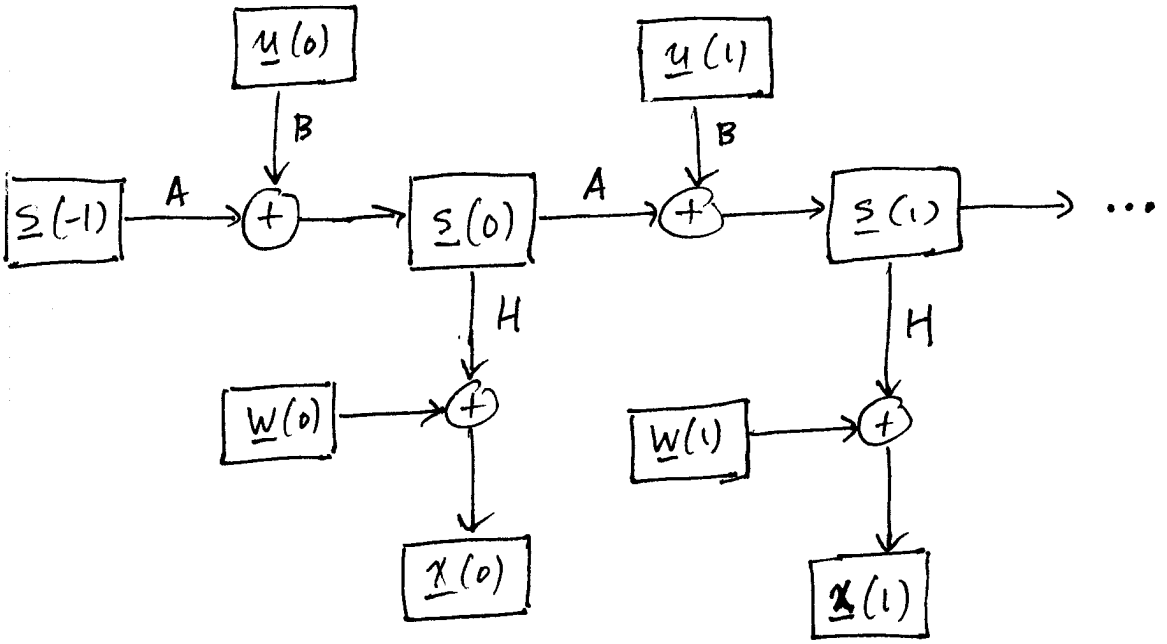
The LMMSE estimator is a Bayesian estimator: we may think of $\underline{\theta} = \underline{s}(n)$ as the parameter and

state equation $\Longleftrightarrow$ prior

observation equation $\Longleftrightarrow$ likelihood

Initial Condition

Our linear dynamical model assumes the system starts at time $n=0$. Hence it is necessary to specify an initial condition $\underline{s}(-1)$, which may be fixed or random with mean $\underline{\mu}_0$, covariance Σ_0.



Example 1st order Gauss-Markov process

Suppose we observe a signal in noise

$$x(n) = s(n) + w(n) \qquad \text{(scalar equation)}$$

where $w(n) \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$.

Furthermore, assume the signal obeys

$$s(n) = a \cdot s(n-1) + u(n)$$

where $u(n) \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2)$, with the initial condition

$$s(-1) \sim \mathcal{N}(\mu_0, \sigma_0^2) \quad \left\{ \text{independent of } \{u(n)\}. \right.$$

We say $\{s(n)\}$ is a <u>first-order Gauss-Markov process.</u>

To write this in the linear dynamical model form,

$$\underline{s}(n) = [s(n)] \quad (1\times1) \qquad \underline{u}(n) = u(n)$$

$$A = [a] \quad (1\times1) \qquad B = [1] \quad (1\times1)$$

$$H = [1] \quad (1\times1) \qquad R_w(n) = [\sigma_w^2] \quad (1\times1)$$

$$Q = [\sigma_u^2] \quad (1\times1)$$

Exercise | p^{th} order Gauss Markov process

Assume the same setup as before, but now $\{s(n)\}$ is described by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + u(n)$$

Express this problem in the linear dynamical model form.

Hint: For the state vector, take

$$\underline{s}(n) = \left[s(n-p+1) \cdots s(n) \right]^T$$

Be careful not to confuse the <u>state</u> $\underline{s}(n)$ with the <u>signal</u> $s(n)$

Solution

$$\begin{bmatrix} s(n-p+1) \\ s(n-p+2) \\ \vdots \\ s(n-1) \\ s(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \begin{bmatrix} s(n-p) \\ s(n-p+1) \\ \vdots \\ s(n-2) \\ s(n-1) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(n)$$

$\underline{s}(n)$ $\quad$ A $\quad$ $\underline{s}(n-1)$ $\quad$ B $\quad$ $\underline{u}(n)$

$$x(n) = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} s(n-p+1) \\ s(n-p+2) \\ \vdots \\ s(n-1) \\ s(n) \end{bmatrix} + w(n)$$

$\underline{x}(n)$ $\quad$ H $\quad$ $\underline{s}(n)$ $\quad$ $\underline{w}(n)$

Note For a p^{th} order G. M. process, the initial condition would be

$$\underline{s}(-1) \sim N(\underline{\mu}_0, \Sigma_0)$$

independent of $\{u(n)\}$

Example: Array processing

$$\underline{x}(n) = \begin{bmatrix} x_1(n) \\ \vdots \\ x_M(n) \end{bmatrix}$$

$x_i(n)$ = signal intensity at receiver i

$$\underline{x}(n) = H \cdot \underline{s}(n) + \underline{w}(n)$$

Distortion-free measurements

$$x_i(n) = s(n) + w_i(n)$$

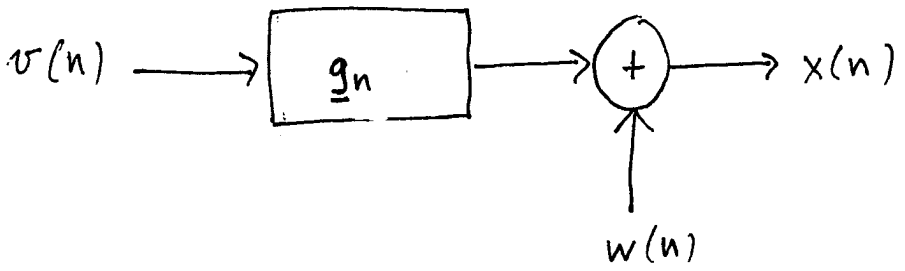
$$\Leftrightarrow \quad H = [1 \; 1 \; \cdots \; 1]^T, \quad \underline{s}(n) = \underset{(1\times 1)}{[s(n)]}$$

Distorted measurements (attenuation / blurring)

$$x_i(n) = \sum_{k=0}^{p-1} h_i(k)\, s(n-k) + w_i(n)$$

$$\Leftrightarrow \quad H = \underset{(M\times p)}{\begin{bmatrix} \underline{h}_1^T \\ \underline{h}_2^T \\ \vdots \\ \underline{h}_M^T \end{bmatrix}}, \quad \underline{s}(n) = \underset{(p\times 1)}{\begin{bmatrix} s(n) \\ s(n-1) \\ \vdots \\ s(n-p+1) \end{bmatrix}}$$

Example System Identification / Channel Estimation



$$x(n) = \sum_{k=0}^{p-1} g_n(k)\, v(n-k)$$

$v(n)$ = known signal (system probe) , $\underline{g}_n$ unknown, time-varying impulse response

$$x(n) = H \cdot \underline{s}(n) + w(n)$$

(a)

$H =$

$\underline{s}(n) =$

The Kalman Recursions

The Kalman filter is given by

$$\hat{\underline{S}}(n) = R_{\underline{S}(n)\underline{X}(n)} R_{\underline{X}(n)\underline{X}(n)}^{-1} \underline{X}(n)$$

The covariance matrices can in principle be computed from the linear dynamical model, but there are two problems:

- these calculations are tedious
- we want an <u>on-line</u> estimator, so that $\hat{\underline{S}}(n)$ can be efficiently updated from $\hat{\underline{S}}(n-1)$

Approach

There are several different ways to derive the Kalman recursions:

- innovations approach
- repeated use of the orthogonality principle (see Kay or Scharf)
- assume joint Gaussianity and apply properties of jointly Gaussian random vectors.

Joint Gaussianity

Since the LMMSE depends only on first and second order moments, we can assume the higher order moments are such that everything is jointly Gaussian.

For our linear dynamical model, this means

$$\underline{u}(n) \sim \mathcal{N}(\underline{0}, Q)$$

$$\underline{w}(n) \sim \mathcal{N}(\underline{0}, R_w(n))$$

$$\underline{s}(-1) \sim \mathcal{N}(\underline{\mu}_0, \Sigma_0)$$

} independent

With these assumptions, the state variable

$$\underline{s}(n) = A\,\underline{s}(n-1) + B\,\underline{u}(n)$$

is said to be a <u>vector Gauss-Markov process</u>.

When deriving the Kalman filter, we will assume $E[\underline{s}(n)] = \underline{0}$, $E[\underline{x}(n)] = \underline{0}$. The equations we derive will also be valid for non-zero means. This can be verified by applying the Kalman filter to $\underline{s}'(n) = \underline{s}(n) - E[\underline{s}(n)]$ and $\underline{x}'(n) = \underline{x}(n) - E[\underline{x}(n)]$, and invoking linearity properties.

Notation

MMSE estimator (our goal)

$$\underline{s}(n) \mid \underline{X}(n) \sim \mathcal{N}\left(\hat{\underline{s}}(n|n), M(n|n)\right)$$

$$\underline{s}(n) \mid \underline{X}(n-1) \sim \mathcal{N}\left(\hat{\underline{s}}(n|n-1), M(n|n-1)\right)$$

MMSE predictor

$$\underline{e}(n) = \underline{s}(n) - \hat{\underline{s}}(n|n-1)$$ "prediction error"

$$R(n) = E\left[\underline{s}(n) \cdot \underline{s}(n)^T\right]$$ "state covariance"

Kalman Filter Derivation

Step 1a: $\underline{s}(n) \mid \underline{X}(n-1) \sim \underline{s}(n) \mid \hat{\underline{s}}(n|n-1)$

This follows from a more general fact:

If $(\underline{\theta}, \underline{Y})$ are jointly Gaussian, zero mean, then

$$\underline{\theta} \mid \underline{Y} \sim \underline{\theta} \mid E[\underline{\theta} \mid \underline{Y}].$$

This seems intuitive because

$$\underline{\theta} \mid \underline{Y} \sim \mathcal{N}\left(E[\underline{\theta} \mid \underline{Y}],\ \mathrm{Cov}(\underline{\theta} \mid \underline{Y})\right)$$

↑ dependence on $\underline{Y}$ manifested in posterior mean

↑ independent of the actual value of $\underline{Y}$

Let's prove it rigorously. Recall

$$E[\underline{\theta} \mid \underline{Y}] = R_{\theta Y} R_{YY}^{-1} \underline{Y}$$

$$\mathrm{Cov}(\underline{\theta} \mid \underline{Y}) = R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta}$$

Note that $E[\underline{\theta} \mid \underline{Y}]$ is a linear transformation of a Gaussian, so it is also Gaussian.

We have

$$\begin{bmatrix} \underline{\theta} \\ E[\underline{\theta}|\underline{Y}] \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & R_{\theta Y} R_{YY}^{-1} \end{bmatrix} \begin{bmatrix} \underline{\theta} \\ \underline{Y} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \underline{\theta} \\ E[\underline{\theta}|\underline{Y}] \end{bmatrix} \sim \mathcal{N}\left(\underline{0}, \begin{bmatrix} R_{\theta\theta} & R_{\theta Y} R_{YY}^{-1} R_{Y\theta} \\ R_{\theta Y} R_{YY}^{-1} R_{Y\theta} & R_{\theta Y} R_{YY}^{-1} R_{Y\theta} \end{bmatrix} \right)$$

$$\Rightarrow \underline{\theta} \,|\, E[\underline{\theta}|\underline{Y}] \sim \mathcal{N}\left(E[\underline{\theta}|\underline{Y}],\ R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta} - \left(R_{\theta Y} R_{YY}^{-1} R_{Y\theta}\right)^{-1} R_{\theta Y} R_{YY}^{-1} R_{Y\theta} \right)$$

$$\sim \mathcal{N}\left(E[\underline{\theta}|\underline{Y}],\ R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta} \right)$$

$$\sim \underline{\theta} \,|\, \underline{Y}$$

Now apply this result with $\underline{\theta} = \underline{s}(n)$, $\underline{Y} = \underline{X}(n-1)$,

and $E[\underline{\theta}|\underline{Y}] = \hat{\underline{s}}(n|n-1) \equiv E\left[\underline{s}(n) \,|\, \underline{X}(n-1)\right]$ ▨

Remark

As a byproduct, we also showed

$$E\left[\hat{\underline{s}}(n|n-1)\, \underline{s}(n)^T\right] = R_{\theta Y} R_{YY}^{-1} R_{Y\theta}$$

$$= R_{\theta\theta} - \left(R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta}\right)$$

$$= R(n) - M(n|n-1)$$

Step 1b: $\underline{s}(n) \mid \underline{X}(n) \sim \underline{s}(n) \mid \hat{\underline{s}}(n|n-1), \underline{x}(n)$

We just showed $\underline{s}(n) \mid \underline{X}(n-1)$ and $\underline{s}(n) \mid \hat{\underline{s}}(n|n-1)$ have the same distribution. Now condition both random variables on $\underline{x}(n)$ ▨

Conclusion from Step 1:

$$\hat{s}(n|n) = E[\underline{s}(n) \mid \underline{X}(n)]$$

$$= E[\underline{s}(n) \mid \hat{\underline{s}}(n|n-1), \underline{x}(n)]$$

↑ we will focus on computing

Step 2

$$\begin{bmatrix} \underline{s}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{x}(n) \end{bmatrix} = \begin{bmatrix} I & I & 0 \\ 0 & I & 0 \\ H & H & I \end{bmatrix} \begin{bmatrix} \underline{e}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{w}(n) \end{bmatrix}$$

because

$$\underline{s}(n) = \left(\underline{s}(n) - \hat{\underline{s}}(n|n-1)\right) + \hat{\underline{s}}(n|n-1)$$

$$= \underline{e}(n) + \hat{\underline{s}}(n|n-1)$$

Outline of rest of derivation:

Step 3: determine covariance of $\begin{bmatrix} \underline{e}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{w}(n) \end{bmatrix}$

Step 4: determine covariance of $\begin{bmatrix} \underline{s}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{x}(n) \end{bmatrix}$

Step 5: compute $\hat{\underline{s}}(n|n) = E\left[\underline{s}(n) \mid \hat{\underline{s}}(n|n-1), \underline{x}(n)\right]$

Step 6: identify recursive structure in ↗

Step 3

$$E\left\{ \begin{bmatrix} \underline{e}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{w}(n) \end{bmatrix} \begin{bmatrix} \underline{e}(n)^T & \hat{\underline{s}}(n|n-1)^T & \underline{w}(n)^T \end{bmatrix} \right\} = \begin{bmatrix} M(n|n-1) & 0 & 0 \\ 0 & R(n)-M(n|n-1) & 0 \\ 0 & 0 & R_w(n) \end{bmatrix}$$

There are nine smaller covariances to compute.

The five involving $\underline{w}(n)$ are easy

$$E\left[\underline{w}(n)\underline{w}(n)^T\right] = R_w(n) \quad \text{by definition of } R_w(n)$$

The other four are 0 because $\underline{e}(n)$ and $\hat{\underline{s}}(n|n-1)$ are linear combinations of $\{\underline{s}(-1), \underline{y}(0), \ldots, \underline{y}(n), \underline{w}(0), \ldots, \underline{w}(n-1)\}$, all of which are independent of $\underline{w}(n)$.

Next, lets show $E\left[\underline{e}(n)\,\hat{\underline{s}}(n|n-1)^T\right] = 0$.

Any guesses?

By the orthogonality principle, the LMMSE estimator is orthogonal to its prediction error. Equivalently, this follows from the Wiener-Hopf equations. If you're still not convinced, you can use properties of Gaussians to show

$$E\left[\underline{s}(n)\,\hat{\underline{s}}(n|n-1)^T\right] = E\left[\hat{\underline{s}}(n|n-1)\,\hat{\underline{s}}(n|n-1)^T\right]$$

using arguments like those in step 1a.

By symmetry, $E\left[\hat{\underline{s}}(n|n-1)\,\underline{e}(n)^T\right] = 0$, so only two terms remain.

Next, we have

$$E\left[\hat{\underline{s}}(n|n-1)\,\hat{\underline{s}}(n|n-1)^T\right] = E\left[\hat{\underline{s}}(n|n-1)\,\underline{s}(n)^T\right] \qquad \text{(from above)}$$

$$= R(n) - M(n|n-1)$$

byproduct of step 1a

Finally,

$$E\left[\underline{e}(n)\,\underline{e}(n)^T\right] = E\left[\left(\underline{s}(n) - \hat{\underline{s}}(n|n-1)\right)\left(\underline{s}(n) - \hat{\underline{s}}(n|n-1)\right)^T\right]$$

$$= E\left[\left(\underline{s}(n) - \hat{\underline{s}}(n|n-1)\right)\underline{s}(n)^T\right] \qquad \text{(from above)}$$

$$= R(n) - \left(R(n) - M(n|n-1)\right) = M(n|n-1)$$

Step 4

$$E\left\{ \begin{bmatrix} \underline{s}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{x}(n) \end{bmatrix} \begin{bmatrix} \underline{s}(n)^T & \hat{\underline{s}}(n|n-1)^T & \underline{x}(n)^T \end{bmatrix} \right\} =$$

$$\begin{bmatrix} I & I & 0 \\ 0 & I & 0 \\ H & H & I \end{bmatrix} \begin{bmatrix} M(n|n-1) & 0 & 0 \\ 0 & R(n)-M(n|n-1) & 0 \\ 0 & 0 & R_w(n) \end{bmatrix} \begin{bmatrix} I & 0 & H^T \\ I & I & H^T \\ 0 & 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} R(n) & \hat{M}(n|n-1) & R(n)\,H^T \\ \hat{M}(n|n-1) & \hat{M}(n|n-1) & \hat{M}(n|n-1)\,H^T \\ H\,R(n) & H\,\hat{M}(n|n-1) & H\,R(n)\,H^T + R_v(n) \end{bmatrix}$$

$$\hat{M}(n|n-1) = R(n) - M(n|n-1)$$

$$= \left[\begin{array}{c:cc} R(n) & \hat{M}(n|n-1) & R(n)\,H^T \\ \hdashline \hat{M}(n|n-1) & \multicolumn{2}{c}{\multirow{2}{*}{$D(n)$}} \\ H\,R(n) & & \end{array}\right]$$

Step 5 : Up to this point we have

$$\underline{s}(n) \mid \underline{X}(n) \sim \mathcal{N}\left(\hat{\underline{s}}(n|n), M(n|n)\right)$$

$$\sim \underline{s}(n) \mid \hat{\underline{s}}(n|n-1), \underline{x}(n) \qquad \boxed{\text{Step 1.}}$$

and

$$\begin{bmatrix} \underline{s}(n) \\ \hat{\underline{s}}(n|n-1) \\ \underline{x}(n) \end{bmatrix} \sim \mathcal{N}\left(\underline{0}, \left[\begin{array}{c:cc} R(n) & \hat{M}(n|n-1) & R(n)H^T \\ \hdashline \hat{M}(n|n-1) & & \\ H\,R(n) & \multicolumn{2}{c}{D(n)} \end{array}\right] \right)$$

$\boxed{\text{Steps 2-4}}$

Therefore

$$\hat{\underline{s}}(n|n) = \begin{bmatrix} \hat{M}(n|n-1) & R(n)H^T \end{bmatrix} \cdot D(n)^{-1} \cdot \begin{bmatrix} \hat{\underline{s}}(n|n-1) \\ \underline{x}(n) \end{bmatrix}$$

$$M(n|n) = R(n) - \begin{bmatrix} \hat{M}(n|n-1) & R(n)H^T \end{bmatrix} D(n)^{-1} \begin{bmatrix} \hat{M}(n|n-1) \\ H\,R(n) \end{bmatrix}$$

Lets compute $D(n)^{-1}$!

Due to the nice structure of $D(n)$,

$$D(n)^{-1} = \begin{bmatrix} \hat{M}(n|n-1)^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -H^T \\ I \end{bmatrix} \left[H M(n|n-1) H^T + R_w(n) \right]^{-1} \begin{bmatrix} -H & I \end{bmatrix}$$

You may verify this yourself.

Plugging in, we get

$$\hat{s}(n|n) = \hat{s}(n|n-1) + K(n) \cdot \left(\underline{x}(n) - H \hat{\underline{s}}(n|n-1) \right)$$

$$M(n|n) = \left(I - K(n) H \right) M(n|n-1)$$

where

$$K(n) = M(n|n-1) \cdot H^T \cdot \left(H M(n|n-1) H^T + R_w(n) \right)^{-1}$$

is called the <u>Kalman gain</u>.

Step 6 Identify recursion.

We are almost there. All we lack are

$$\hat{\underline{s}}(n|n-1) = E\left[\underline{s}(n) \mid \underline{X}(n-1)\right]$$

$$= E\left[A \cdot \underline{s}(n-1) + B\,\underline{u}(n) \mid \underline{X}(n-1)\right]$$

$$= A\,E\left[\underline{s}(n-1) \mid \underline{X}(n-1)\right] + B\,E\left[\underline{u}(n) \mid \underline{X}(n-1)\right]$$

$$= A\,\hat{\underline{s}}(n-1|n-1) + \underline{0}.$$

and

$$M(n|n-1) = A\,M(n-1|n-1)\,A^T + B\,Q\,B^T$$

by a similar application of the definition of the vector Gauss-Markov process $\underline{s}(n)$

In particular

$$M(n|n-1) = E\left[\left(\underline{s}(n) - \hat{\underline{s}}(n|n-1)\right)\left(\underline{s}(n) - \hat{\underline{s}}(n|n-1)\right)^T \,\middle|\, \underline{X}(n-1)\right]$$

$$= E\left[\left(A\underline{s}(n-1) + B\underline{u}(n) - A\cdot\hat{\underline{s}}(n-1|n-1)\right)\cdot\right.$$

$$\left.\left(A\underline{s}(n-1) + B\underline{u}(n) - A\hat{\underline{s}}(n-1|n-1)\right)^T \,\middle|\, \underline{X}(n-1)\right]$$

$$= E\left\{\left(A\left[\underline{s}(n-1) - \hat{\underline{s}}(n-1|n-1)\right] + B\underline{u}(n)\right)\cdot\right.$$

$$\left.\left(A\left[\underline{s}(n-1) - \hat{\underline{s}}(n-1|n-1)\right] + B\underline{u}(n)\right)^T \,\middle|\, \underline{X}(n-1)\right\}$$

$$= A\cdot E\left[\left(\underline{s}(n-1) - \hat{\underline{s}}(n-1|n-1)\right)\cdot\left(\underline{s}(n-1) - \hat{\underline{s}}(n-1|n-1)\right)^T \,\middle|\, \underline{X}(n-1)\right]$$

$$+ B\cdot E\left[\underline{u}(n)\cdot\underline{u}(n)^T\right]B^T$$

independence assumption

$$= A\,M(n-1|n-1)\,A^T + B\,Q\,B^T$$

Kalman Filter Equations

1. $\hat{\underline{s}}(n|n-1) = A\,\hat{\underline{s}}(n-1|n-1)$

2. $M(n|n-1) = A \cdot M(n-1|n-1)\,A^T + B\,Q\,B^T$.

3. $K(n) = M(n|n-1)\,H^T \left(H\,M(n|n-1)\,H^T + R_w(n)\right)^{-1}$

4. $\hat{\underline{s}}(n|n) = \hat{\underline{s}}(n|n-1) + K(n) \cdot \left(\underline{x}(n) - H\,\hat{\underline{s}}(n|n-1)\right)$

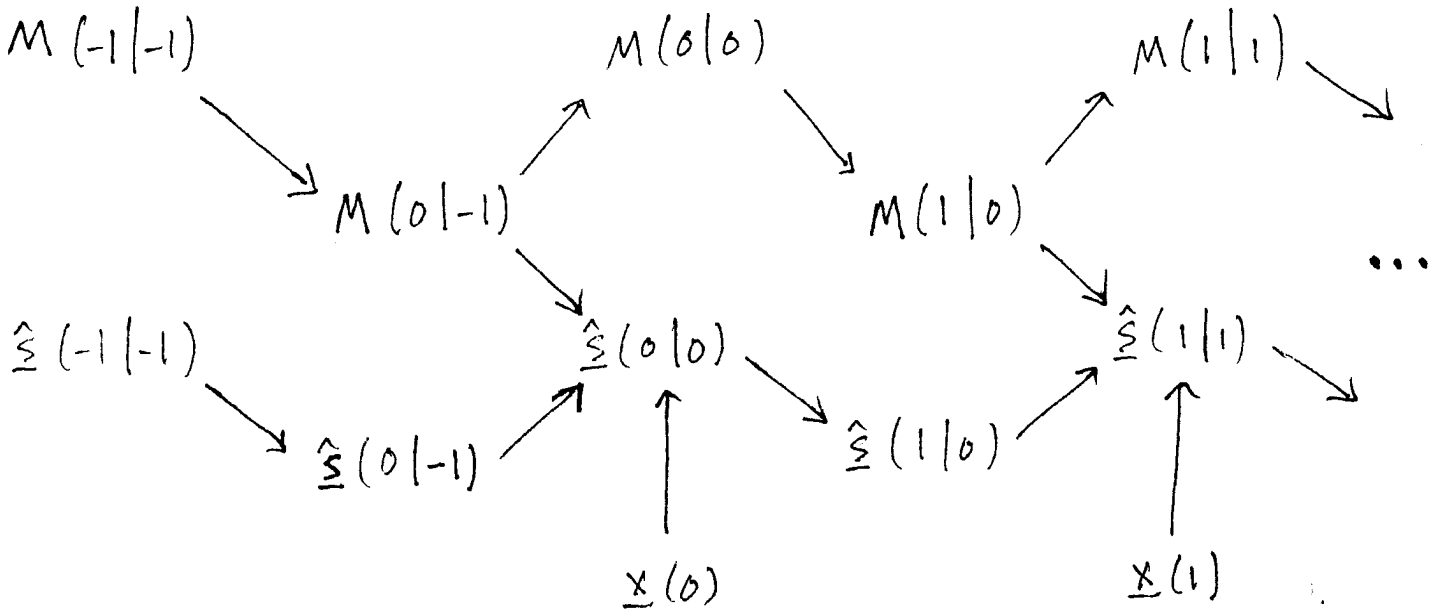
5. $M(n|n) = \left(I - K(n)\,H\right) M(n|n-1)$

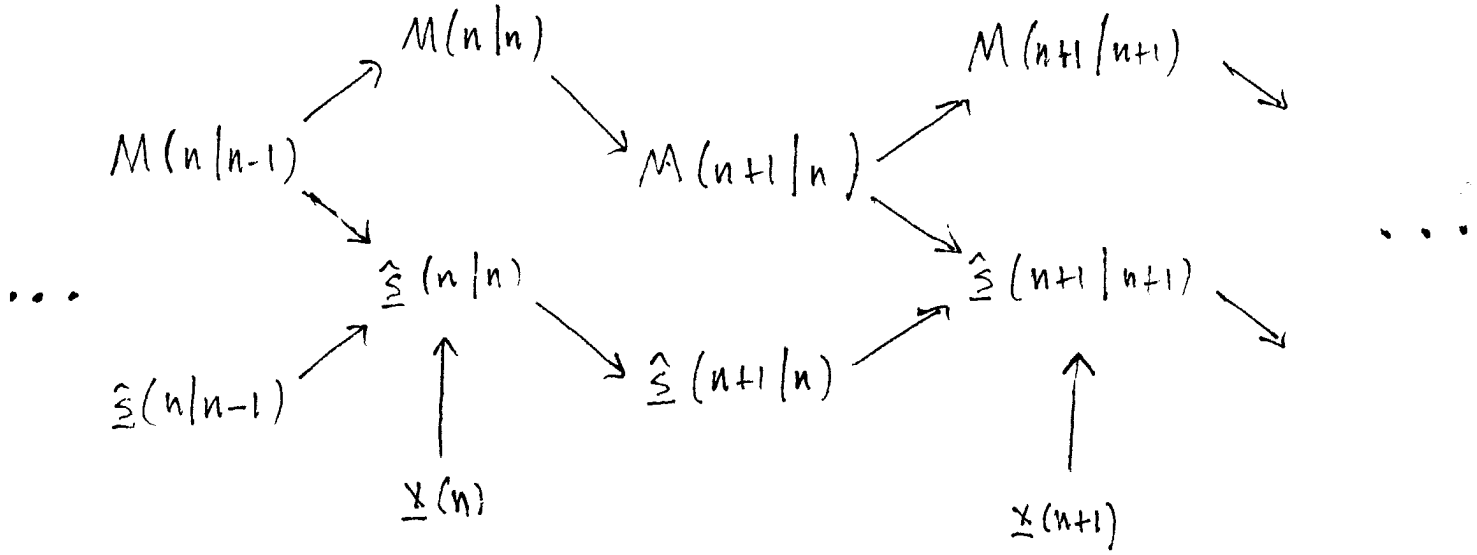
Initialization

$\hat{\underline{s}}(-1|-1) = \underline{\mu}_0$

$\hat{M}(-1|-1) = \Sigma_0$

Flow Diagram





Two stages:

I. **Update** predictor variables $\hat{\underline{s}}(n|n-1)$, $M(n|n-1)$

II. **Update** estimator variables $\hat{\underline{s}}(n|n)$, $M(n|n)$

Remarks

- The Kalman filter is the MMSE estimator when the data / state are jointly Gaussian.
- If not, the Kalman filter is still the LMMSE estimator.
- In practice, the state transition model or observation model may be innaccurate. The Kalman filter may still return a useful estimator.
- The Kalman recursions are still valid when we allow A, B, H, and Q to be time varying.

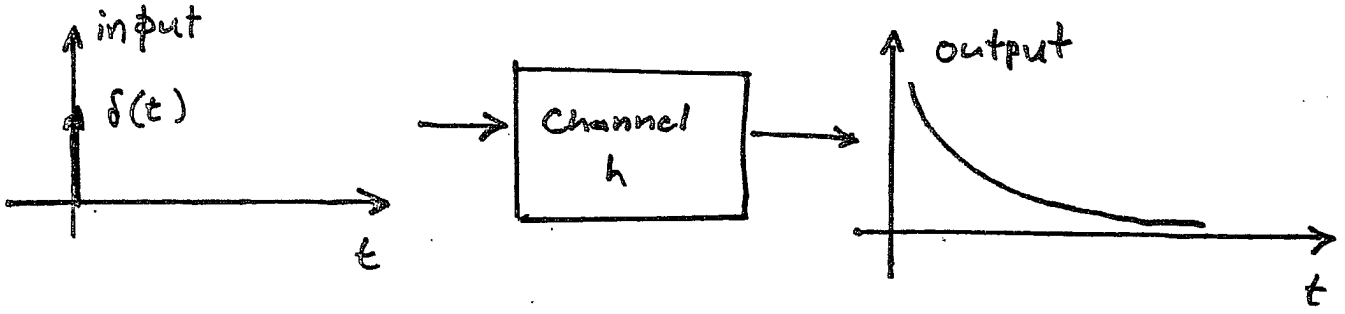
Key

a. $H = \begin{bmatrix} v(n-p+1) & \cdots & v(n) \end{bmatrix}$, $\underline{s}(n) = \begin{bmatrix} g_n(0) \\ g_n(1) \\ \vdots \\ g_n(p-1) \end{bmatrix}$

APPLICATION: CHANNEL ESTIMATION

Kalman Filtering for Time-Varying Channel Estimation

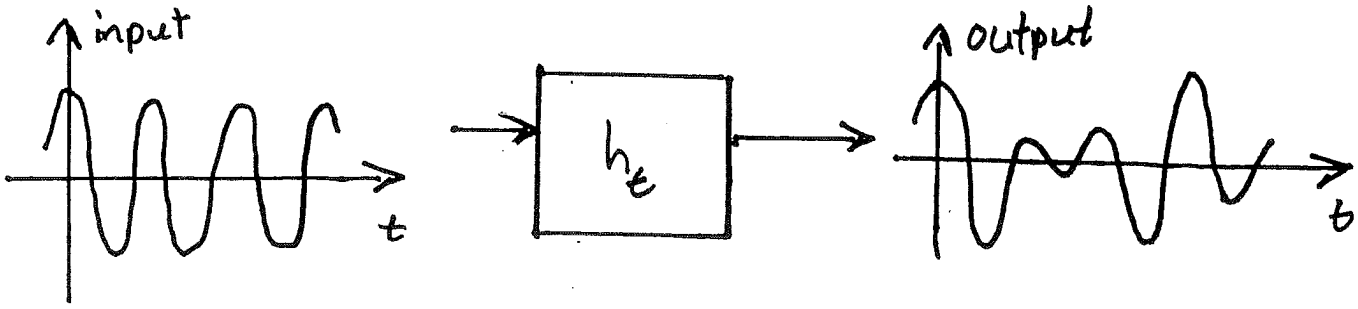
Multipath Communications Channel:



This effect is the result of many propagation paths, each of which delays and attenuates the input signal.

Additionally, the channel is time-varying due to movement of the source, receiver, and/or scatters. Therefore, the channel is acting like a linear time varying filter

Due to the time-varying nature of the channel, a sinusoidal input does not produce a pure sinusoid at the output:



Instead, the output is a narrowband process. Assuming the channel is relatively slowly varying (compare to the frequency of the input) we can view the input sinusoid as being amplitude modulated by the time-varying channel. This effect is referred to as <u>fading</u> and such channels are called <u>fading multipath</u> channels.

If we sample the output of the channel, then a very good model is the low-pass tapped delay line model:

$$y(n) = \sum_{k=0}^{p-1} h_n(k)\, v(n-k)$$

$y(n)$: output

$v(n-k)$: input

$h_n(k)$: impulse response depends on time n

This is simply an FIR filter with time-varying coefficients. In practice we wouldn't observe this perfect output, but rather a noise-corrupted version of it:

$$x(n) = \sum_{k=0}^{p-1} h_n(k)\, v(n-k) + w(n)$$

where $w(n)$ is observation noise.

The goal of channel estimation is to determine the linear time-varying filter $h_n(k)$ based on the input $v(n)$ and measured output $x(n)$. Is this possible?

Assume $v(n) = 0$ for $n < 0$. Then

$$x(0) = h_0(0)\,v(0) + h_0(1)\,v(-1) + w(0)$$
$$= h_0(0)\,v(0) + w(0)$$

$$x(1) = h_1(0)\,v(1) + h_1(1)\,v(0) + w(1)$$

$$x(2) = h_2(0)\,v(2) + h_2(1)\,v(1) + w(2)$$

$$\vdots$$

For each $n \geq 1$ we have two new parameters we must estimate!

Even in the absence of measurement noise we have more unknowns than equations and we can't determine the filter.

What can we do?

Well, suppose that the filter weights are not changing too rapidly from sample to sample. This is known as a slow-fading channel model.

Probabilistically, we can view the slowly varying channel as vector-valued Gauss-Markov process:

$$\underline{h}_{n+1} = A\,\underline{h}_n + \underline{u}_n$$

where $\underline{h}_n = [h_n(0), \ldots, h_n(p-1)]^T$, A is a $p \times p$ matrix designed to reflect the correlation expected between filter weights at different time samples, and $\underline{u}_n$ is a white Gaussian noise vector process with covariance Q.

That is,

$\ldots, \underline{u}_{n-1}, \underline{u}_n, \underline{u}_{n+1}, \ldots$ are iid vectors

and

$$\underline{u}_n \sim N(\underline{0}, Q)$$

A standard simplifying assumption is to assume that A and Q are diagonal $\Rightarrow$ filter weights are uncorrelated with each other.

This is called an <u>uncorrelated scattering</u> model.

The measurement/observation model in vector form is

$$x_n = \underline{v}_n^T \underline{h}_n + w_n$$

where $\underline{v}_n = [v(n), v(n-1), \ldots, v(n-p+1)]^T$.

With this notation and our Gauss-Markov model for the time-varying filter, we can now devise a Kalman filter to estimate and <u>track</u> the channel.

In the case we have the

State equation:

$$\underline{h}_{n+1} = A\,\underline{h}_n + \underline{u}_n \,, \quad n \geq 0$$

(with $\underline{h}_n$ in place of $\underline{s}_n$ now)

Furthermore, assume that

$$\underline{h}_0 \sim N(\underline{0}, R_0)$$

with R_0 also diagonal.

Measurement equation:

$$x_n = \underline{v}_n^T \underline{h}_n + w_n$$

Note that $\underline{v}_n$ is known, but time-varying. In our earlier discussion the H matrix of the observation model was constant. The Kalman filter still is applicable here, we just replace H with $\underline{v}_n^T$.

Kalman Filter

$$\left(\underline{\Sigma}_n = \underline{h}_n, \quad H = \underline{v}_n^T, \quad B = I\right)$$

$$\hat{\underline{h}}_{n|n-1} = A\,\hat{\underline{h}}_{n-1|n-1}$$

$$M_{n|n-1} = A\,M_{n-1|n-1}\,A^T + Q$$

$$K_n = \frac{M_{n|n-1}\cdot \underline{v}_n}{\underline{v}_n^T M_{n|n-1}\,\underline{v}_n + \sigma_w^2}$$

$$\hat{\underline{h}}_{n|n} = \hat{\underline{h}}_{n|n-1} + K_n\left(x_n - \underline{v}_n^T\,\hat{\underline{h}}_{n|n-1}\right)$$

$$M_{n|n} = \left(I - K_n\,\underline{v}_n^T\right) M_{n|n-1}$$

Example ($p = 2$)

Slow-Fading Model Parameters

$$A = \begin{bmatrix} 0.999 & 0 \\ 0 & 0.999 \end{bmatrix}$$ ← state transition matrix

$$Q = \begin{bmatrix} 0.0001 & 0 \\ 0 & 0.0001 \end{bmatrix}$$ ← driving noise covariance

to reflect little or no knowledge about the initial state of the channel

$$\underline{h}_0 \sim N\left(\underline{0}, \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}\right)$$

Channel model:

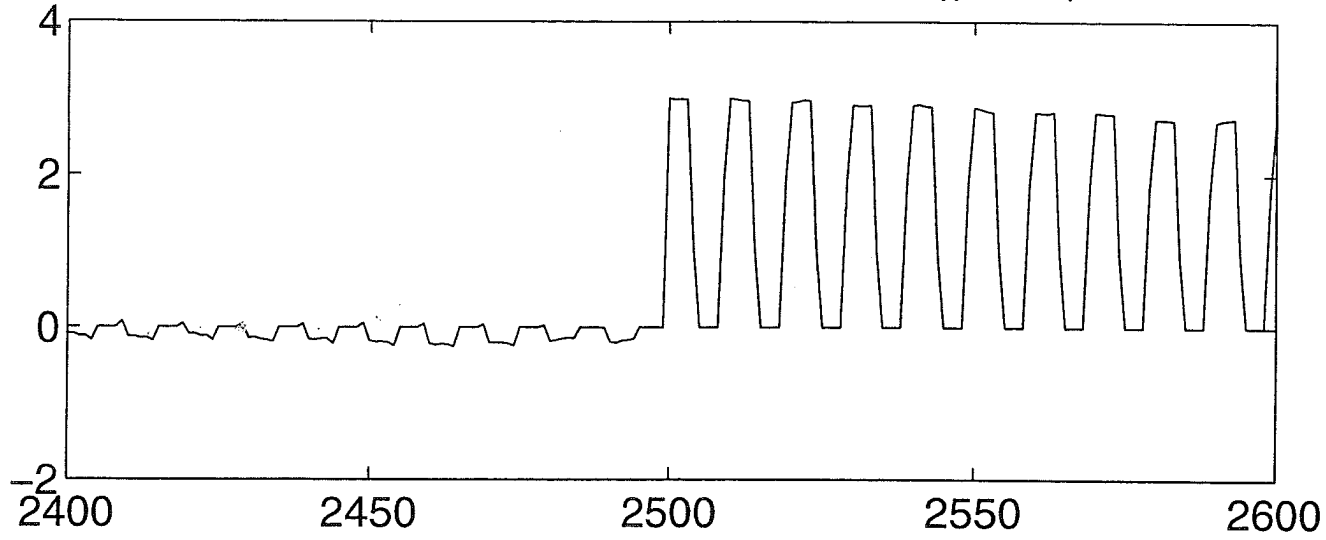
$$x(n) = h_n(0)\, v(n) + h_n(1)\, v(n-1) + w(n)$$

$v(n)$ = input to channel, known square wave

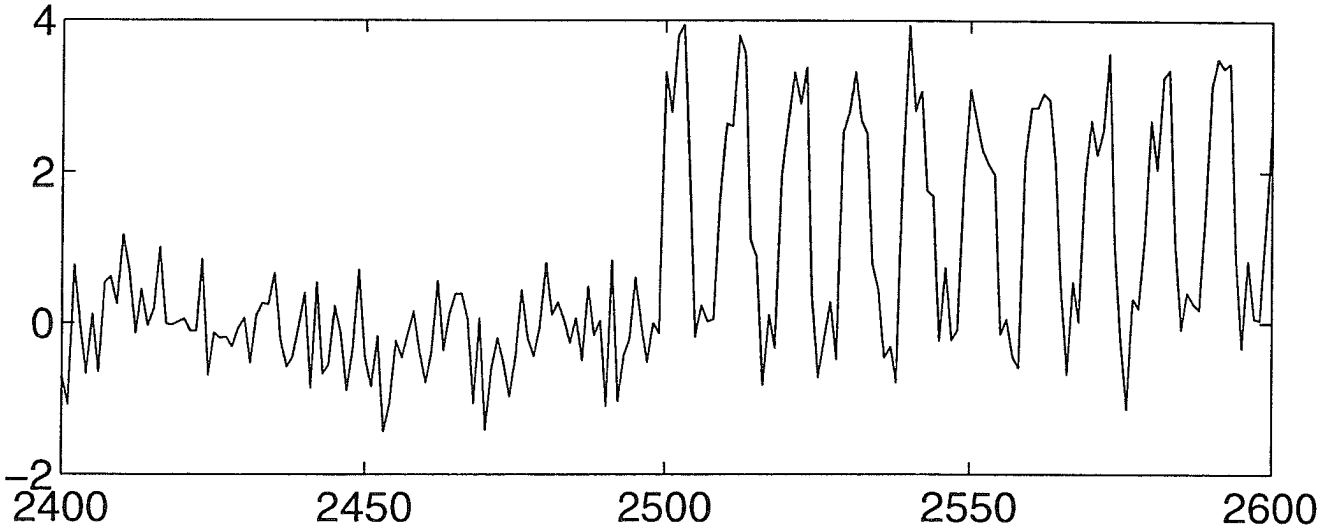
$h_n(k)$ = time-varying channel model (linear time-varying FIR filter)

$w(n)$ = white Gaussian observation noise

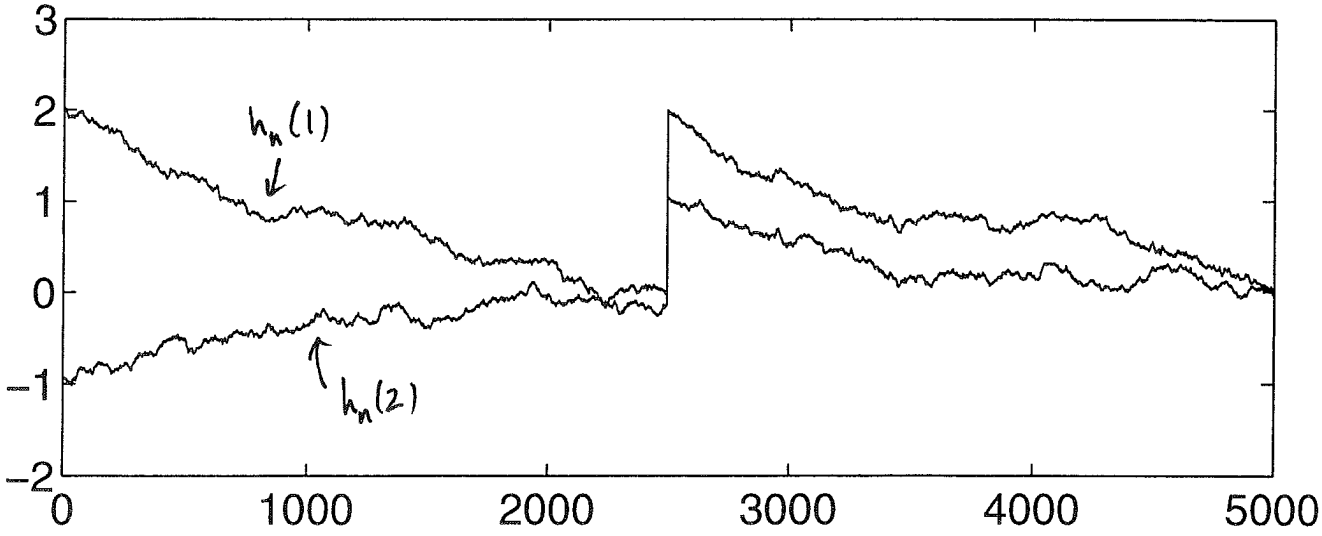
noise–free channel output $y_n = \underline{v}_n^T \underline{h}_n$



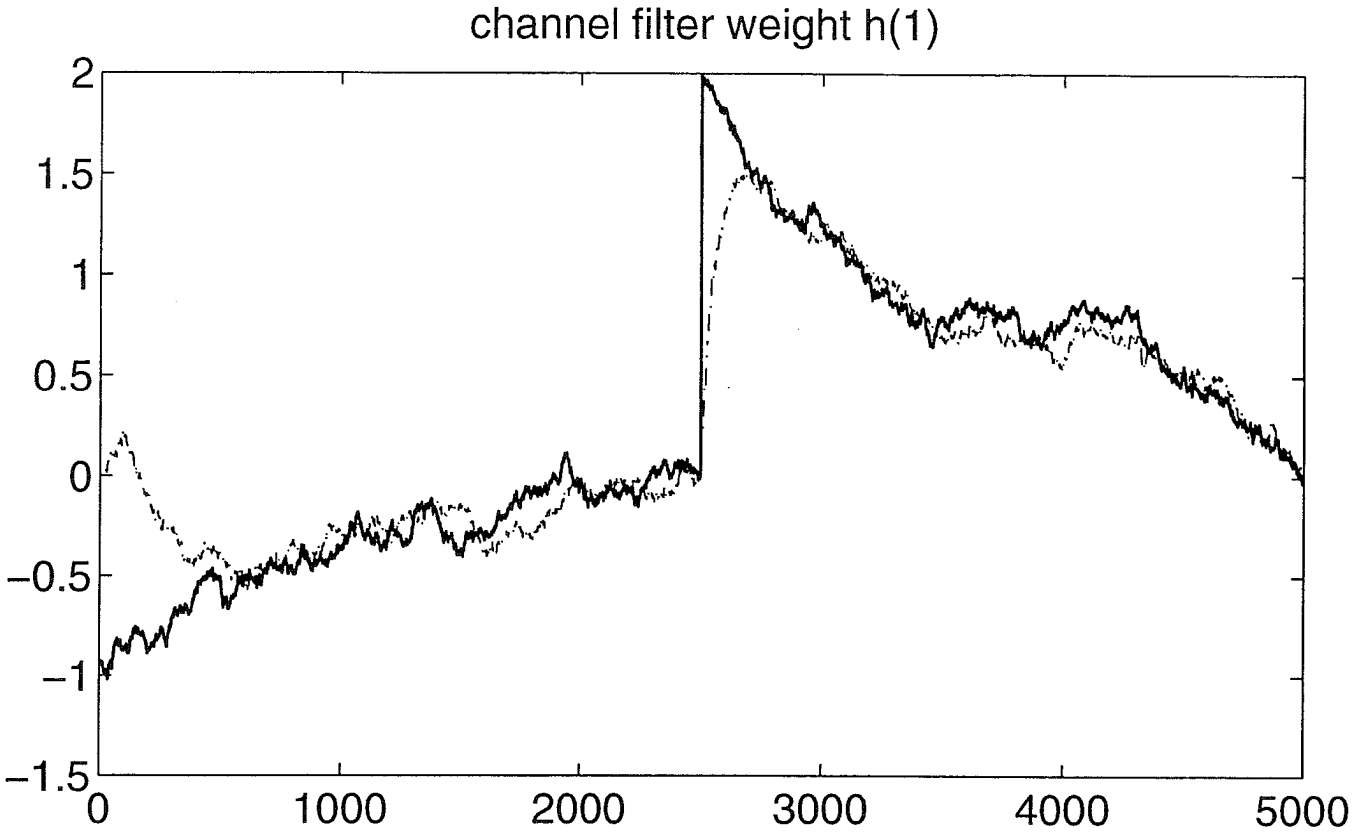
observation channel output x

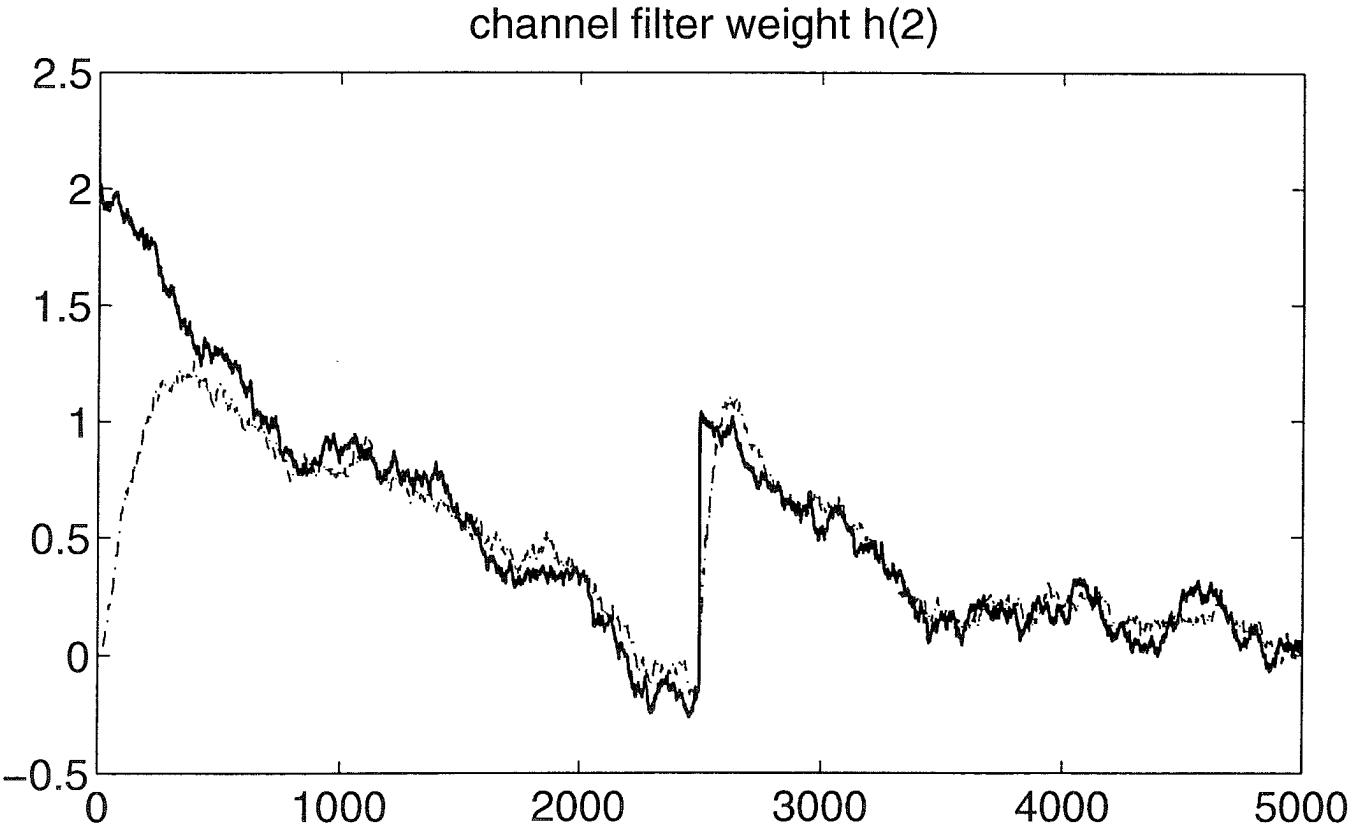


channel h



↑ filter changes drastically
(e.g., car drives out of a tunnel)


channel filter weight h(1)
2
1.5
1
0.5
0
−0.5
−1
−1.5
0
1000
2000
3000
4000
5000


channel filter weight h(2)
2.5
2
1.5
1
0.5
0
−0.5
0
1000
2000
3000
4000
5000

DETECTION THEORY

Hypothesis Testing

In a hypothesis testing problem, we are given data $\underline{x} = [x_1 \dots x_N]^T$, and we must decide which of two or more hypotheses best fits the data.

Example

Suppose $x_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$

but we are not sure about θ.

we suspect either $\theta = -1$ or $\theta = 1$:

$\mathcal{H}_0: \quad \theta = -1$

$\mathcal{H}_1: \quad \theta = 1$

Binary vs. M-ary Test

In general, there can be $M \geq 2$ hypotheses, $\mathcal{H}_1, \ldots, \mathcal{H}_M$.

Example

Suppose we want to devise a speech recognition system to recognize the digits $0, 1, \ldots, 9$:

$$\mathcal{H}_0 : \underline{x} \sim \text{"zero"}$$
$$\mathcal{H}_1 : \underline{x} \sim \text{"one"}$$
$$\vdots$$
$$\mathcal{H}_9 : \underline{x} \sim \text{"nine"}$$

If $M=2$, we have a <u>binary</u> testing problem.

Example

$$\mathcal{H}_0 : \underline{x} \sim \text{"no"}$$
$$\mathcal{H}_1 : \underline{x} \sim \text{"yes"}$$

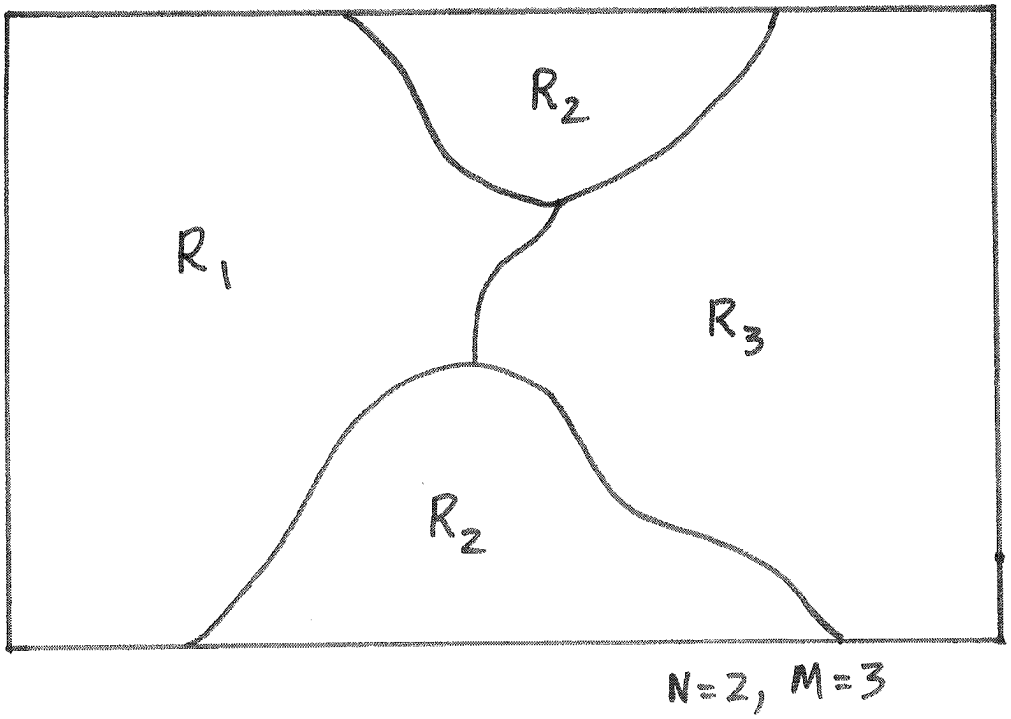
In hypothesis testing the goal is to construct a **decision rule** / detector / test, i.e. a mapping

$$h: \mathbb{R}^N \longrightarrow \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$$

assigning an observation $\underline{x}$ to a hypothesis.

A decision rule partitions the input space into **decision regions**

$$R_k = \{\underline{x} \in \mathbb{R}^N : h(\underline{x}) = \mathcal{H}_k\}.$$



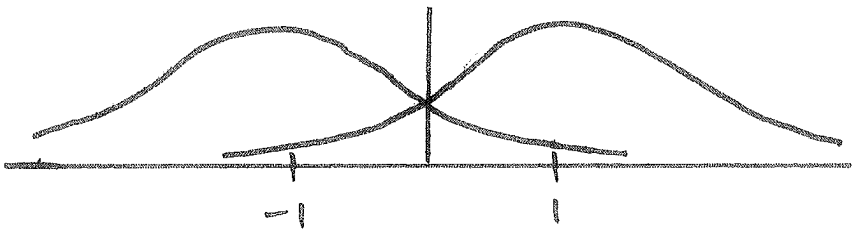
$N=2,\ M=3$

Ex Suppose $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, where

$$x_n \overset{iid}{\sim} N(\theta, 1)$$

and consider the testing problem

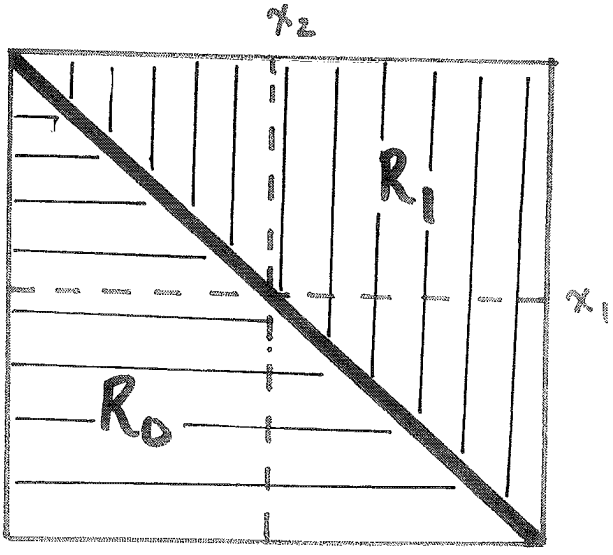
$$H_0: \theta = -1$$
$$H_1: \theta = 1$$



A reasonable test is

$$h(\underline{x}) = \begin{cases} H_0 & \text{if } \bar{x} \leq 0 \\ H_1 & \text{if } \bar{x} > 0 \end{cases}$$

where $\bar{x} = \frac{1}{2}(x_1 + x_2)$



Simple vs. Composite Hypotheses

If a hypothesis $\mathcal{H}$ specifies a unique distribution for $\underline{x}$, we say $\mathcal{H}$ is **simple**.

If $\mathcal{H}$ specifies a class of possible distributions for $\underline{x}$, we say $\mathcal{H}$ is **composite**

Example

Suppose we want to detect a sinusoid

$$s_n = \cos(2\pi f_0 n + \phi), \quad n = 0, 1, \ldots, N-1$$

where f_0 is known but ϕ is unknown.

Let $w_n \sim \mathcal{N}(0, \sigma^2)$, σ^2 known.

$$\mathcal{H}_0 : \underline{x} = \underline{w}$$

$$\mathcal{H}_1 : \underline{x} = \underline{s} + \underline{w}$$

Here $\mathcal{H}_0$ is simple and

$\mathcal{H}_1$ is composite (since ϕ unknown)

Terminology

Null hypothesis: An interesting event did not happen

One-Sided Test

$$\mathcal{H}_0: \quad \theta \leq \theta_0 \qquad\qquad \theta \in \mathbb{R}$$
$$\mathcal{H}_1: \quad \theta > \theta_0$$

Two-Sided Test

$$\mathcal{H}_0: \quad \underline{\theta} = \underline{\theta}_0$$
$$\mathcal{H}_1: \quad \underline{\theta} \neq \underline{\theta}_0 \qquad\qquad \underline{\theta} \in \mathbb{R}^N$$

"hypothesis $\mathcal{H}_1$ lies on both sides of $\mathcal{H}_0$"

Thresholding Rule

$$T(\underline{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma$$

$T(\underline{x})$ is a scalar statistic.

Detection Theory

Detection theory is another name for hypothesis testing in the context of signal processing problems.

In the next few weeks, we will study the basic theory of hypothesis testing, and apply it to various signal detection problems.

BAYES RISK DETECTION

Consider a binary hypothesis testing problem involving simple hypotheses:

$$H_0: \underline{X} \sim f_0(\underline{x})$$

$$H_1: \underline{X} \sim f_1(\underline{x})$$

We assume for every observation $\underline{x}$, exactly one of the two models is true.

Let's view the "active hypothesis" as a random event:

$\pi_0 :=$ probability that H_0 is in effect

$\pi_1 :=$ " " H_1 " " "

Obviously we have

$$\pi_0 + \pi_1 = 1.$$

The Bayes Risk

How should we measure the performance of a decision rule?

Suppose we have a decision rule defined by the <u>decision regions</u> R_0 and R_1.

$$\underline{x} \in R_0 \iff \text{declare } H_0 \text{ is in effect}$$

$$\underline{x} \in R_1 \iff \text{declare } H_1 \text{ is in effect}$$

There are four possible outcomes:

decision \ truth	H_0	H_1
$\underline{x} \in R_0$	(0,0)	(0,1)
$\underline{x} \in R_1$	(1,0)	(1,1)

Suppose we are able to specify

$$c_{i,j} := \underline{\text{cost}} \text{ of declaring } H_i \text{ when } H_j \text{ true}$$

To be sensible, we should have

$$c_{i,i} < c_{i,j}, \quad i \neq j$$

Define the Bayes Risk

$$\bar{C} = \text{expected cost of a decision}$$

$$= \sum_{i,j=0}^{1} c_{ij} \cdot P(\text{declare } H_i, H_j \text{ true})$$

$$= \sum_{i,j=0}^{1} c_{ij} \cdot P(H_j \text{ true}) \cdot P(\text{declare } H_i \mid H_j \text{ true})$$

$$= \sum_{i,j=0}^{1} c_{ij}\, \pi_j\, P(H_i \mid H_j)$$

where

$$P(H_i \mid H_j) := \text{probability that } \underline{X} \in R_i \text{ when } \underline{X} \sim f_j(\underline{x})$$

Remark

It is helpful to think of (declared hypothesis, true hypothesis) as a jointly distributed random pair.

Example | Consider a scalar observation

$$H_0 : X \sim \mathcal{N}(-1, 1)$$

$$H_1 : X \sim \mathcal{N}(1, 1)$$

If our decision regions are

$$R_0 = (-\infty, 0]$$

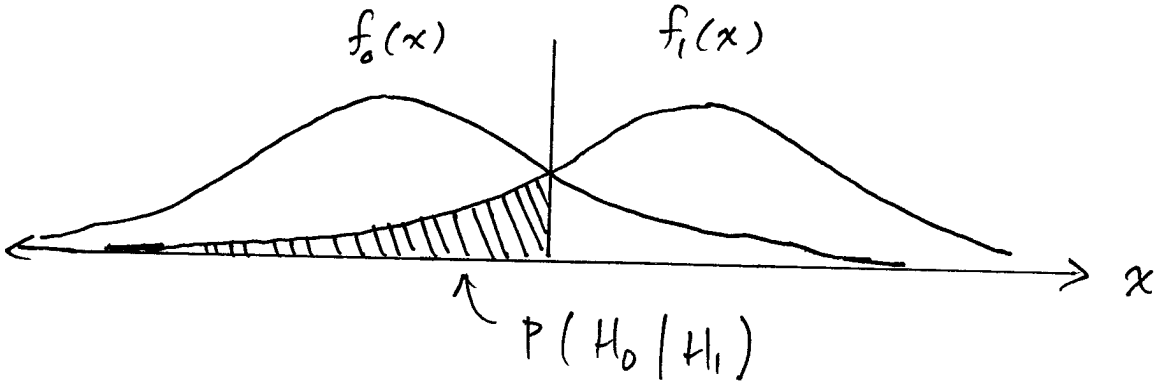
$$R_1 = (0, \infty)$$

then

$$P(H_0 \mid H_1) = P(X \in R_0 \mid H_1)$$

(a) $=$

Picture:



The decision rule minimizing the Bayes risk is called the Bayes risk detector

Bayes Risk Detector: Continuous Case

Assume $\underline{X}$ is continuous under both H_0 and H_1. That is, assume $f_0(\underline{x})$ and $f_1(\underline{x})$ are densities.

We may write

$$\bar{C} = \sum_{i,j=0}^{1} c_{ij}\, \pi_j\, P(H_i \mid H_j)$$

$$= \sum_{i,j} c_{ij}\, \pi_j \int_{R_i} f_j(\underline{x})\, d\underline{x}$$

$$= \int_{R_0} \left(c_{00}\, \pi_0\, f_0(\underline{x}) + c_{01}\, \pi_1\, f_1(\underline{x}) \right) d\underline{x}$$

$$+ \int_{R_1} \left(c_{10}\, \pi_0\, f_0(\underline{x}) + c_{11}\, \pi_1\, f_1(\underline{x}) \right) d\underline{x}$$

How should we choose R_0 and R_1 to minimize this expression?

Recall:

$$\left.\begin{aligned} R_0 \cap R_1 &= \emptyset \\ R_0 \cup R_1 &= \mathbb{R}^N \end{aligned}\right\} \text{Partition of } \mathbb{R}^N$$

So every $\underline{x} \in \mathbb{R}^N$ is in one and only one R_i.

To minimize the Bayes risk, therefore, choose $\underline{x} \in R_i$ when the corresponding integrand is smaller.

That is, choose

$$\underline{x} \in R_0 \iff c_{00}\pi_0 f_0(\underline{x}) + c_{01}\pi_1 f_1(\underline{x}) < c_{10}\pi_0 f_0(\underline{x}) + c_{11}\pi_1 f_1(\underline{x})$$

$$\iff \frac{f_1(\underline{x})}{f_0(\underline{x})} < \frac{\pi_0}{\pi_1} \cdot \frac{(c_{10} - c_{00})}{(c_{01} - c_{11})}$$

More concisely, we may express the Bayes risk detector as:

$$\boxed{\frac{f_1(\underline{x})}{f_0(\underline{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\pi_0}{\pi_1} \cdot \frac{(c_{10} - c_{00})}{(c_{01} - c_{11})}}$$

Likelihood Ratio Tests

The Bayes risk detector is an example of a likelihood ratio test (LRT)

A LRT has the form

$$\Lambda(\underline{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

where

$$\Lambda(\underline{x}) := \frac{f_1(\underline{x})}{f_0(\underline{x})}$$

is the likelihood ratio and $\eta > 0$ is a threshold.

Bayes Risk Detector: Discrete Case

Now suppose f_0 and f_1 are mass functions, and let $\mathcal{X}$ denote the domain of $\underline{X}$.

Then

$$\bar{c} = \sum_{i,j} c_{ij}\, \pi_j\, P(H_i \mid H_j)$$

(b)
$$= \sum_{i,j} c_{ij}\, \pi_j$$

$$= \sum_{\underline{x} \in \mathcal{X} \cap R_0} \left(c_{00}\, \pi_0\, f_0(\underline{x}) + c_{01}\, \pi_1\, f_1(\underline{x}) \right)$$

$$+ \sum_{\underline{x} \in \mathcal{X} \cap R_1} \left(c_{10}\, \pi_0\, f_0(\underline{x}) + c_{11}\, \pi_1\, f_1(\underline{x}) \right)$$

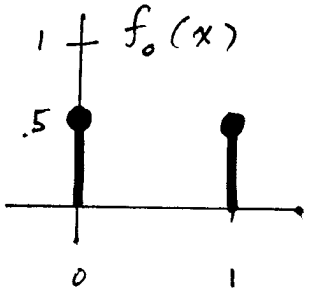
Choosing R_0, R_1 to minimize this expression we once again obtain

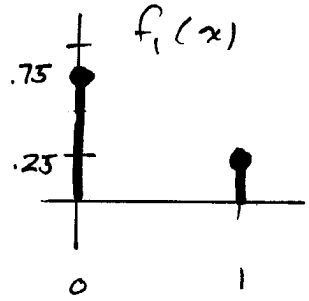
$$\frac{f_1(\underline{x})}{f_0(\underline{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\pi_0}{\pi_1} \cdot \frac{(c_{10} - c_{00})}{(c_{01} - c_{11})}$$

Exercise (a) Give an example of a discrete problem where $\Lambda(\underline{X}) = \eta$ occurs with probability > 0. (b) Same as (a), but for a continuous problem (c) What is the optimal decision in such cases?

Solution (a) Essentially any discrete problem will suffice provided η is chosen appropriately.

For example, suppose $\mathcal{X} = \{0, 1\}$





$\pi_1 = \frac{2}{5}$, $\pi_0 = \frac{3}{5}$, and $c_{ij} = 1 - \delta_{ij}$

Then

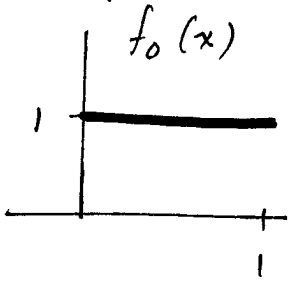
$$\frac{f_1(0)}{f_0(0)} = \frac{.75}{.5} = \frac{3}{2} =: \eta$$

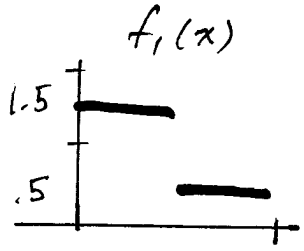
occurs with probability

$$\frac{3}{5} \cdot (.5) + \frac{2}{5} \cdot (.75) > 0$$

(b) Taking $f_1(x)$ and $f_0(x)$ to be piecewise constant does the trick, although there are other ways.

For example





Then $\forall\, x \in (0, \frac{1}{2})$ we have

$$\frac{f_1(x)}{f_0(x)} = \frac{1.5}{1} = \frac{3}{2}$$

Now select $\pi_1 = \frac{2}{5}$, $\pi_0 = \frac{3}{5}$, $c_{ij} = 1 - \delta_{ij}$.

Then $\eta = \frac{3}{2}$, so

$$\frac{f_1(x)}{f_0(x)} = \eta$$

occurs with probability

$$\frac{2}{5} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{3}{4} > 0 .$$

(c) It doesn't matter what you decide. Either decision contributes equally to the Bayes risk.

Minimum Probability of Error Detector

An important special case of the Bayes detector occurs when

$$c_{ij} = 1 - \delta_{ij} = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

Then the Bayes risk is

$$\bar{C} = \sum_{ij} c_{ij} P(\text{declare } H_i, H_j \text{ true})$$

$$= P(\text{declare } H_0, H_1 \text{ true})$$

$$+ P(\text{declare } H_1, H_0 \text{ true})$$

$$= P(\text{decision} \neq \text{truth})$$

$$= \underline{\text{probability of error}} =: P_E$$

The "min P_E" detector is therefore

Example Consider the problem of detecting a DC signal with amplitude $A > 0$ in additive white Gaussian noise.

$$H_0: \; X_i = W_i, \qquad i = 1, \ldots, N$$

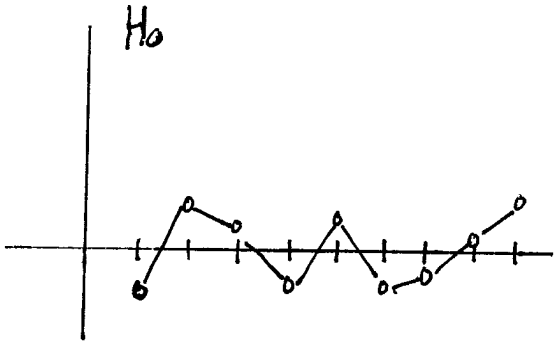
$$H_1: \; X_i = A + W_i, \qquad i = 1, \ldots, N$$

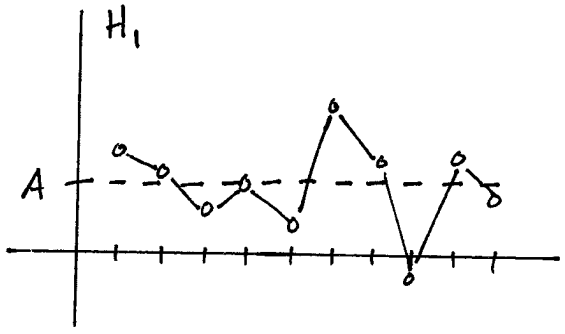
where $W_i \overset{iid}{\sim} N(0, \sigma^2)$ and A, σ^2 are known.

Equivalently we could write

$$H_0: \; \underline{X} \sim N(\underline{0}, \sigma^2 I)$$

$$H_1: \; \underline{X} \sim N(A \cdot \underline{1}, \sigma^2 I)$$





What is the Bayes risk detector? Any guesses?

We have

$$\Lambda(\underline{x}) = \frac{f_1(\underline{x})}{f_0(\underline{x})} = \frac{\prod_{n=1}^{N} f_1(x_n)}{\prod_{n=1}^{N} f_0(x_n)}$$

$$= \frac{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x_n - A)^2}{2\sigma^2}\right\}}{\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x_n^2}{2\sigma^2}\right\}}$$

$$= \frac{\exp\left\{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} (x_n - A)^2\right\}}{\exp\left\{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} x_n^2\right\}}$$

$$= \exp\left\{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} (-2x_n A + A^2)\right\}$$

$$= \exp\left\{\frac{A}{\sigma^2} \sum_{n=1}^{N} x_n - \frac{NA^2}{2\sigma^2}\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Can we simplify the detector further?

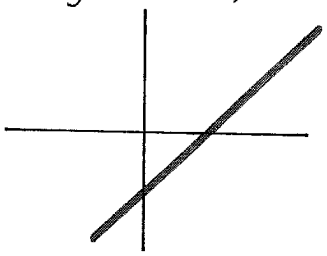
Monotonic Transformations

If we apply a monotonically increasing function to both sides of the LRT, the decision regions remain the same.

Examples

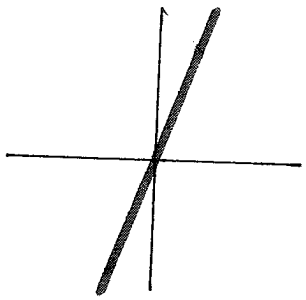
1. adding a number

$y = x + c, \; c \in \mathbb{R}$



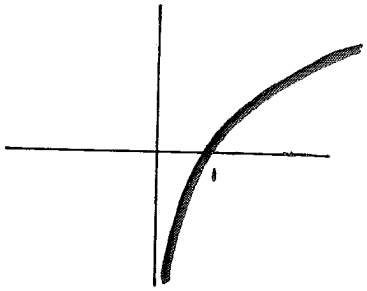
2. multiplying by a positive number

$y = cx, \; c > 0$



3. natural logarithm

$y = \ln(x)$



Commonly used, since many densities and mass functions have an exponential form

(DC signal detection, continued)

$$\exp\left\{ \frac{A}{\sigma^2} \sum_{n=1}^{N} x_n - \frac{NA^2}{2\sigma^2} \right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

$$\Updownarrow$$

$$\frac{A}{\sigma^2} \sum_{n=1}^{N} x_n - \frac{NA^2}{2\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \ln(\eta)$$

$$\Updownarrow$$

$$\frac{1}{N} \sum_{n=1}^{N} x_n \underset{H_0}{\overset{H}{\gtrless}} \frac{\sigma^2}{NA} \ln(\eta) + \frac{A}{2} \equiv \gamma$$

Thus, the detector reduces to a simple <u>thresholding test</u> involving the sample mean.

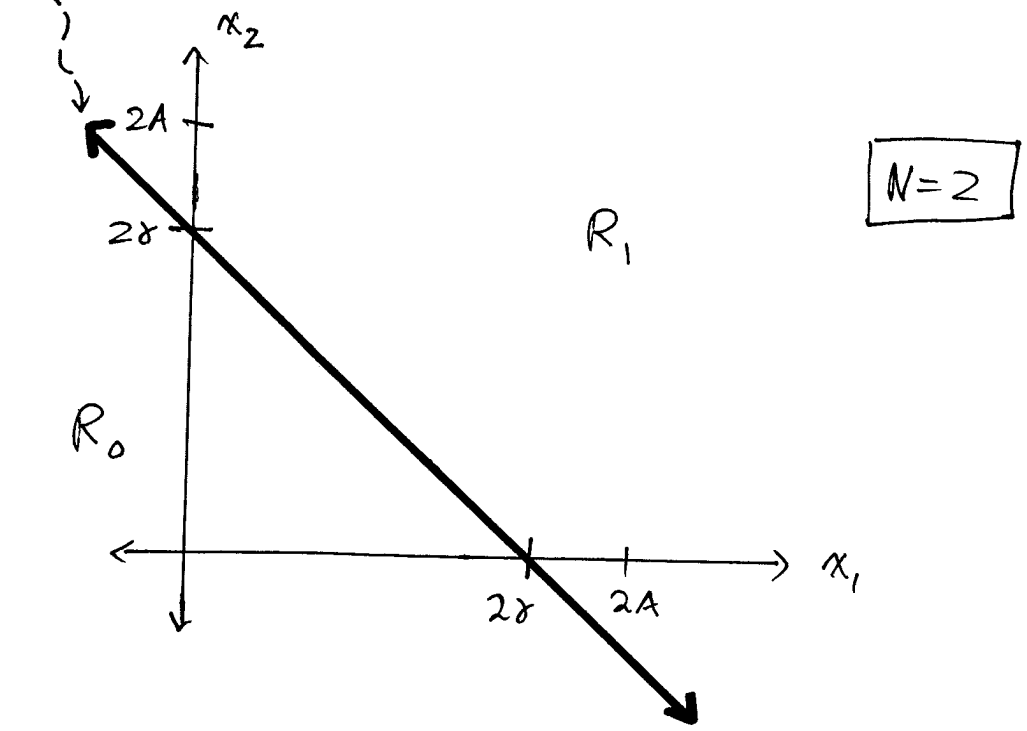
<u>Note</u> | If $\eta = 1$ $(\pi_0 = \pi_1 = \frac{1}{2})$, then $\gamma = \frac{A}{2}$, and σ^2 need not be known.

Where did we use the assumption $A > 0$?

Note that the Bayes risk detector is a linear detector, meaning it is obtained by thresholding a linear function of the data.

Equivalently, the decision boundary is a hyperplane

$$\text{decision boundary} = \left\{ \underline{x} \in \mathbb{R}^N : \langle \underline{x}, \tfrac{1}{N}\underline{1} \rangle = \gamma \right\}$$



Calculating Error Probabilities

How can we calculate $P(H_i / H_j)$ to assess the performance of a detector?

For example, the probability of error is

$$P_E = P(H_0, H_1) + P(H_1, H_0)$$

$$= \pi_1 P(H_0 | H_1) + \pi_0 P(H_1 | H_0)$$

$$= \pi_1 \int_{R_0} f_1(\underline{x})\, d\underline{x} + \pi_0 \int_{R_1} f_0(\underline{x})\, d\underline{x}$$

We must compute N-dimensional integrals.

This is a daunting task, even for the relatively simple case of Gaussian noise and linear decision boundaries.

Fortunately, we can use

- simplified test statistics
- monotone transformations

to make our lives easier.

In the previous example

$$t = \frac{1}{N} \sum_{i=1}^{N} x_i$$

is an example of a <u>test statistic</u>.

More generally, a test statistic is simply a statistic (i.e., a function of the data) that is used in a test/detector.

The importance of test statistics for error calculation is that often they

- are 1-dimensiond
- have known distributions

and can therefore be used in place of the N-dimensional data.

Example (continued)

We had

$$\frac{1}{N}\sum_{n=1}^{N} x_n = t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma = \frac{\sigma^2}{NA}\ln(\eta) + \frac{A}{2}$$

Recall

$$\underline{x} \sim N(\underline{0}, \sigma^2 I) \quad \text{under } H_0$$

$$\underline{x} \sim N(A\underline{1}, \sigma^2 I) \quad \text{under } H_1$$

Now

$$t = \underline{B}\underline{x}, \quad \underline{B} = \left[\frac{1}{N} \cdots \frac{1}{N}\right]$$

(d) So

$$T \sim$$

$$\sim \quad \text{under } H_0$$

and

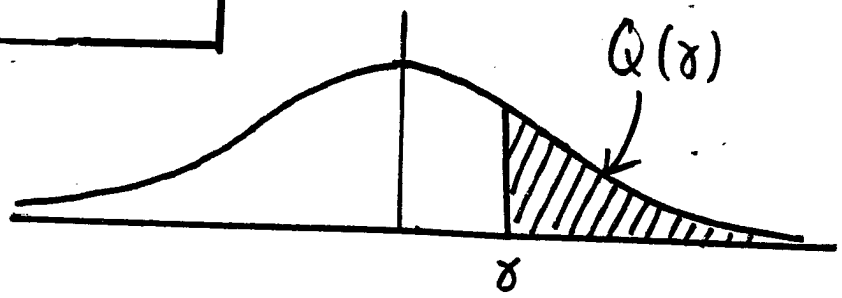
$$T \sim$$

$$\sim \quad \text{under } H_1$$

The Q-function

Let $X \sim N(0,1)$. Define

$$Q(\gamma) \equiv P(X \geq \gamma) = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$



If $X \sim N(\mu, \sigma^2)$, then

$$P(X \geq \gamma) = Q\left(\frac{\gamma - \mu}{\sigma}\right)$$

← show this by change of variables argument

Note: $Q : \mathbb{R} \to (0,1)$ is monotonically decreasing, so it has an inverse.

In Matlab

$$Q(\gamma) = \frac{1}{2}\left(1 - \text{erf}\left(\frac{\gamma}{\sqrt{2}}\right)\right)$$

$$Q^{-1}(\alpha) = \sqrt{2}\ \text{erfinv}(1-2\alpha)$$

Under H_0, $T \sim N(0, \frac{\sigma^2}{N})$, so

so

$$P(H_1|H_0) = P(T > \gamma \mid H_0)$$

$$= Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right)$$

Under H_1, $T \sim N(A, \frac{\sigma^2}{N})$, so

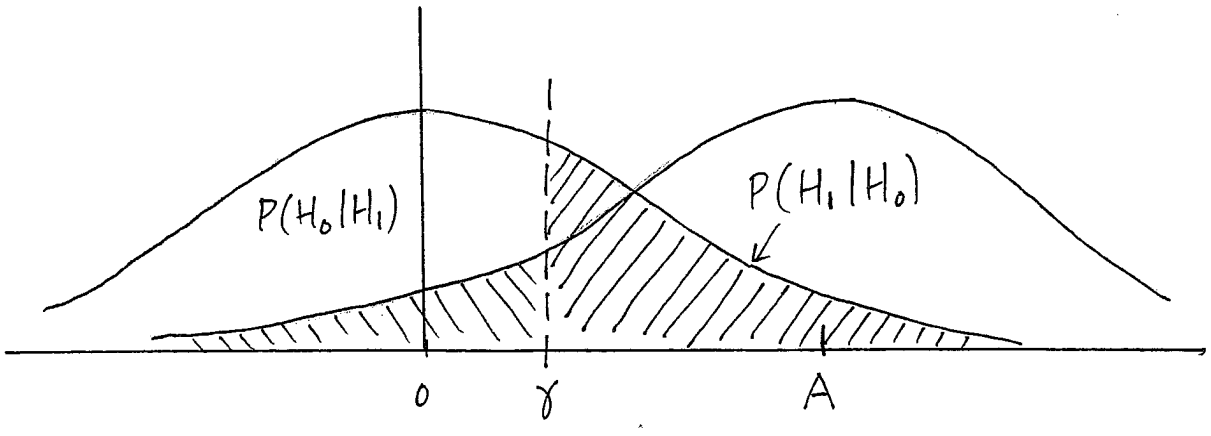
$$P(H_0|H_1) = P(T < \gamma \mid H_1)$$

$$= 1 - Q\left(\frac{\gamma - A}{\sigma/\sqrt{N}}\right).$$

Therefore,

$$P_E = \pi_0 P(H_1|H_0) + \pi_1 P(H_0|H_1)$$

$$= \pi_0 Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right) + \pi_1 \left(1 - Q\left(\frac{\gamma - A}{\sigma/\sqrt{N}}\right)\right)$$



Recall $\gamma = \frac{\sigma^2}{NA} \ln(\eta) + \frac{A}{2}$.

If $\pi_0 = \pi_1 = 1/2$ $(\eta = 1)$, then

$$\boxed{P_E = Q\left(\frac{A\sqrt{N}}{2\sigma}\right)}$$

For this problem we may define the <u>signal to noise ratio</u>

$$SNR = \frac{A^2 N}{\sigma^2}$$

Then

smaller $P_E \iff$ larger SNR

The MAP Detector

Instead of minimizing P_E, we could maximize P_c:

P_c = probability of a correct decision

$$= P(H_0, H_0) + P(H_1, H_1)$$

$$= \pi_0 \int_{R_0} f_0(\underline{x})\, d\underline{x} + \pi_1 \int_{R_1} f_1(\underline{x})\, d\underline{x}$$

So we would choose

$$\underline{x} \in R_i \iff \pi_i f_i(\underline{x}) \text{ is maximal}$$

Note: this just another way of writing the LRT

By Bayes rule,

$$P(\mathcal{H}_i|\underline{x}) = \frac{P(\mathcal{H}_i)\cdot f(\underline{x}|\mathcal{H}_i)}{f(\underline{x})}$$

$$= \frac{\pi_i f_i(\underline{x})}{f(\underline{x})}$$

same thing, different notation

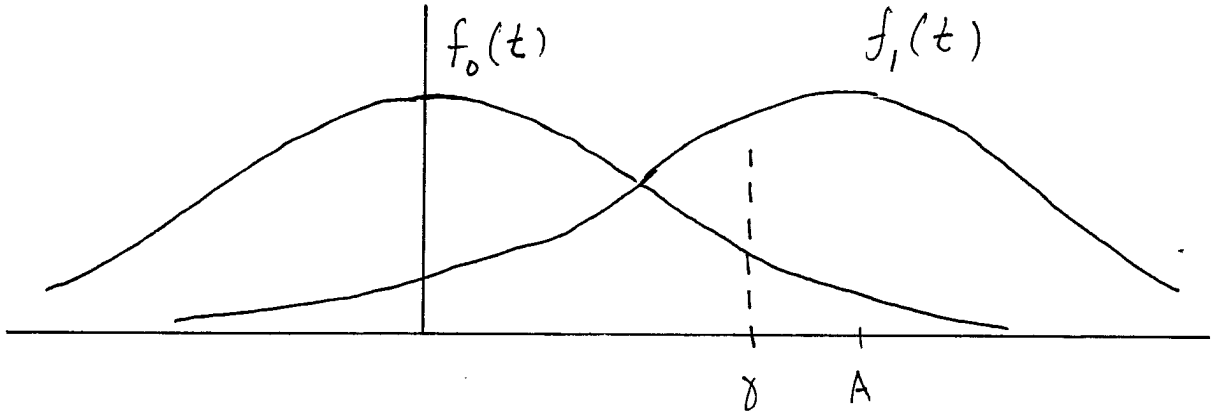
We call $P(\mathcal{H}_i|\underline{x})$ the <u>a posteriori</u> (or <u>posterior</u>) <u>probability</u> of hypothesis $\mathcal{H}_i$.

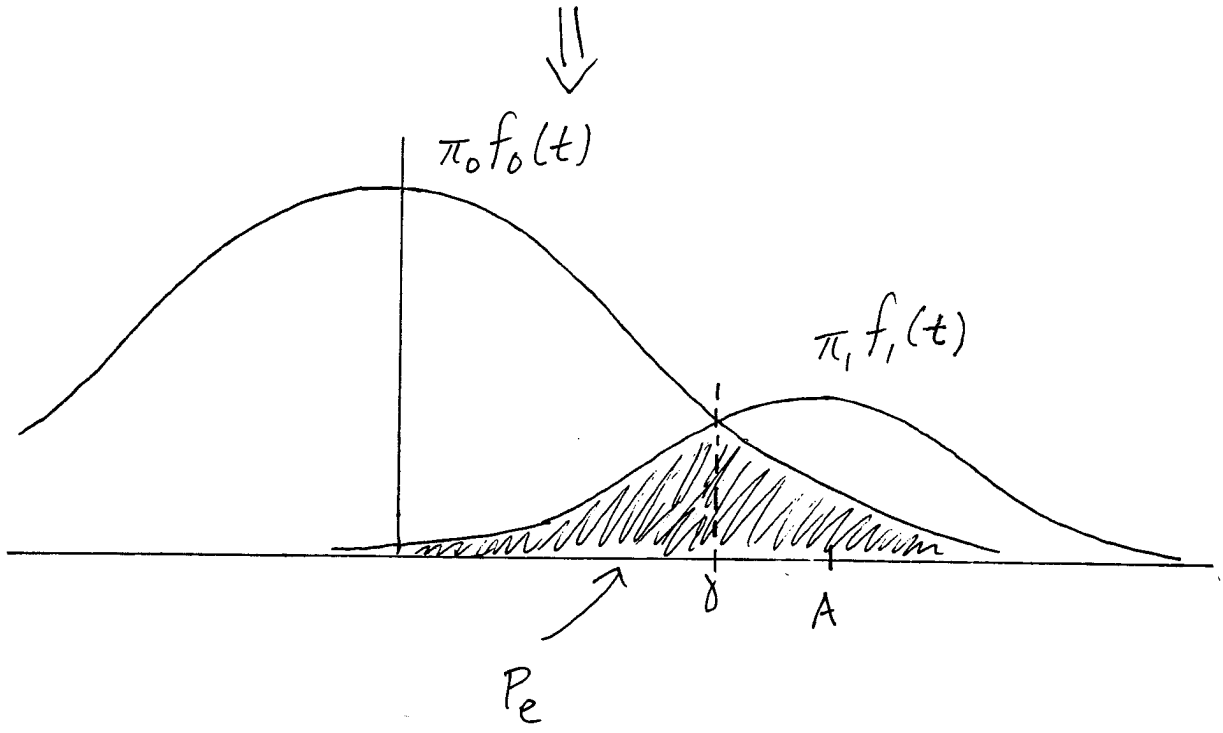
Since $f(\underline{x})$ is independent of i, maximizing $\pi_i f_i(\underline{x})$ is equivalent to maximizing $P(\mathcal{H}_i|\underline{x})$. This gives rise to the <u>maximum a posteriori probability</u> (MAP) detector:

$$\underline{x} \in R_i \iff P(\mathcal{H}_i|\underline{x}) \text{ is maximal}$$

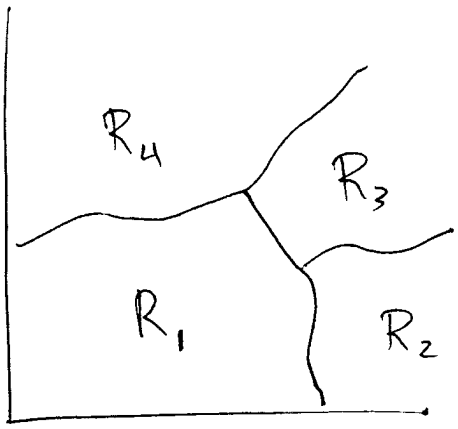
Example

Assume $\pi_0 > \pi_1$





Multiple Hypotheses



Consider

$$\mathcal{H}_1 : \underline{X} \sim f_1(\underline{x})$$

$$\vdots$$

$$\mathcal{H}_M : \underline{X} \sim f_M(\underline{x})$$

Then $$P_e = 1 - P_c = 1 - \left(\sum_i \int_{R_i} \pi_i f_i(\underline{x}) \, d\underline{x} \right)$$

$\Rightarrow$ MAP detector is optimal:

$$x \in R_i \iff \pi_i f_i(\underline{x}) \text{ is maximal}$$

Example $$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}(\underline{m}_i, I) \quad , \quad \pi_i = \frac{1}{3}$$

m_1 . . m_3

. m_2

Remark

The MAP detector is a special case of the MAP estimator.

Here, $\theta = H_0$ or $\theta = H_1$

and π_0, π_1, determine the prior.

This illustrates a more general fact: detection is a special case of estimation where the range of possible parameter values is finite.

Summary

- Bayes detector: minimizes Bayes risk
  = expected cost of a decision

- Min P_e detector = special case of
  Bayes detector

- LRT = form of Bayes detector
  for <u>binary</u> tests

- MAP rule: form of Min P_e detector
  for $M \geq 2$ hypotheses

- All the above rules assume
  $\pi_i = P(H_i)$ is known.

- Next lecture: what if π_i is not known?

Key

a. $\int_{-\infty}^{0} f_1(x)\,dx$

b. $\sum_{\underline{x} \in \mathcal{X} \cap R_i} f_j(\underline{x})$

c. $\frac{f_1(\underline{x})}{f_0(\underline{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\pi_0}{\pi_1}$

d. $T \sim N(B \cdot \underline{0}, B \cdot \sigma^2 I \cdot B^T)$

$\sim N(0, \frac{\sigma^2}{N})$ under H_0

$T \sim N(B \cdot A\underline{1}, B \cdot \sigma^2 I \cdot B^T)$

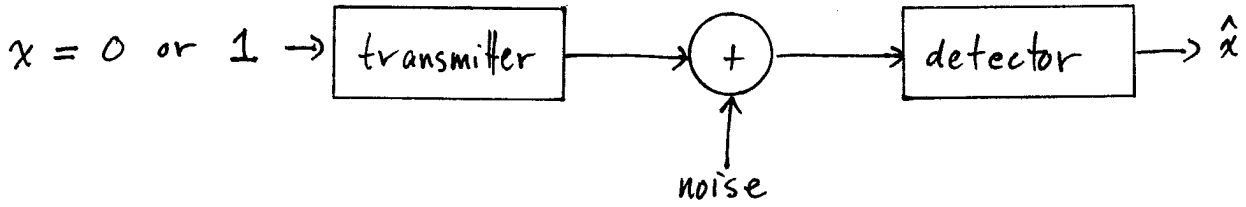
$\sim N(A, \frac{\sigma^2}{N})$ under H_1

NEYMAN-PEARSON DETECTION

In deriving the Bayes detector we assumed $\pi_i = P(H_i)$ to be known. Some times this is a reasonable assumption, other times it isn't.

Examples

1. Binary communication channel



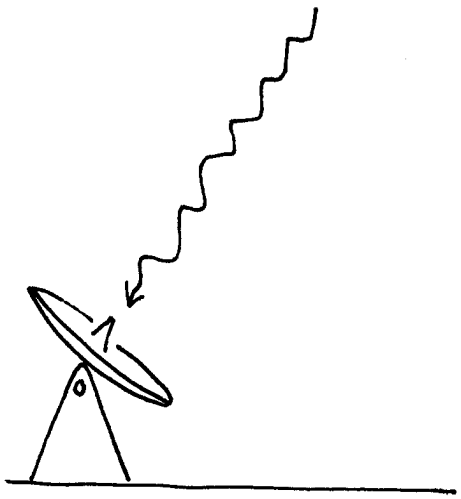
$\pi_0 = \pi_1 = \frac{1}{2}$

2. Search for extra-terrestrial life

H_0: $\underline{X}$ ~ cosmic radiation

H_1: $\underline{X}$ ~ cosmic radiation + intelligent signal

$P(H_1) = ?$



In a binary hypothesis testing problem

$$H_0 : \underline{X} \sim f_0(\underline{x})$$

$$H_1 : \underline{X} \sim f_1(\underline{x})$$

we assign names to the four possible outcomes

decision \ truth	H_0	H_1
H_0	rejection	miss
H_1	false alarm	detection

$$P_D = P(H_1 \mid H_1) \quad \text{"detection probability"}$$

$$P_M = P(H_0 \mid H_1) \quad \text{"miss"}$$

$$P_F = P(H_1 \mid H_0) \quad \text{"false alarm"}$$

$$P_R = P(H_0 \mid H_0) \quad \text{"rejection"}$$

Note that

$$P_D = 1 - P_M$$

$$P_F = 1 - P_R$$

so there are only <u>two degrees of freedom</u> for evaluating a hypothesis test.

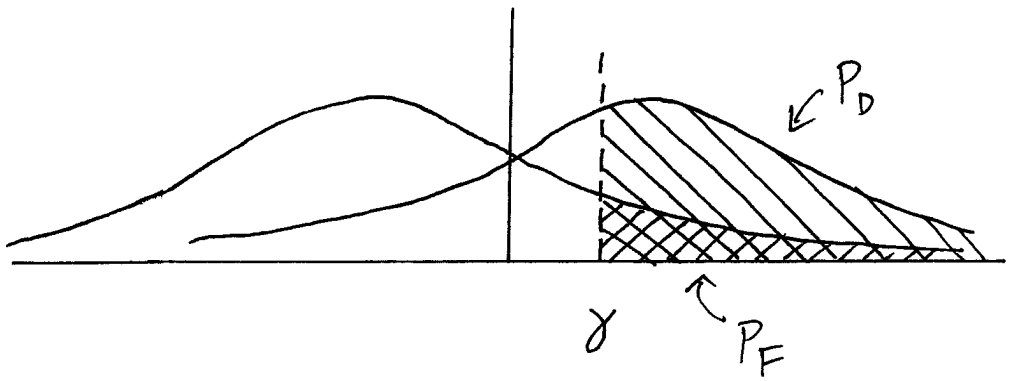
Also note: P_D and P_F do not involve <u>prior probabilities</u> on $\mathcal{H}_0, \mathcal{H}_1$.

Idea: formulate a detection criterion in terms of P_D, P_F.

Example

$H_0: X \sim \mathcal{N}(-1, 1)$

$H_1: X \sim \mathcal{N}(1, 1)$



Consider the decision rule

$$x \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

As γ increases,

P_F decreases (good)

P_D decreases (bad)

More generally, P_F and P_D are indirectly related through the <u>decision region</u> R_1:

$$P_D = \int_{R_1} f_1(\underline{x})\, d\underline{x}$$

$$P_F = \int_{R_1} f_0(\underline{x})\, d\underline{x}$$

As R_1 <u>expands</u>, P_D and P_F increase

As R_1 <u>shrinks</u>, P_D and P_F decrease

Ideally, we would like

$$P_D = 1 \quad , \quad P_F = 0$$

but this is only possible when

(a)

So what is the best way to choose R_1?

The Neyman-Pearson Criterion

The Neyman-Pearson (NP) detector solves the following optimization problem:

$$\max \quad P_D$$
$$\text{s.t. } P_F \leq \alpha$$

In words, the NP detector has the largest detection probability among all detectors with false alarm probability no greater than α.

Terminology

P_D = power

P_F = size

So the NP detector is the most powerful test of size (not exceeding) α.

The Neyman-Pearson Lemma

Let $\alpha \in [0,1]$. The NP detector is

$$\Lambda(\underline{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

where $\Lambda(\underline{x}) = \dfrac{f_1(\underline{x})}{f_0(\underline{x})}$ and η is chosen such that

$$P_F = \int_{\Lambda(\underline{x})>\eta} f_0(\underline{x})\, d\underline{x} = \alpha$$

Note! It may not always be possible to set η such that $P_F = \alpha$, such as in the case of discrete data. We will return to this case later. For now, assume $P_F = \alpha$ is achievable.

So the optimal detector is once again a <u>likelihood ratio test.</u>

The Bayes detector and NP detector lead to the same test. The difference is in how we select the <u>threshold</u> η.

- For the Bayes detector:

$$\eta = \frac{\pi_0}{\pi_1} \cdot \frac{(c_{10} - c_{00})}{(c_{01} - c_{11})}$$

- For the NP detector

$$\eta = P_F^{-1}(\alpha)$$

where

$$P_F(\eta) = \int_{\Lambda > \eta} f_0(\underline{x})\, d\underline{x}$$

We can prove the NP Lemma using the theory of constrained optimization (Lagrange multiplier theory)

Constrained Optimization

Consider the problem

$$\max_{x \in \Omega} h(x) \quad \text{subject to } g(x) \leq C,$$

where $h, g : \Omega \to \mathbb{R}$, $C \in \mathbb{R}$, and Ω is an arbitrary set.

Theorem Let $\lambda \geq 0$ and suppose $x_0(\lambda) \in \Omega$ maximizes

$$L(x, \lambda) \equiv h(x) - \lambda g(x)$$

for each λ. Then $x_0(\lambda)$ maximizes h over all x such that $g(x) \leq g(x_0(\lambda))$.

Corollary: If λ^* is such that $g(x_0(\lambda^*)) = C$, then $x_0(\lambda^*)$ maximizes $h(x)$ over all x such that $g(x) \leq C$.

Proof of Theorem:

By assumption, $x_0 = x_0(\lambda)$ satisfies

$$h(x_0) - \lambda g(x_0) \geq h(x) - \lambda g(x)$$

for all $x \in \Omega$. Equivalently,

$$h(x_0) - h(x) \geq \lambda(g(x_0) - g(x_0))$$

for all $x \in \Omega$. Define

$$S = \{x \in \Omega \mid g(x) \leq g(x_0)\}.$$

Then for all $x \in S$,

$$h(x_0) - h(x) \geq \lambda(g(x_0) - g(x)) \geq 0$$

so $h(x_0) \geq h(x) \quad \forall x \in S.$ ∎

Proof of Neyman-Pearson Lemma

Apply constrained optimization theorem with

$h = P_D$

$g = P_F$

$C = \alpha$

$x = R_1$

Ω = all possible decision regions

To do this, we must find $R_1 = R_1(\lambda)$ that maximizes

$$L = P_D - \lambda P_F$$

$$= \int_{R_1} f_1(\underline{x})\,dx - \lambda \int_{R_1} f_0(\underline{x})\,d\underline{x}$$

$$= \int_{R_1} [f_1(\underline{x}) - \lambda f_0(\underline{x})]\,d\underline{x}$$

So choose

$$R_1 = \left\{ \underline{x} : \frac{f_1(\underline{x})}{f_0(\underline{x})} > \lambda \right\}.$$

Now, to maximize P_D over all R_1 such that $P_F \leq \alpha$, we take λ such that

$$P_F = \int_{\Lambda > \lambda} f_0(\underline{x}) d\underline{x} = \alpha$$

■

Setting the threshold

Computing η such that $P_F = \alpha$ is not always easy. It usually requires the use of monotonic transformations and test statistics as the following example demonstrates.

Example DC signal in AWGN

$$H_0: \underline{X} \sim \mathcal{N}(\underline{0}, \sigma^2 \underline{\underline{I}})$$

$$H_1: \underline{X} \sim \mathcal{N}(A\underline{1}, \sigma^2 \underline{\underline{I}}) \quad , \quad A > 0$$

Let's design a NP detector

From the last lecture we saw

$$\Lambda(\underline{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

$$\Updownarrow$$

$$\frac{1}{N}\sum_{n=1}^{N} x_n \equiv t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \equiv \frac{\sigma^2}{NA}\ln(\eta) + \frac{A}{2}$$

Recall $t \sim \mathcal{N}(0, \frac{\sigma^2}{N})$ under H_0

$t \sim \mathcal{N}(A, \frac{\sigma^2}{N})$ under H_0

Exercise (a) Use the Q function to express P_F, P_D in terms of γ and known quantities (b) Find γ for the NP detector of size α (c) Express P_D in terms of P_F and SNR.

Solution

$$P_F = \text{Prob}(t > \gamma \mid H_0)$$

$$= Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right) \leq \alpha$$

$$P_D = \text{Prob}(t > \gamma \mid H_1)$$

$$= Q\left(\frac{\gamma - A}{\sigma/\sqrt{N}}\right)$$

To set the threshold, we take

$$P_F = \alpha$$

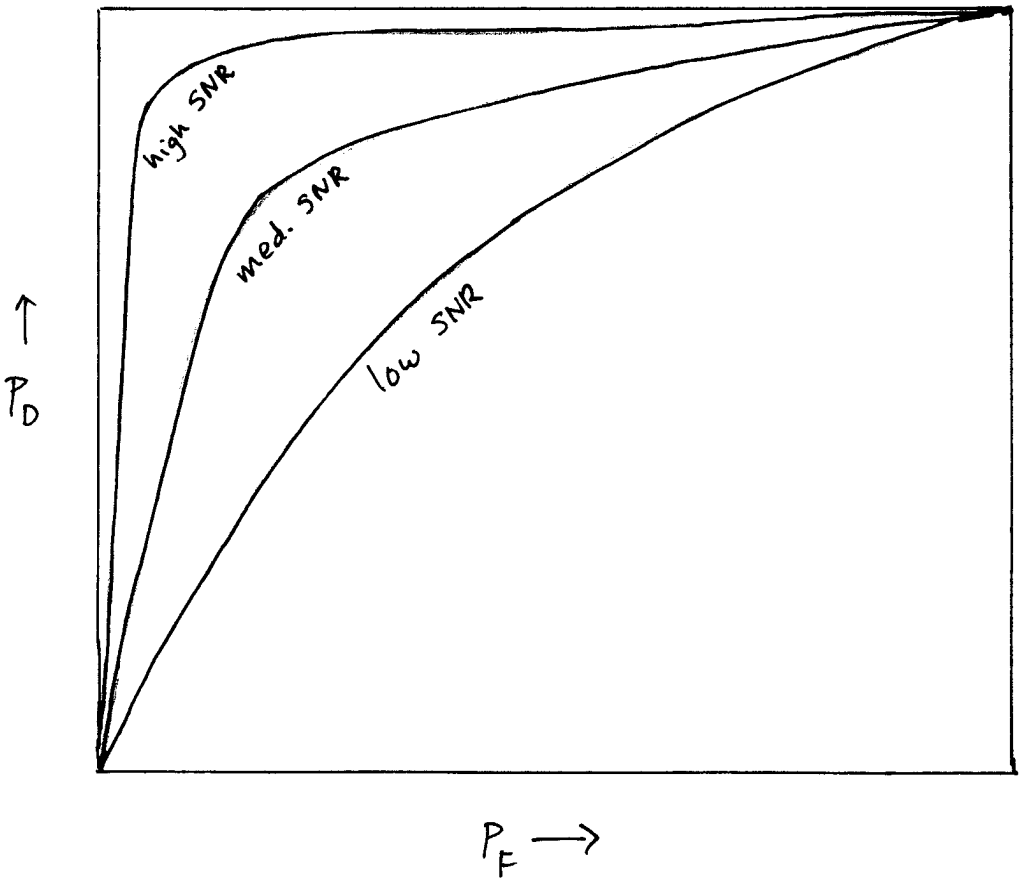
$$\Rightarrow \quad \gamma = \frac{\sigma}{\sqrt{N}} Q^{-1}(\alpha)$$

$$\Rightarrow \quad P_D = Q\left(Q^{-1}(P_F) - \frac{A\sqrt{N}}{\sigma}\right)$$

$$= Q\left(Q^{-1}(P_F) - \sqrt{SNR}\right)$$

The Receiver Operating Characteristic

The ROC of a detector is a plot of P_D vs. P_F.



Ex

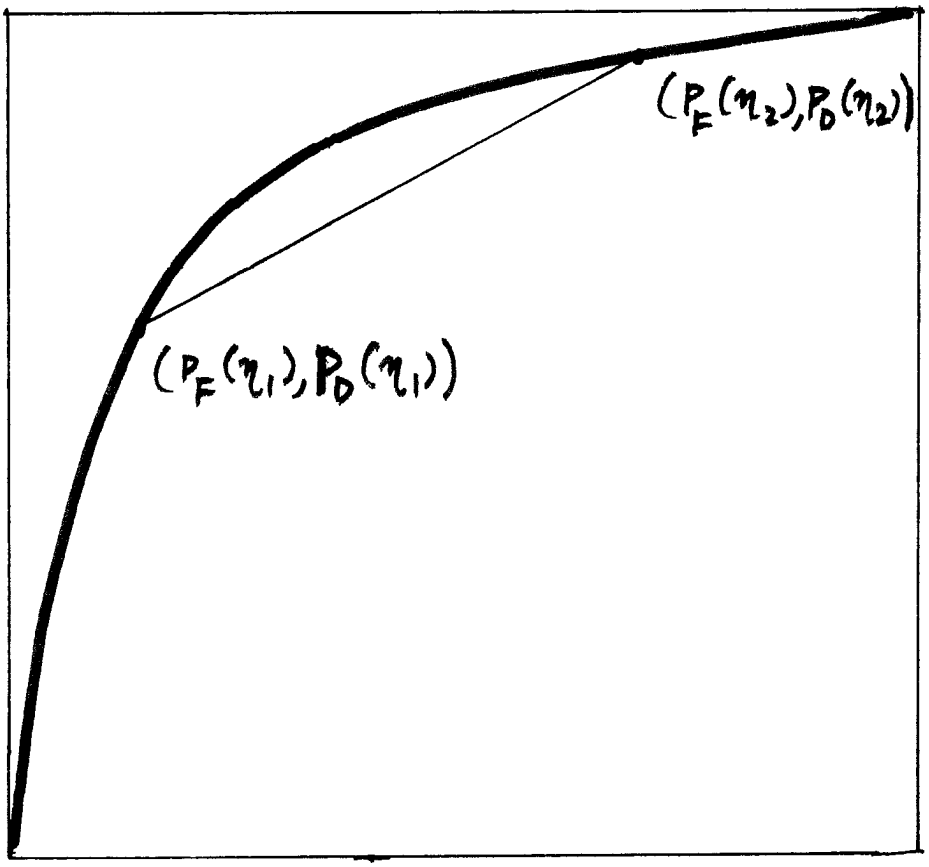
$$P_D = Q\left(Q^{-1}(P_F) - \frac{A\sqrt{N}}{\sigma}\right)$$

$$SNR = \frac{A^2 N}{\sigma^2}$$

Fact The ROC of the LRT is **concave**.

$$P_D(\eta) = P_r(\Lambda(\underline{x}) > \eta \mid H_1)$$

$$P_F(\eta) = P_r(\Lambda(\underline{x}) > \eta \mid H_0)$$



In other words, for all $\eta_1, \eta_2 \geq 0$, the line segment

$$\left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} P_F(\eta_1) \\ P_D(\eta_1) \end{bmatrix} + (1-\lambda) \begin{bmatrix} P_F(\eta_2) \\ P_D(\eta_2) \end{bmatrix}, \lambda \in [0,1] \right\}$$

is **below** the ROC.

Let's prove this. Suppose it's not true.

Then $\exists\, \eta_1, \eta_2$ and $\lambda \in [0,1]$ such that

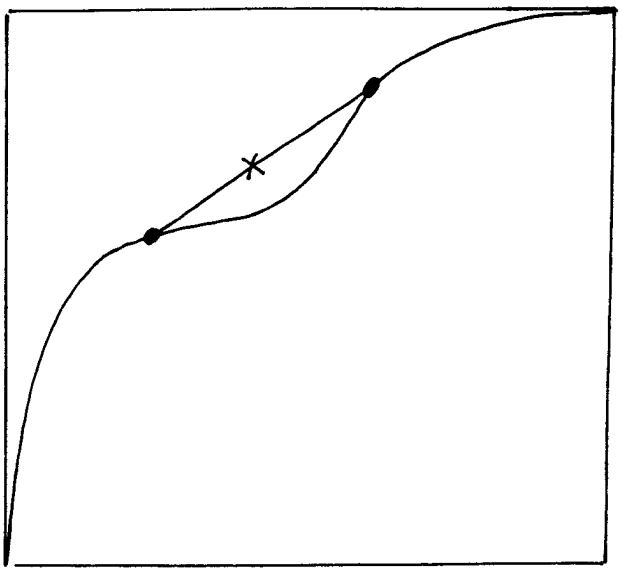
$$\lambda \begin{bmatrix} P_F(\eta_1) \\ P_D(\eta_1) \end{bmatrix} + (1-\lambda) \begin{bmatrix} P_F(\eta_2) \\ P_D(\eta_2) \end{bmatrix}$$

is <u>above</u> the ROC.

Consider the rule

$$\begin{cases} \text{use} \quad \Lambda(\underline{x}) \gtrless \eta_1 & \text{with prob. } \lambda \\ \text{use} \quad \Lambda(\underline{x}) \gtrless \eta_2 & \text{with prob } 1-\lambda \end{cases}$$

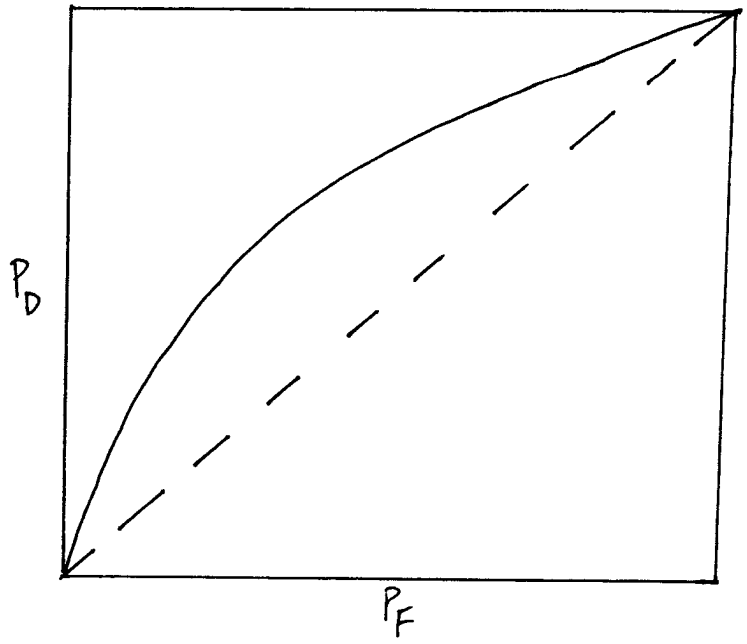
Then
$$P_F = \lambda P_F(\eta_1) + (1-\lambda) P_F(\eta_2)$$
$$P_D = \lambda P_D(\eta_1) + (1-\lambda) P_D(\eta_2)$$



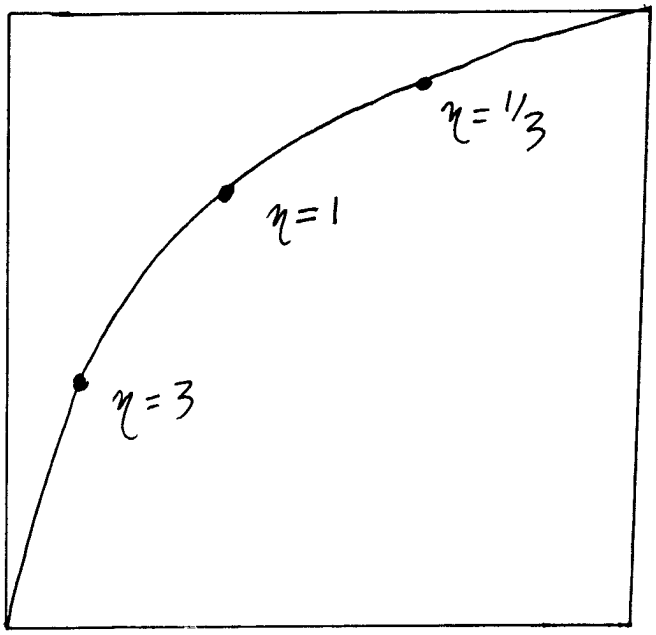
This <u>contradicts</u> the optimality of the LRT.

Fact 2 The ROC of the LRT is above the line $P_D = P_F$.

(b) **Proof?**



Fact 3 The <u>slope</u> of the ROC (for the LRT) at a point $(P_F(\eta), P_D(\eta))$ is η.



That is,

$$\frac{dP_D}{dP_F} = \eta$$

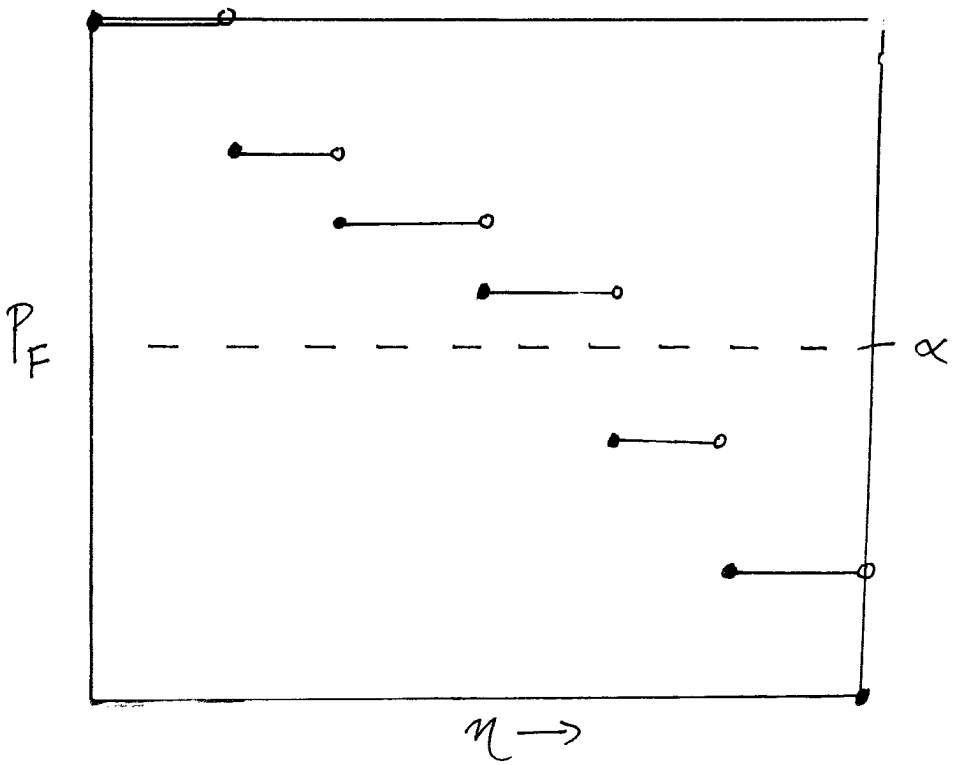
Discrete data

Suppose the data $\underline{X}$ is discrete.

Then

$$P_F = \sum_{\underline{x}:\Lambda(\underline{x})>\eta} f_0(\underline{x})$$

So it may not be possible to have $P_F = \alpha$ for all α in the current setup.



What if we choose η so that P_F is as large as possible? Does the LRT still solve

$$\max P_D \quad ?$$
$$\text{s.t. } P_F \leq \alpha$$

Not quite.

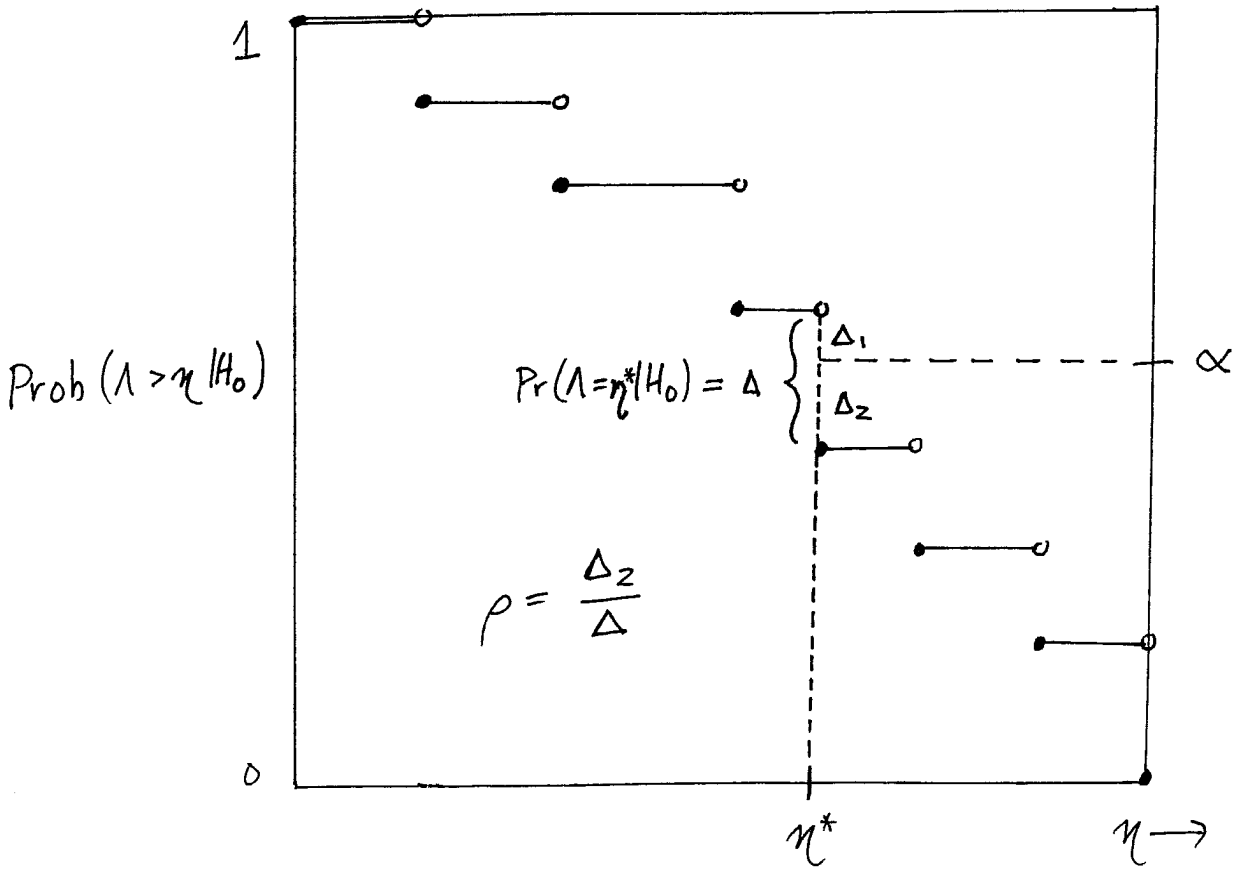
We must now be concerned with the case $\Lambda(\underline{x}) = \eta$, which can occur with <u>nonzero</u> probability.

Let $\alpha \in [0,1]$, and let η^* be as small as possible such that

$$Pr(\Lambda > \eta^* | H_0) < \alpha$$

Choose $\rho \in [0,1)$ such that

$$Pr(\Lambda > \eta^* | H_0) + \rho Pr(\Lambda = \eta^* | H_0) = \alpha$$



Consider the decision rule

$$\begin{cases} \text{declare } H_1 & \text{if } \Lambda(\underline{x}) > \eta^* \\ \text{flip a "}\rho\text{-coin"} & \text{if } \Lambda(\underline{x}) = \eta^* \\ \text{declare } H_0 & \text{if } \Lambda(\underline{x}) < \eta^* \end{cases}$$

Then $P_F = \alpha$

A "ρ-coin" turns up heads (H_1) with probability ρ.

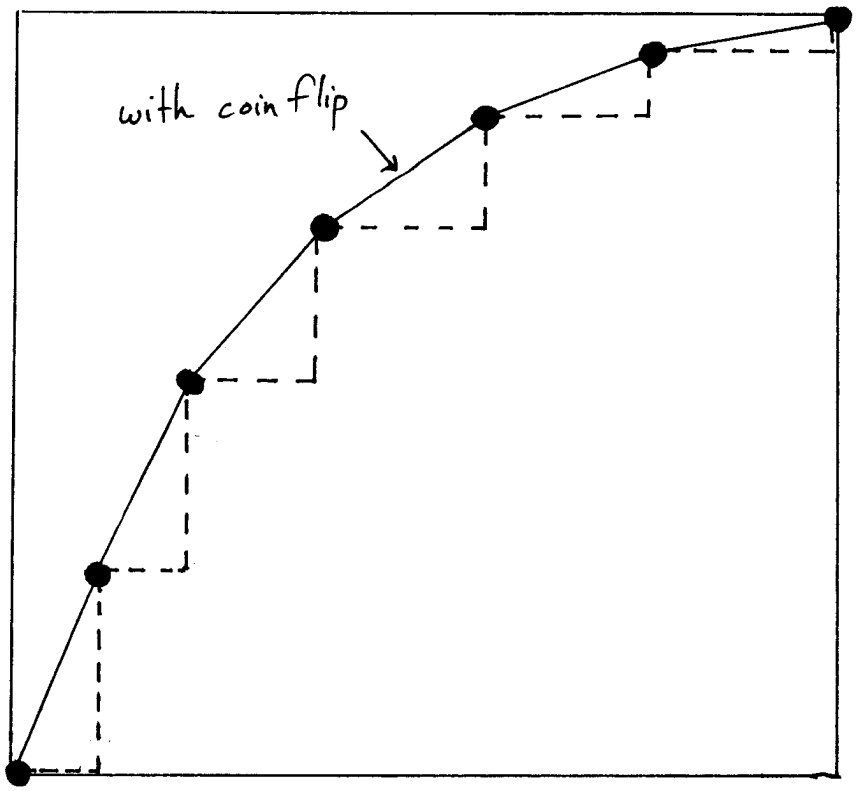
If you think back to the proof of the NP Lemma, we can redistribute

$$\{\underline{x} : \Lambda(\underline{x}) = \eta\}$$

as we see fit and still maximize the Lagrangian. We use just enough probability mass to bring P_F up to α and use the rest to increase P_D.

The <u>modified</u> LRT is now <u>optimal</u> for the case of <u>disrete</u> data.

Intuition:



Old LRT: $\Lambda \gtrless \eta$

operates at discrete set of (P_F, P_D)

New LRT (with coin flip)

ROC = "convex hull" of old ROC

Summary

- Neyman-Pearson detector:
  maximizes P_D s.t. $P_F \leq \alpha$

- NP criterion does not assume knowledge of prior probabilities of each hypothesis (frequentist)

- The Bayes risk detector does (Bayesian)

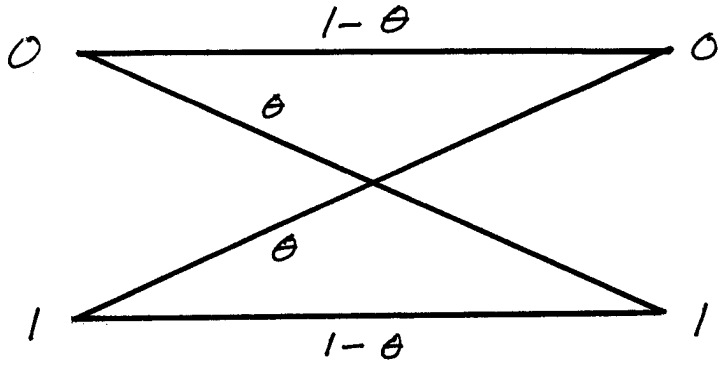
- Optimal detector for both criteria given by LRT.

Key a. $\text{support}(f_0) \cap \text{support}(f_1) = \emptyset$

b. The line $P_D = P_F$ corresponds to random guessing, and the NP detector does at least as well as that.

APPLICATION: BINARY SYMMETRIC CHANNEL

In a binary symmetric channel (BSC), whenever we transmit a bit (0 or 1), the bit is <u>flipped</u> with probability $\theta \in [0,1]$.



BSC is a very common model in digital comm. systems

If we denote

x = transmitted bit

y = received bit

then

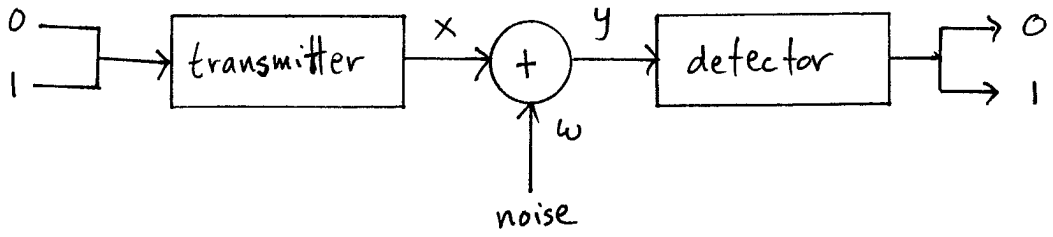
$$y = x + w$$

where

$$w \sim \text{Bernoulli}(\theta)$$

modular arithmetic: $1+1=0$

Our job is to build a detector to determine the transmitted bit from the received bit



Let's set this up as a hypothesis testing problem:

$$H_0: y = 0 + w$$

$$H_1: y = 1 + w$$

Equivalently

$$H_0: y \sim \text{Bernoulli}(\theta)$$

$$H_1: y \sim$$

(a)

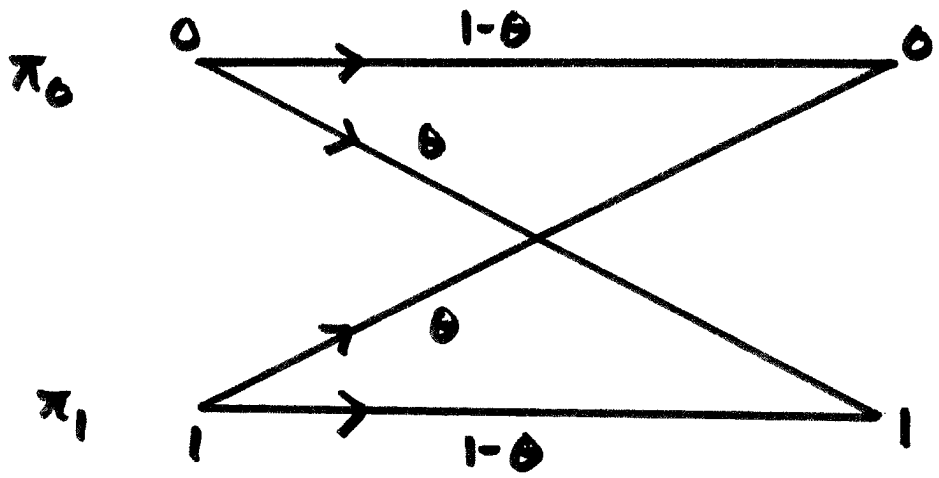
Our goal is effectively to form an estimate $\hat{x}$ of the transmitted bit x.

Let's assume $\theta < \frac{1}{2}$: more often than not a bit is uncorrupted.

Also assume prior probabilities

$$\pi_i = \Pr\{X = i\}$$

are known.



How should we decode / detect y so as to minimize the probability of error?

We want to minimize $P\{\hat{x} \neq x\}$.

Since this is a very simple detection problem, we can write down all 4 possible decision rules.

1. $0 \to 0$, $1 \to 1$

2. $0 \to 0$, $1 \to 0$

3. $0 \to 1$, $1 \to 1$

4. $0 \to 1$, $1 \to 0$

$$P_e = P\{\hat{x} \neq x\} = P(H_0, H_1) + P(H_1, H_0)$$
$$= \pi_1 P(H_0 \mid H_1) + \pi_0 P(H_1 \mid H_0)$$

(b)

$$= \begin{cases} 1. \\ 2. \\ 3. \\ 4. \end{cases}$$

Which is smallest?

Since $\theta < \frac{1}{2}$, we know $\theta < 1-\theta$,
so the 4th rule never has smallest P_e.

So there are 3 cases:

1) If $\theta < \pi_1, \pi_2$ $\quad \hat{x} =$

2) If $\pi_1 < \theta$ $\quad \hat{x} =$

3) If $\pi_0 < \theta$ $\quad \hat{x} =$

In general, it will be extremely tedious, if not impossible, to enumerate all possible decision rules.

Fortunately, there is a systematic way to find the decision rule with smallest P_e.

P_e is minimized by the LRT:

$$\Lambda(y) = \frac{f_1(y)}{f_0(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\pi_0}{\pi_1} = \eta$$

where

$$f_1(y) = \begin{cases} 1-\theta & \text{if } y=1 \\ \theta & \text{if } y=0 \end{cases}$$

$$f_0(y) = \begin{cases} \theta & \text{if } y=1 \\ 1-\theta & \text{if } y=0 \end{cases}$$

So

(d) $\Lambda(y) =$

$=$

Let's apply <u>monotonic transformations</u> to simplify the LRT:

$$\frac{\theta}{1-\theta}\cdot\left(\frac{1-\theta}{\theta}\right)^{2y} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

$$\Updownarrow$$

$$\left(\frac{1-\theta}{\theta}\right)^{2y} \underset{H_0}{\overset{H_1}{\gtrless}} \eta\cdot\frac{1-\theta}{\theta}$$

$$\Updownarrow$$

$$2y\ln\left(\frac{1-\theta}{\theta}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \ln(\eta)+\ln\left(\frac{1-\theta}{\theta}\right)$$

$$\Updownarrow$$

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2}+\frac{1}{2}\cdot\frac{\ln(\eta)}{\ln\left(\frac{1-\theta}{\theta}\right)} \equiv \gamma$$

Note: $\theta<\frac{1}{2} \Rightarrow \frac{1-\theta}{\theta}>1 \Rightarrow \ln\left(\frac{1-\theta}{\theta}\right)>0$

Three cases:

threshold	decision rule
$\gamma > 1$	$\hat{x} = 0$
$\gamma < 0$	$\hat{x} = 1$
$0 < \gamma < 1$	$\hat{x} = y$

Question Do we need to worry about $\gamma = 0$ or $\gamma = 1$?

Let's consider each case: Recall $\gamma = \frac{1}{2} + \frac{1}{2}\frac{\ln(\eta)}{\ln\left(\frac{1-\theta}{\theta}\right)}$

$$\boxed{\gamma > 1} \iff \frac{\ln(\eta)}{\ln\left(\frac{1-\theta}{\theta}\right)} > 1$$

$$\iff \ln(\eta) > \ln\left(\frac{1-\theta}{\theta}\right)$$

$$\iff \eta > \frac{1-\theta}{\theta} \qquad \left[\text{Recall } \eta = \frac{\pi_0}{\pi_1} = \frac{1-\pi_1}{\pi_1}\right]$$

$$\iff \frac{1-\pi_1}{\pi_1} > \frac{1-\theta}{\theta}$$

$$\iff \frac{1}{\pi_1} - 1 > \frac{1}{\theta} - 1$$

$$\iff \frac{1}{\pi_1} > \frac{1}{\theta} \iff \boxed{\pi_1 < \theta}$$

$\boxed{\gamma < 0} \iff \frac{\ln(\eta)}{\ln\left(\frac{1-\theta}{\theta}\right)} < -1$

$\iff \ln(\eta) < \ln\left(\frac{\theta}{1-\theta}\right)$

$\iff \eta < \frac{\theta}{1-\theta}$

$\iff \frac{\pi_0}{1-\pi_0} < \frac{\theta}{1-\theta}$

$\iff \frac{1-\pi_0}{\pi_0} > \frac{1-\theta}{\theta}$

$\iff \frac{1}{\pi_0} - 1 > \frac{1}{\theta} - 1$

$\iff \frac{1}{\pi_0} > \frac{1}{\theta} \iff \boxed{\pi_0 < \theta}$

$\boxed{0 < \gamma < 1}$ From the previous two cases, we conclude

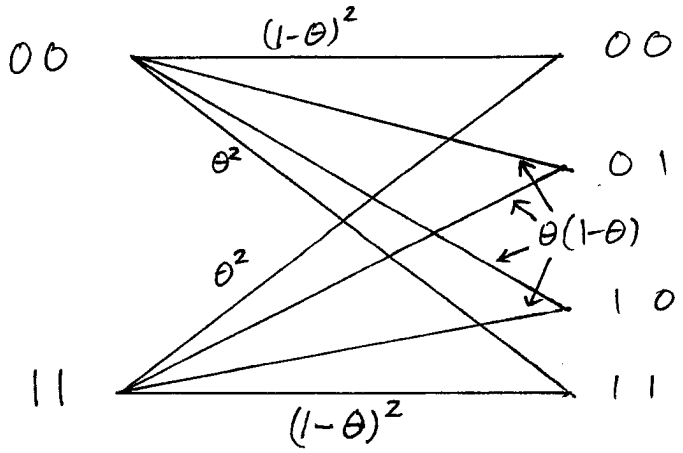
$0 < \gamma < 1 \iff \boxed{\theta < \pi_1 \text{ and } \theta < \pi_2.}$

Conclusion. The LRT coincides with the minimum probability of error detector we derived by hand, for each case of θ, π_1, π_2.

Binary Repitition Code

In an effort to decrease P_e, let's send N copies of each "information" bit:

Example $N = 2$



As a hypothesis testing problem, we have

$$H_0: \underline{y} = 00\cdots0 + \underline{\omega}$$

$$H_1: \underline{y} = 11\cdots1 + \underline{\omega}$$

where

$$\underline{y} = (y_1, \cdots, y_N)^T$$

$$\underline{\omega} = (\omega_1, \cdots, \omega_N)^T$$

$$\omega_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

The brute force approach is impractical for large N.

(e) The total number of possible decision rules is _____.

Exercise] Apply the LRT to determine the min P_E detector. Express your answer in terms of a one dimensional test statistic. Derive a formula for P_E in terms of N, θ, and $\eta = \frac{\pi_0}{\pi_1}$

Solution

$$\Lambda(\underline{y}) = \frac{f_1(\underline{y})}{f_0(\underline{y})} = \frac{\prod_{n=1}^{N} (1-\theta)^{y_n} \theta^{1-y_n}}{\prod_{n=1}^{N} \theta^{y_n} (1-\theta)^{1-y_n}}$$

$$= \frac{(1-\theta)^k \theta^{N-k}}{\theta^k (1-\theta)^{N-k}} = \left(\frac{1-\theta}{\theta}\right)^{2k-N}$$

$k = \sum_{n=1}^{N} y_n$ ← sufficient statistic

After some simplification

$$\Lambda(\underline{y}) \underset{H_0}{\overset{H_1}{\lessgtr}} \eta \iff (2k-N) \ln\left(\frac{1-\theta}{\theta}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \ln(\eta)$$

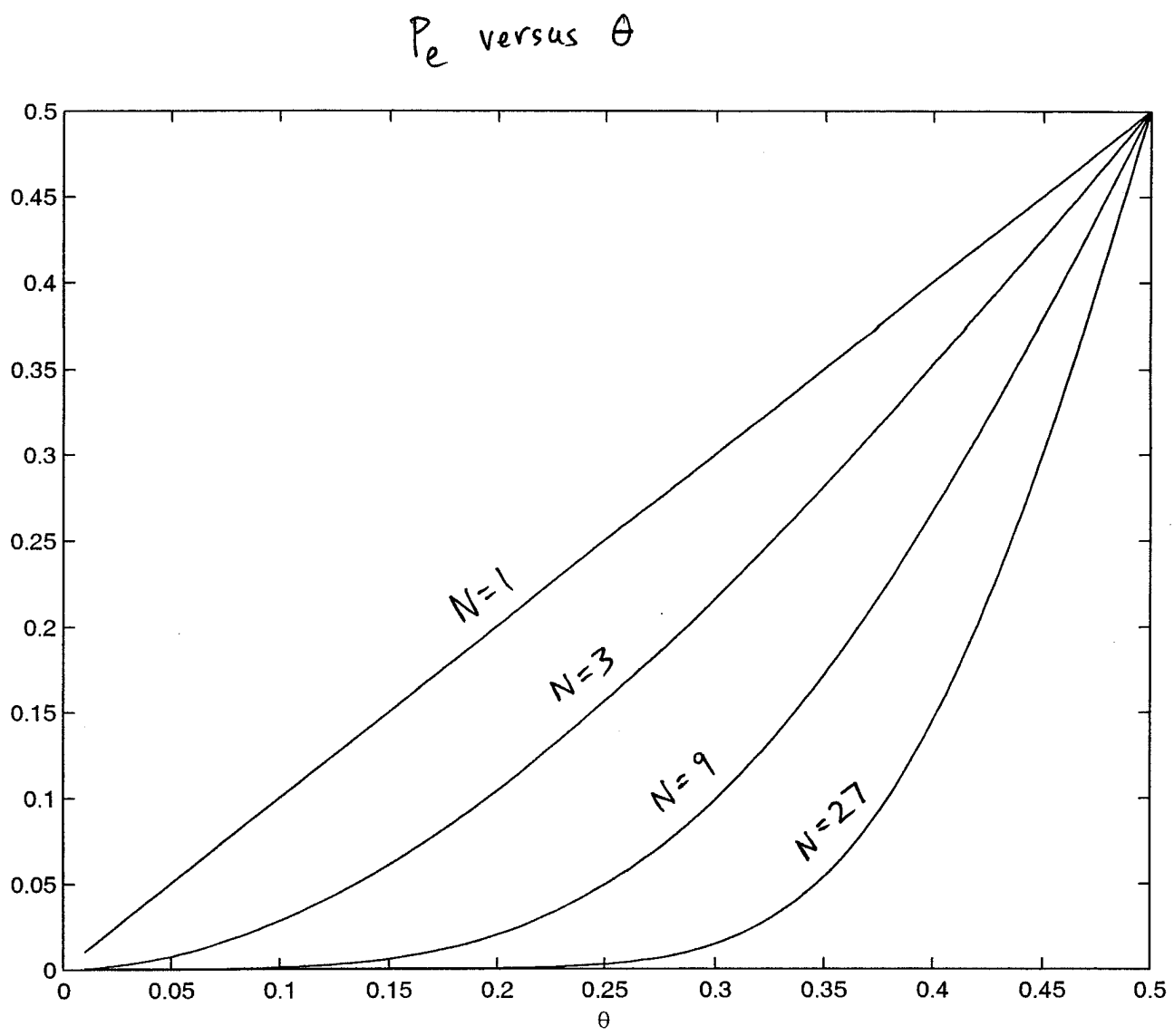
$$\iff k \underset{H_0}{\overset{H_1}{\gtrless}} \frac{N}{2} + \frac{1}{2} \frac{\ln(\eta)}{\ln\left(\frac{1-\theta}{\theta}\right)} \equiv \gamma$$

In the special case $\eta = 1$ $\left(\pi_0 = \pi_1 = \frac{1}{2}\right)$,

$$k \underset{H_0}{\overset{H_1}{\gtrless}} \frac{N}{2}$$

"majority vote"

In case $k = \gamma$, let's declare H_0.


P_e versus θ
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
θ
N=1
N=3
N=9
N=27

```
% Binary repitition code

close all

for N=[1 3 9 27]

    Npts=100;
    p=linspace(.01,.5,Npts);

    pi0=1/2;
    pi1=1-pi0;
    eta=pi0/pi1;

    for j=1:Npts
        gam=N/2 +.5*(log(eta)/log((1-p)/p));

        pmf0=zeros(1,N+1);
        for k=0:N
            pmf0(k+1)=nchoosek(N,k)*(p(j)^k)*((1-p(j))^(N-k));
        end

        Pf = sum(pmf0(find(0:N>gam)));

        pmf1=zeros(1,N+1);
        for k=0:N
            pmf1(k+1)=nchoosek(N,k)*((1-p(j))^k)*(p(j)^(N-k));
        end

        Pm = sum(pmf1(find(0:N<=gam)));

        Pe(j)=pi0*Pf+pi1*Pm;
    end

    plot(p,Pe)
    hold on
end

xlabel('\theta');
```

Now let's design a Neyman-Pearson detector.

To be concrete, let's say $N=15$, $\theta = \frac{1}{4}$, $\alpha = 0.2$

We need to find γ, ρ such that

$$P(k > \gamma \mid H_0) + \rho P(k = \gamma \mid H_0) = \alpha$$

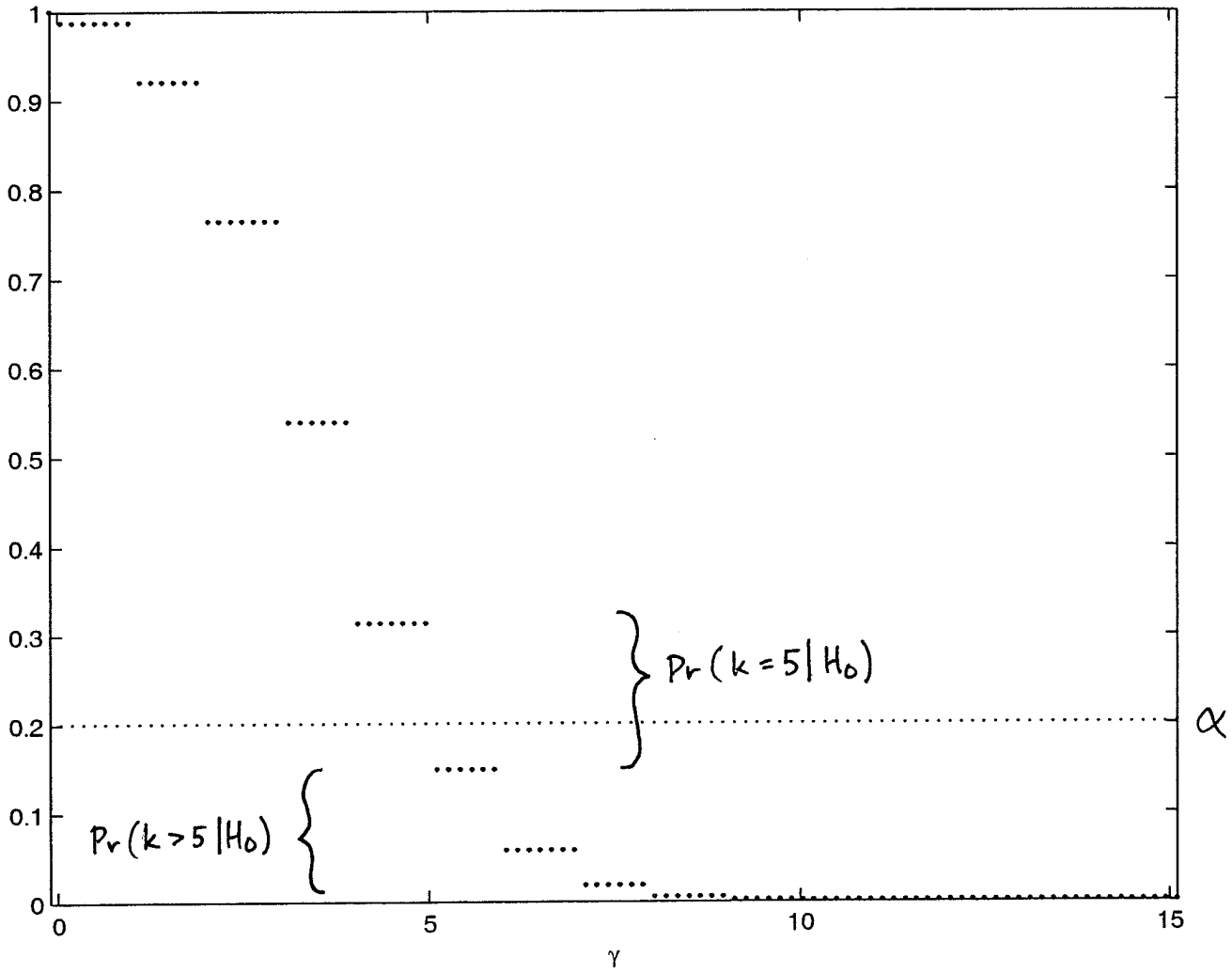
where $\gamma \in \{0, 1, \ldots, N\}$, $\rho \in [0, 1)$.

[see plot on next page]

So the NP detector is

- If $k > 5$, declare H_1
- If $k < 5$, declare H_0
- If $k = 5$, flip a coin that turns up heads (H_1) with probability $\rho = .3126$

$$Pr(k > \gamma \mid H_0)$$



$$\rho = \frac{\alpha - Pr(k > 5 \mid H_0)}{Pr(k=5)} = .3126$$

$$\Longleftrightarrow \quad P_F = Pr(k > 5 \mid H_0) + \rho Pr(k=5 \mid H_0) = \alpha$$

```
% Binary repitition code

close all

N=15;
alpha=0.2;
p=.25;

pmf0=zeros(1,N+1);
for k=0:N
    pmf0(k+1)=nchoosek(N,k)*(p^k)*((1-p)^(N-k));
end

Npts=100;
gam = linspace(-.1,N+.1,Npts);
for j=1:Npts
    Pf(j) = sum(pmf0(find(0:N>gam(j))));
end

plot(gam,Pf,'.')
hold on
plot(gam,alpha,'--')

xlabel('\gamma');
axis tight

% gamma = 5 by inspection of graph
rho = (alpha - sum(pmf0(0:N>5)))/pmf0(5+1)  % recall Matlab indexing starts↙
 at 1
```

Error Correcting Codes

There are more sophisticated ways of introducing redundancy in order to reduce the probability of error per "information bit."

Example Hamming (7,4) code

$\underline{x}_1$	0000000	$\underline{x}_9$	1000011
$\underline{x}_2$	0001111	$\underline{x}_{10}$	1001100
$\underline{x}_3$	0010110	$\underline{x}_{11}$	1010101
$\underline{x}_4$	0011001	$\underline{x}_{12}$	1011010
$\underline{x}_5$	0100101	$\underline{x}_{13}$	1100110
$\underline{x}_6$	0101010	$\underline{x}_{14}$	1101001
$\underline{x}_7$	0110011	$\underline{x}_{15}$	1110000
$\underline{x}_8$	0110100	$\underline{x}_{16}$	1111111

("information bits" ↙ ↘ "parity check bits")

M-ary hypothesis test

$$H_1 : \underline{y} = \underline{x}_1 + \underline{w}$$

$$\vdots$$

$$H_M : \underline{y} = \underline{x}_M + \underline{w}$$

Recall the MAP detector:

> Choose H_i such that $\pi_i f_i(\underline{y})$ is maximal

In communication systems, we often know

$$\pi_1 = \pi_2 = \cdots = \pi_M = \frac{1}{M}$$

in which case we get the maximum likelihood detector:

> Choose H_i such that $f_i(\underline{y})$ is maximal

In general, $f_i(\underline{y})$ is easily computed as a product of Bernoulli likelihoods, since $\underline{x}_i$ is known.

Key

a. Bernoulli $(1-\theta)$

b. 1. $\pi_1 \theta + \pi_0 \cdot \theta = \theta$

2. $\pi_1 \cdot 1 + \pi_0 \cdot 0 = \pi_1$

3. $\pi_1 \cdot 0 + \pi_0 \cdot 1 = \pi_0$

4. $\pi_1 (1-\theta) + \pi_0 (1-\theta) = 1-\theta$

c. 1. $\hat{x} = y$

2. $\hat{x} = 0$

3. $\hat{x} = 1$

d. $\Lambda(y) = \dfrac{(1-\theta)^y \theta^{1-y}}{\theta^y (1-\theta)^{1-y}} = \dfrac{\theta}{1-\theta} \cdot \left(\dfrac{1-\theta}{\theta}\right)^{2y}$

e. 2^{2^N}

f. Under H_0, $k \sim \text{Binom}(N, \theta)$

$$f_0(k) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$$P_F = \sum_{k>\gamma} \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Under H_1, $k \sim \text{Binom}(N, 1-\theta)$

SIGNAL DETECTION IN GAUSSIAN NOISE

The Gaussian noise model is the most common noise model in detection problems..

We will study the detection problem

$$H_1: \underline{x} = \underline{s}_1 + \underline{w}$$

$$\vdots$$

$$H_M: \underline{x} = \underline{s}_M + \underline{w}$$

where

$$\underline{w} \sim N(\underline{0}, R).$$

and R is symmetric, positive definite.

R and $\underline{s}_1, \ldots, \underline{s}_M$ are assumed <u>known</u>.

IID Gaussian Noise

For now, assume $R = \sigma^2 I_{N\times N}$.

Let's also assume $M=2$. and $\underline{s}_0 = \underline{0}$.

$$H_0 : \underline{x} = \underline{w}$$

$$H_1 : \underline{x} = \underline{s} + \underline{w}$$

The optimal detector is

$$\Lambda(\underline{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

where

$$\Lambda(\underline{x}) = \frac{(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{\frac{-1}{2\sigma^2}(\underline{x}-\underline{s})^T(\underline{x}-\underline{s})\right\}}{(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{\frac{-1}{2\sigma^2}\underline{x}^T\underline{x}\right\}}$$

$$= \exp\left\{\frac{-1}{2\sigma^2}\left[(\underline{x}-\underline{s})^T(\underline{x}-\underline{s}) - \underline{x}^T\underline{x}\right]\right\}$$

In terms of the log likelihood ratio, we have

$$\log \Lambda(\underline{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \log \eta$$

where

$$\log \Lambda(\underline{x}) = \frac{-1}{2\sigma^2}\left[(\underline{x}-\underline{s})^T(\underline{x}-\underline{s}) - \underline{x}^T\underline{x}\right]$$

$$= \frac{1}{\sigma^2}\left[\underline{s}^T\underline{x} - \frac{\underline{s}^T\underline{s}}{2}\right]$$

Rearranging terms, we have

$$\underline{s}^T\underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2 \log \eta + \frac{\underline{s}^T\underline{s}}{2} \equiv \gamma$$

Remarks

1. If we rewrite the detection problem

$$H_0 : \underline{X} \sim \mathcal{N}(\theta \underline{s}, \sigma^2 \underline{\underline{I}}), \quad \theta = 0$$

$$H_1 : \underline{X} \sim \mathcal{N}(\theta \underline{s}, \sigma^2 \underline{\underline{I}}), \quad \theta = 1$$

Then $\underline{s}^T \underline{x}$ is a <u>sufficient statistic</u> for θ.

2. $\underline{s}^T \underline{s} = \sum_{n=0}^{N-1} s(n)^2 =$ <u>signal energy</u>

3. The operator

$$\underline{x} \longmapsto \underline{s}^T \underline{x} = \sum_{n=0}^{N-1} x(n)\, s(n)$$

is called a <u>correlator</u>.

Projection Interpretation

We may rewrite the LRT as

$$\frac{\underline{s}^T \underline{x}}{\underline{s}^T \underline{s}} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma^2 \log \eta}{\underline{s}^T \underline{s}} + \frac{1}{2} = \gamma'$$

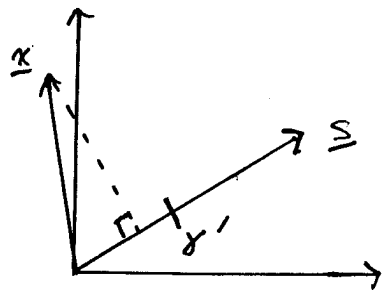
Let $S = \langle \underline{s} \rangle$. Then

$$P_S(\underline{x}) = \underline{s}(\underline{s}^T \underline{s})^{-1} \underline{s}^T \underline{x}$$

$$= \left(\frac{\underline{s}^T \underline{x}}{\underline{s}^T \underline{s}} \right) \underline{s}$$

Coefficient given by pseudo-inverse $(\underline{s}^T \underline{s})^{-1} \underline{s}^T \underline{x}$

So the LR detector depends only on the projection of the data onto the signal subspace. Other components of $\underline{x}$ are "filtered out" and do not factor into the decision.



DSP Interpretation

$$\underline{s}^T \underline{x} = \sum_{k=0}^{N-1} x(k)s(k)$$

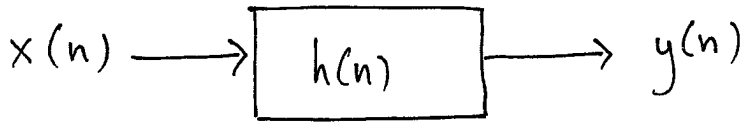
$$= \sum_{k=0}^{N-1} x(k)h(n-k) \Big|_{n=N-1}$$

sample at time $n=N-1$

where $h(k) = s(N-1-k)$

Expressed as a convolution,

$$\underline{s}^T \underline{x} = x(n) * h(n) \Big|_{n=N-1}$$

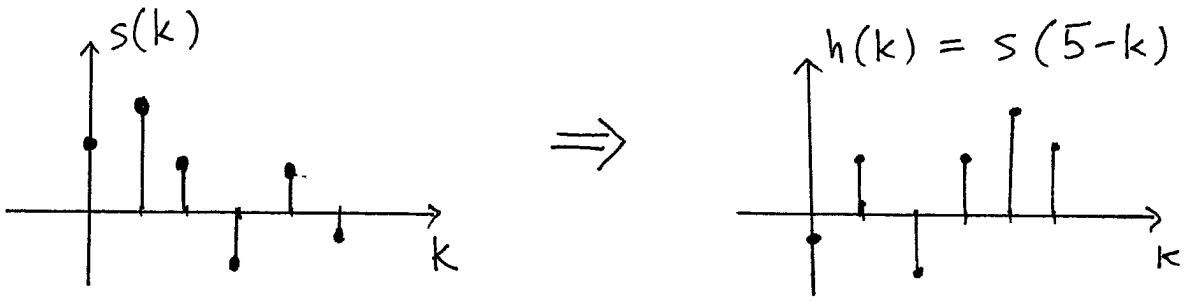


FIR filter

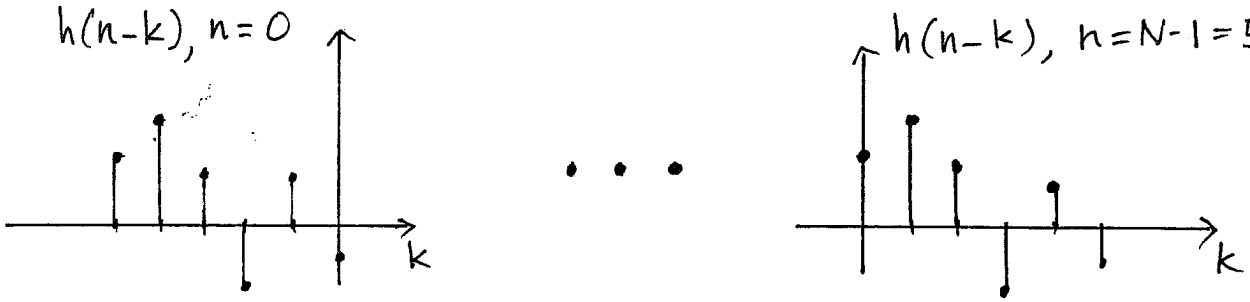
$$y(N-1) = \underline{s}^T \underline{x}$$

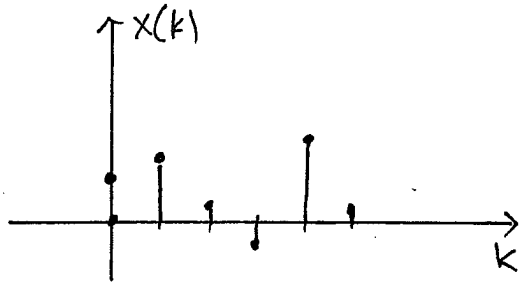
In this context, the detector is called a matched filter, because we pass the data $\underline{x}$ through a filter whose impulse response "matches" the signal $\underline{s}$.

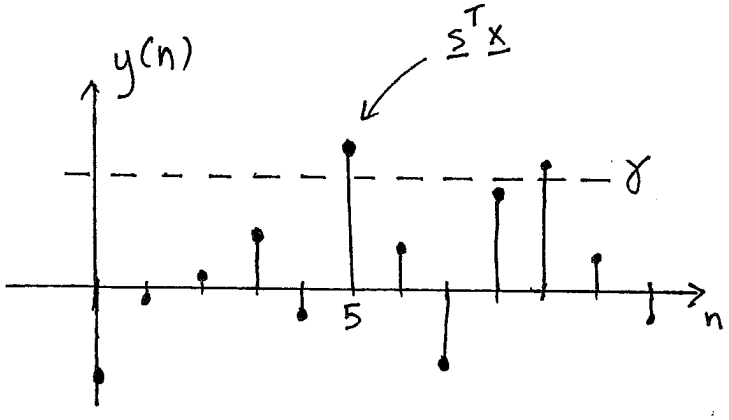
Picture: $N = 6$



Now to convolve with $h(k)$, we flip $h(k)$

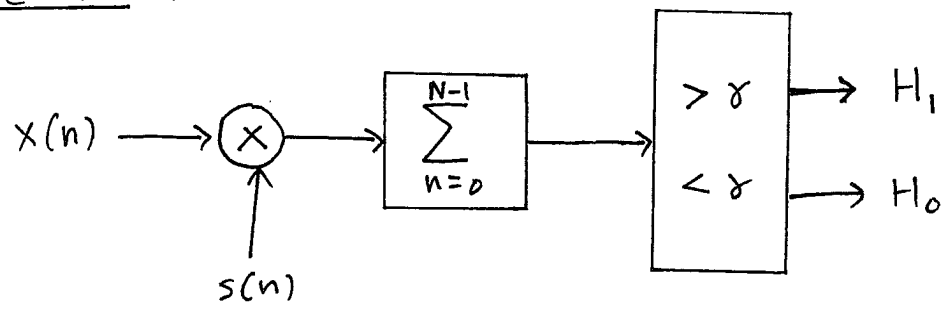




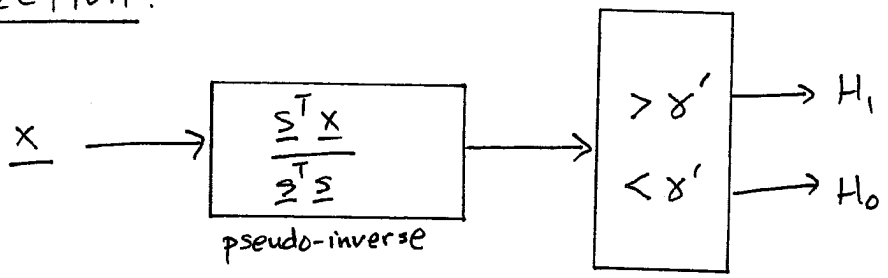


Block Diagrams

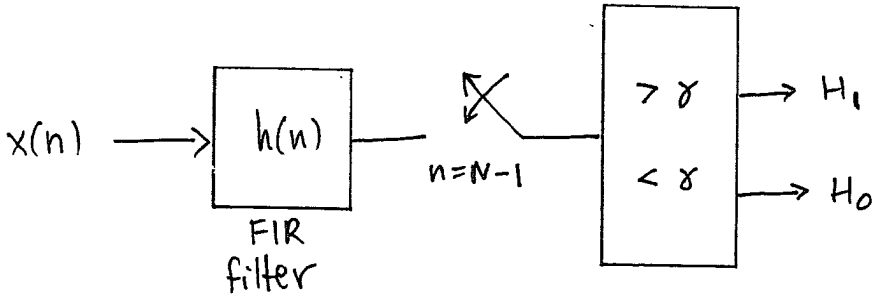
Correlator:



Projection:



Matched Filter:



Special case: DC signal

$$H_0: \quad x(n) = w(n), \quad n = 0, 1, \ldots, N-1$$

$$H_1: \quad x(n) = A + w(n), \quad n = 0, 1, \ldots, N-1$$

$$w(n) \sim \mathcal{N}(0, \sigma^2)$$

$$\underline{s} = [A \; A \; \cdots \; A]^T$$

Optimal detector:

$$\underline{s}^T \underline{x} = A \sum_{n=0}^{N-1} x(n)$$

$$\gamma = \sigma^2 \log \eta + \frac{\underline{s}^T \underline{s}}{2} = \sigma^2 \log \eta + \frac{NA^2}{2}$$

$$\Rightarrow \quad \frac{1}{N} \sum_{n=0}^{N-1} x(n) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma^2}{NA} \log \eta + \frac{A}{2}$$

Just what we derived earlier.

Let's consider the slightly more general problem

$$H_0: \underline{x} = \underline{s}_0 + \underline{w}$$

$$H_1: \underline{x} = \underline{s}_1 + \underline{w}$$

$$\underline{w} \sim \mathcal{N}(\underline{0}, \sigma^2 \underline{\underline{I}})$$

Now the transmitted signal is <u>nonzero</u> under <u>both</u> hypotheses.

<u>Log-likelihood ratio</u>:

$$\log \Lambda(\underline{x}) = \frac{-1}{2\sigma^2}\left[(\underline{x}-\underline{s}_1)^T(\underline{x}-\underline{s}_1) - (\underline{x}-\underline{s}_0)^T(\underline{x}-\underline{s}_0)\right]$$

$$= \frac{1}{\sigma^2}\left[(\underline{s}_1-\underline{s}_0)^T\underline{x} - \frac{\underline{s}_1^T\underline{s}_1}{2} + \frac{\underline{s}_0^T\underline{s}_0}{2}\right]$$

so the optimal test is:

$$(\underline{s}_1 - \underline{s}_0)^T\underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2 \log(\eta) + \frac{\underline{s}_1^T\underline{s}_1}{2} - \frac{\underline{s}_0^T\underline{s}_0}{2} \equiv \gamma$$

Projection Interpretation

Consider a minimum probability of error detector. Assume

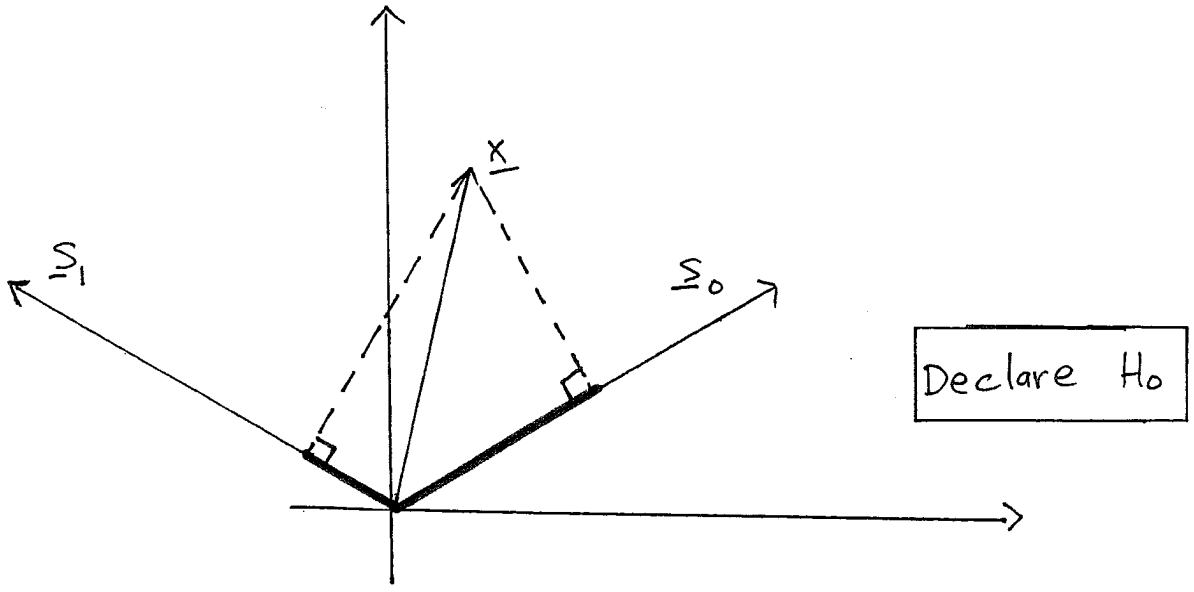
- $\pi_0 = \pi_1 = \frac{1}{2}$ $(\eta = 1)$
- $\|s_0\|^2 = \|s_1\|^2$

Then the detector reduces to

(a)

$$\underset{H_0}{\overset{H_1}{\gtrless}}$$

Recall $s_i^T x \propto$ coefficient of x projected onto s_i.



Exercise | Performance analysis.

(a) Calculate P_E, P_F, P_M as functions of γ

(b) Determine γ in terms of desired false alarm rate α.

(c) Express P_D as function of P_F. (d) Identify SNR.

Solution Recall $\underline{X} \sim N(\underline{s}_i, \sigma^2 I)$ under H_i.

Under H_0,

$$(\underline{s}_1 - \underline{s}_0)^T \underline{X} \sim N\left((\underline{s}_1 - \underline{s}_0)^T \underline{s}_0, \sigma^2 \|\underline{s}_1 - \underline{s}_0\|^2\right)$$

Under H_1

$$(\underline{s}_1 - \underline{s}_0)^T \underline{X} \sim N\left((\underline{s}_1 - \underline{s}_0)^T \underline{s}_1, \sigma^2 \|\underline{s}_1 - \underline{s}_0\|^2\right)$$

(a) $$P_F = P\left((\underline{s}_1 - \underline{s}_0)^T \underline{X} > \gamma \mid H_0\right)$$

$$= Q\left(\frac{\gamma - (\underline{s}_1 - \underline{s}_0)^T \underline{s}_0}{\sigma \|\underline{s}_1 - \underline{s}_0\|}\right)$$

$$P_M = P\left((\underline{s}_1 - \underline{s}_0)^T \underline{X} < \gamma \mid H_1\right)$$

$$= 1 - Q\left(\frac{\gamma - (\underline{s}_1 - \underline{s}_0)^T \underline{s}_1}{\sigma \|\underline{s}_1 - \underline{s}_0\|}\right)$$

$$P_E = \pi_0 P_F + \pi_1 P_M, \quad \eta = \frac{\pi_0}{\pi_1}$$

(b) $$\gamma = \sigma \|\underline{s}_1 - \underline{s}_0\| \cdot Q^{-1}(\alpha) + (\underline{s}_1 - \underline{s}_0)^T \underline{s}_0$$

(c) $$P_D = P\left((\underline{s}_1 - \underline{s}_0)^T \underline{X} > \gamma \mid H_1\right)$$

$$= Q\left(\frac{\gamma - (\underline{s}_1 - \underline{s}_0)^T \underline{s}_1}{\sigma \|\underline{s}_1 - \underline{s}_0\|}\right)$$

$$= Q\left(\frac{\sigma \|\underline{s}_1 - \underline{s}_0\| \cdot Q^{-1}(\alpha) + (\underline{s}_1 - \underline{s}_0)^T \underline{s}_0 - (\underline{s}_1 - \underline{s}_0)^T \underline{s}_1}{\sigma \|\underline{s}_1 - \underline{s}_0\|}\right)$$

$$= Q\left(Q^{-1}(\alpha) - \frac{\|\underline{s}_0 - \underline{s}_1\|}{\sigma} \right)$$

(d) $SNR = \dfrac{\|\underline{s}_0 - \underline{s}_1\|^2}{\sigma^2}$

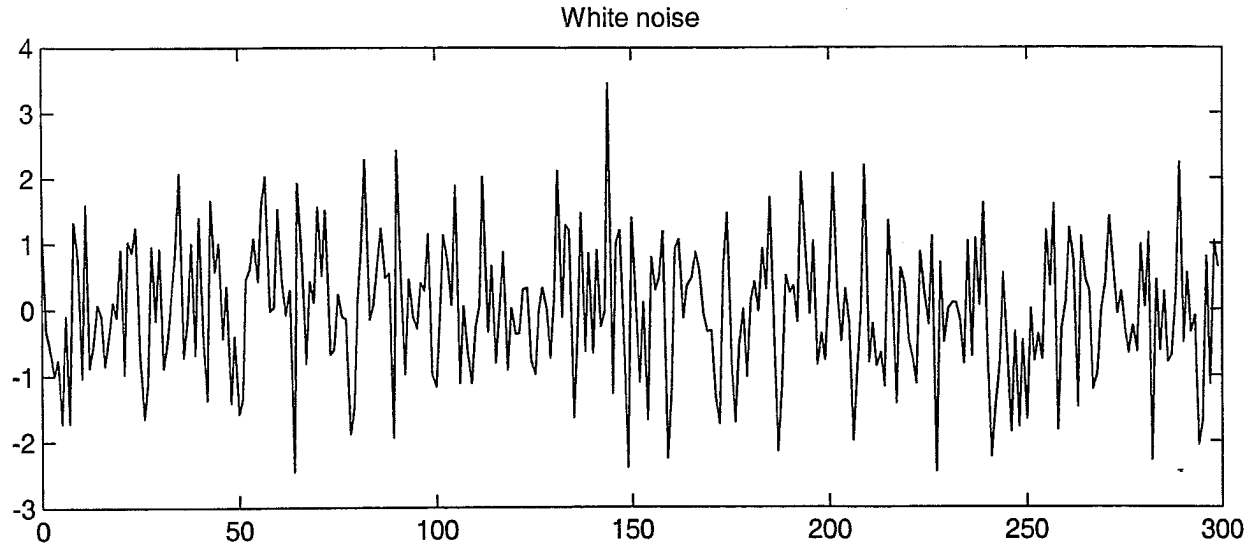
The Gaussian Assumption

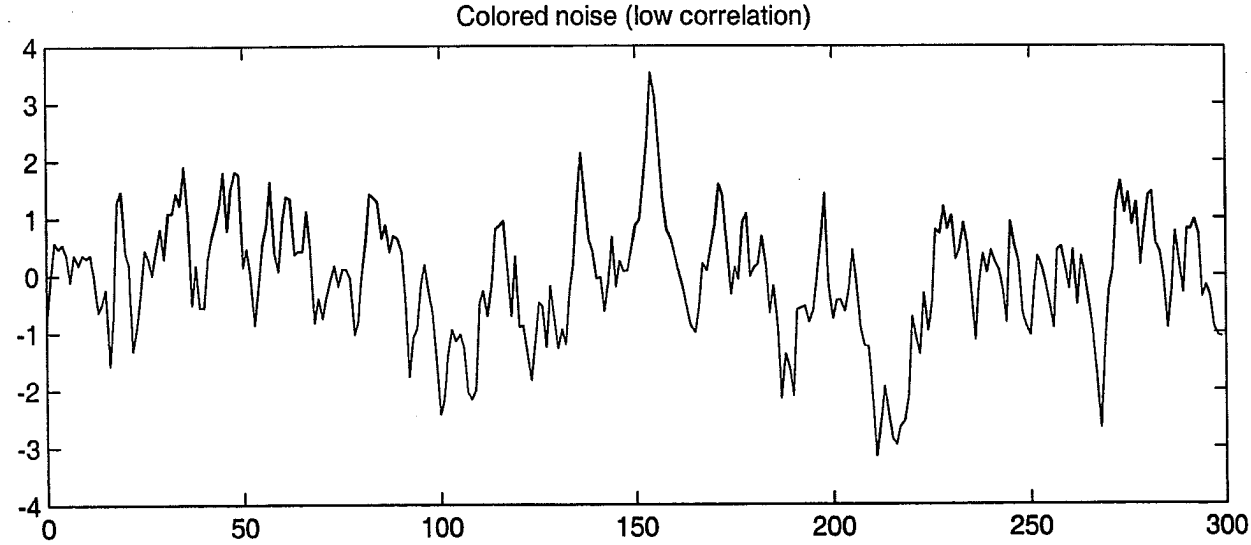
Is real world noise Gaussian? Is it white?

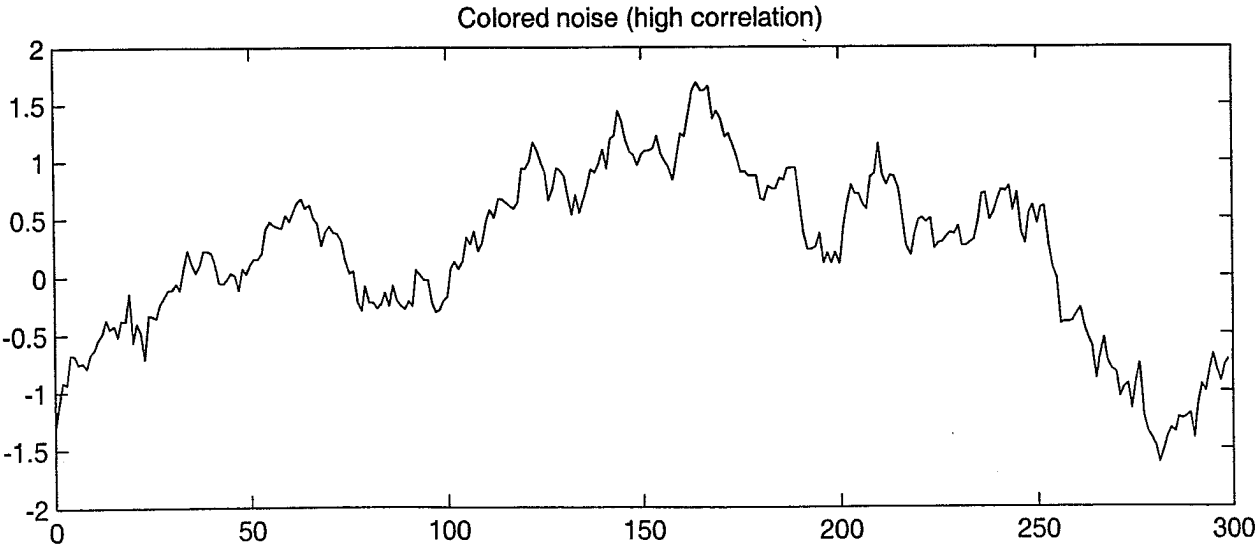
If the noise arises from a large number of random events, the central limit theorem suggests that the noise will be Gaussian.

Example | In communication systems, electronic noise is due to the aggregate effect of huge numbers of charge carriers undergoing random motion.

However, the "whiteness" assumption is often violated. That is, errors tend to be correlated from sample to sample.


White noise
4
3
2
1
0
-1
-2
-3
0
50
100
150
200
250
300


Colored noise (low correlation)
4
3
2
1
0
-1
-2
-3
-4
0
50
100
150
200
250
300


Colored noise (high correlation)
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
50
100
150
200
250
300

```
clear all
close all

% White versus colored Gaussian noise

N = 300;

% white noise
w = randn(N,1);

subplot(3,1,1);
plot(0:N-1,w);
title('White noise')

% colored noise (low correlation)
q=.8;
r=q.^(0:N-1);
R = Toeplitz(r); % covariance matrix

[U,D]=eig(R);
w=U*sqrt(D)*randn(N,1); % randn generates IID samples.
                        % this correlates the samples
subplot(3,1,2)
plot(0:N-1,w);
title('Colored noise (low correlation)');

% colored noise (high correlation)
q=.99;
r=q.^(0:N-1);

R = Toeplitz(r); % covariance matrix

[U,D]=eig(R);
w=U*sqrt(D)*randn(N,1); % randn generates IID samples.
                        % this correlates the samples
subplot(3,1,3)
plot(0:N-1,w);
title('Colored noise (high correlation)');

orient tall
```

Colored Gaussian Noise

$$H_0: \underline{x} = \underline{s}_0 + \underline{w}$$

$$H_1: \underline{x} = \underline{s}_1 + \underline{w}$$

$$\underline{W} \sim \mathcal{N}(\underline{0}, R)$$

$$f(\underline{w}) = \frac{1}{(2\pi)^{\frac{N}{2}} |R|^{\frac{1}{2}}} e^{-\frac{1}{2}\underline{w}^T R^{-1} \underline{w}}$$

Log LRT

$$(\underline{x}-\underline{s}_1)^T R^{-1}(\underline{x}-\underline{s}_1) - (\underline{x}-\underline{s}_0)^T R^{-1}(\underline{x}-\underline{s}_0) \underset{H_0}{\overset{H_1}{\gtrless}} 2\log\eta$$

or

$$(\underline{s}_1-\underline{s}_0)^T R^{-1} \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \log\eta + \frac{\underline{s}_1^T R^{-1}\underline{s}_1}{2} - \frac{\underline{s}_0^T R^{-1}\underline{s}_0}{2} \equiv \gamma$$

↑ sufficient statistic: t

Performance :

$$H_0: \quad t \sim \mathcal{N}\left((\underline{s}_1 - \underline{s}_0)^T R^{-1} \underline{s}_0 \,,\; (\underline{s}_1 - \underline{s}_0)^T R^{-1} (\underline{s}_1 - \underline{s}_0) \right)$$

$$H_1: \quad t \sim \mathcal{N}\left((\underline{s}_1 - \underline{s}_0)^T R^{-1} \underline{s}_1 \,,\; (\underline{s}_1 - \underline{s}_0)^T R^{-1} (\underline{s}_1 - \underline{s}_0) \right)$$

$$P_F = Q\left(\frac{\gamma - (\underline{s}_1 - \underline{s}_0)^T R^{-1} \underline{s}_0}{\left[(\underline{s}_1 - \underline{s}_0)^T R^{-1} (\underline{s}_1 - \underline{s}_0) \right]^{1/2}} \right)$$

$$P_D = Q\left(\frac{\gamma - (\underline{s}_1 - \underline{s}_0)^T R^{-1} \underline{s}_1}{\left[(\underline{s}_1 - \underline{s}_0)^T R^{-1} (\underline{s} - \underline{s}_0) \right]^{1/2}} \right)$$

$$= Q\left(Q^{-1}(P_F) - \sqrt{SNR} \right)$$

where $SNR \equiv (\underline{s}_1 - \underline{s}_0)^T R^{-1} (\underline{s}_1 - \underline{s}_0)$

All of the detectors studied so far (today) can be deduced as special cases of this general form.

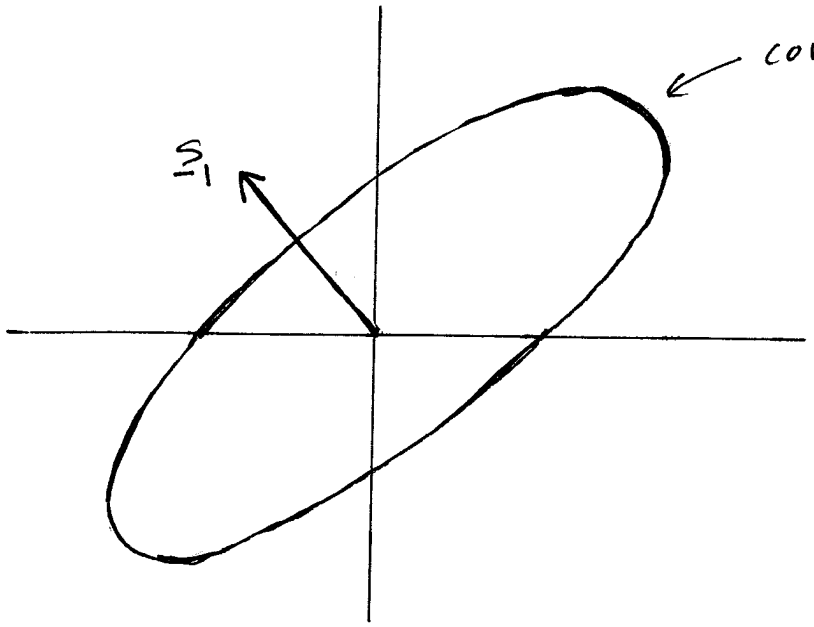
Signal Design

Suppose $\underline{s}_0 = \underline{0}$. How should we choose $\underline{s}_1$ to maximize SNR?

$$\underline{s}_1 = \arg\max_{\underline{s}} \frac{\underline{s}^T R^{-1} \underline{s}}{\underline{s}^T \underline{s}}$$

← Rayleigh quotient

$\Rightarrow$ $\underline{s}_1$ = eigenvector associated to smallest eigenvalue of R.



Prewhitening

Suppose we have colored noise

$$\underline{w} \sim \mathcal{N}(\underline{0}, R) \quad , \quad R \text{ known}$$

Since R is symmetric, we can write

$$R = U \Lambda U^T$$

where $UU^T = U^T U = I_{N \times N}$ and

$$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

Since R is positive definite, $\lambda_i > 0 \quad \forall i$.

Then we can write

$$\tilde{U} R \tilde{U}^T = I$$

where

$$\tilde{U} = \Lambda^{-\frac{1}{2}} U^T \quad , \quad \Lambda^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\lambda_N}} \end{pmatrix}$$

Suppose we modify our observation according to

$$\tilde{\underline{x}} = \tilde{\mathcal{U}}\,\underline{x}$$

If $\underline{X} \sim \mathcal{N}(\underline{s}_i, R)$ under H_i,

then $\tilde{\underline{X}} \sim \mathcal{N}(\tilde{\underline{s}}_i, I)$, were $\tilde{\underline{s}}_i = \tilde{\mathcal{U}}\,\underline{s}_i$

In other words, if we know the covariance matrix R, we can reduce the problem of detecting a signal in <u>colored noise</u> to the problem of detecting a signal in <u>IID</u> noise

This process is called <u>prewhitening</u>, and the transformation $\tilde{\mathcal{U}}$ is called the <u>whitening filter.</u>

So we can reduce the colored noise problem to

$$H_0: \underline{x} = \underline{s}_0 + \underline{w}$$
$$H_1: \underline{x} = \underline{s}_1 + \underline{w}, \quad \underline{w} \sim \mathcal{N}(\underline{0}, I).$$

Can we make any more reductions?

Define $\underline{\tilde{x}} = \underline{x} - \underline{s}_0$. Then the problem reduces to

$$H_0: \underline{\tilde{x}} = \underline{w}$$
$$H_1: \underline{\tilde{x}} = \underline{\tilde{s}} + \underline{w}$$

where $\underline{\tilde{s}} = \underline{s}_1 - \underline{s}_0$. This was the first problem we considered.

Conclusion Any binary detection problem involving Gaussian noise can be reduced to a "signal present vs. signal absent" problem in white Gaussian noise.

Multiple Hypotheses

Consider the problem of detecting M hypotheses in additive Gaussian noise.

$$H_1: \underline{x} = \underline{s}_1 + \underline{w}$$

$$H_2: \underline{x} = \underline{s}_2 + \underline{w}$$

$$\vdots$$

$$H_M: \underline{x} = \underline{s}_M + \underline{w}$$

Assume data is prewhitened

where $\underline{w} \sim N(\underline{0}, \underline{I})$

Recall the MAP detector:

> Choose H_i such that $\pi_i f_i(\underline{x})$ is maximal

For simplicity, assume $\pi_k = \frac{1}{M}$, $k = 1, 2, \ldots, M$.

Then the minimum error probability is achieved by the maximum likelihood detector:

> Choose H_i such that $f_i(\underline{x})$ is maximal

Now

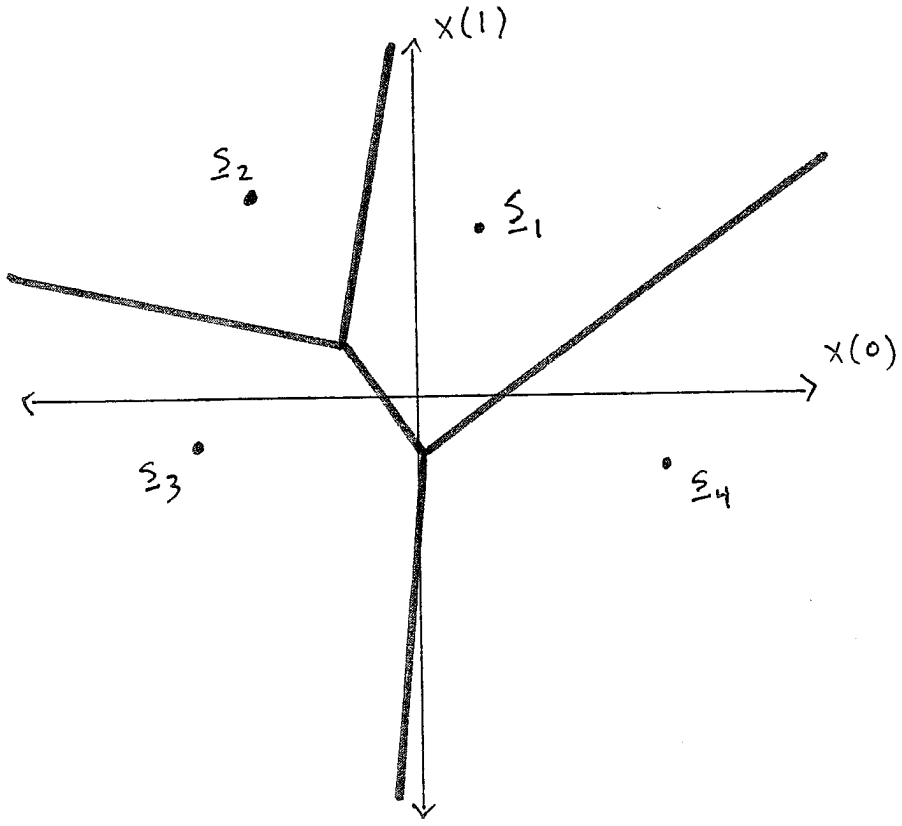
$$f_i(\underline{x}) = \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{s}_i)^T(\underline{x}-\underline{s}_i)}$$

So maximizing $f_i(\underline{x})$ is equivalent to minimizing

$$(\underline{x}-\underline{s}_i)^T(\underline{x}-\underline{s}_i) = \|\underline{x}-\underline{s}_i\|^2$$

and the optimal detector reduces to a nearest-neighbor detector:

> Choose H_i such that $\|\underline{x}-\underline{s}_i\|^2$ is minimal

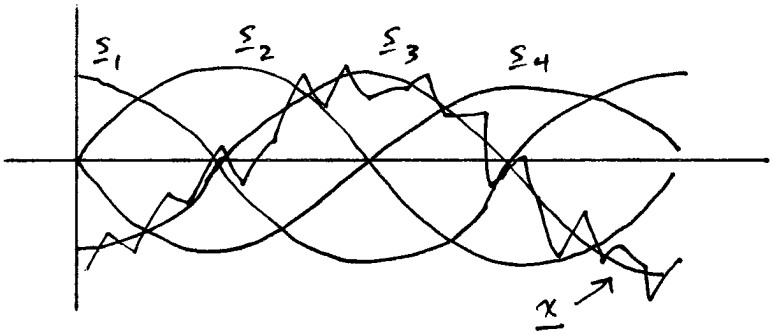


N=2

Equivalenty, this may be thought of as a maximum correlation detector:

> Choose H_i such that $\underline{s}_i^T \underline{x}$ is maximal

provided all $\underline{s}_i$ have the same energy.



$\Rightarrow$ declare H_3

Summary

- Detection in Gaussian noise $\Rightarrow$ many intuitive rules and interpretations:
  - max correlation
  - matched filter
  - projection
  - nearest neighbor
- Prewhitening reduces colored noise problem to white noise problem.

Key

a. $\underline{s}_1^T \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \underline{s}_0^T \underline{x}$

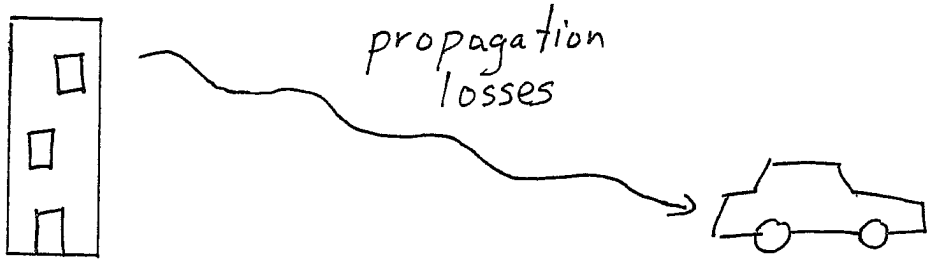
UMP TESTS & THE KARLIN-RUBIN THEOREM

Signal Detection in the Presence of Unknowns

In many real world detection problems, the characteristics of the signal and/or noise are not perfectly known.

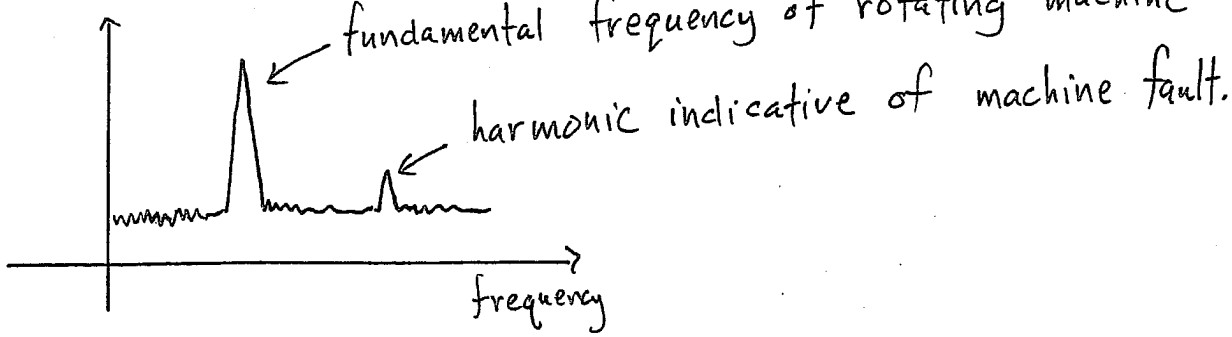
Ex 1 | Unknown signal amplitude

(a) Wireless comm:



received signal attenuated by unknown factor

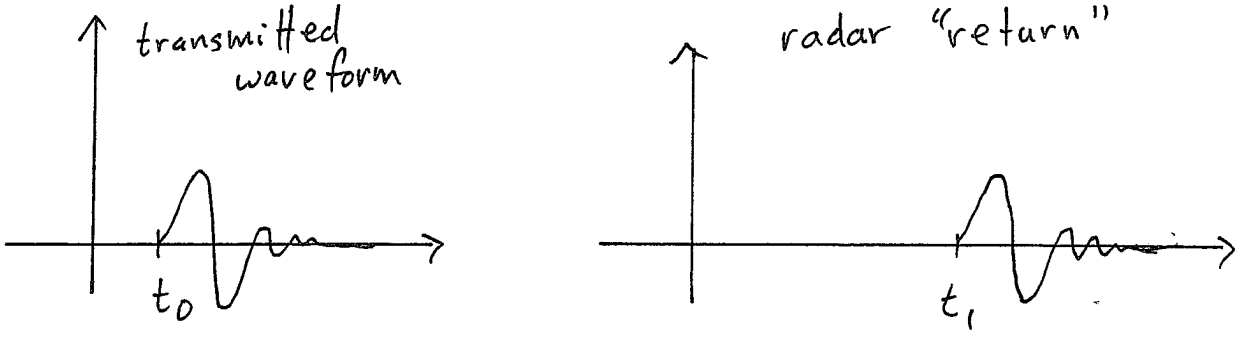
(b) machine fault detection



strength of harmonic distortion is uncertain

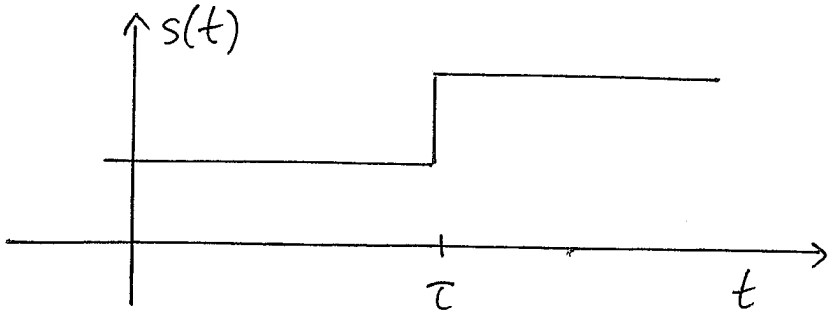
Ex 2] Unknown location / delay

(a) Radar:



$d = t_1 - t_0$ is unknown

(b) Step-change detection:



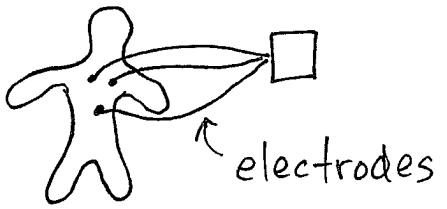
τ unknown

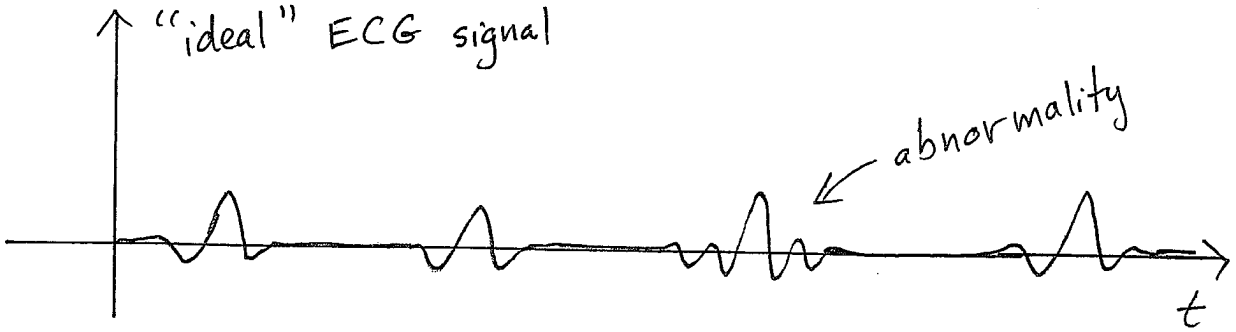
In 2-d, this amounts to <u>edge-detection</u>:
where do edges occur?

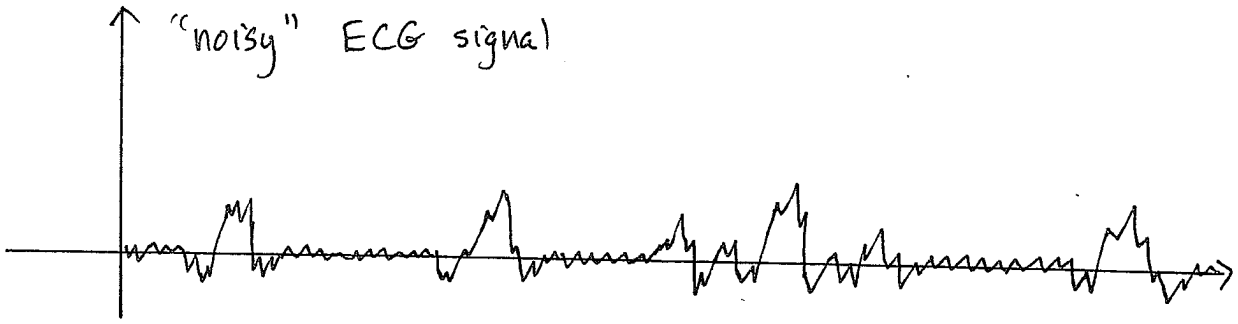
Ex 3| Unknown noise power

We may know noise is white, but what is σ^2?

Heart monitoring:







Goal: detect abnormal heartbeat.

Noise level depends on patient, electrode placement, and ambient noise from environment, all of which may be unknown.

Modeling Uncertainty

① Parametric uncertainty: unknown parameters in pdf or pmf of observation.

② Nonparametric uncertainty: we don't even know the functional form of the pdf or pmf

Non parametric uncertainty is much more challenging. We will focus on parametric uncertainty in this course.

Composite Hypothesis Testing

$$H_0 : \underline{X} \sim f(\underline{x}\,;\,\underline{\theta}_0), \qquad \underline{\theta}_0 \in \Theta_0$$

$$H_1 : \underline{X} \sim f(\underline{x}\,;\,\underline{\theta}_1), \qquad \underline{\theta}_1 \in \Theta_1$$

So far, we have only considered <u>simple</u> hypotheses: $\Theta_0 = \{\underline{\theta}_0\}$, $\Theta_1 = \{\underline{\theta}_1\}$. When $|\Theta_k| > 1$, H_k is a <u>composite</u> hypothesis.

<u>Ex 1</u> Unknown mean:

$$H_0: \quad X \sim \mathcal{N}(0,1)$$

$$H_1: \quad X \sim \mathcal{N}(\mu,1), \qquad \mu > 0$$

$\Rightarrow H_1$ is composite

<u>Ex 2</u> Unknown noise power

$$H_0: \quad x(n) = s_0(n) + w(n)$$

$$H_1: \quad x(n) = s_1(n) + w(n)$$

$w(n) \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$, σ^2 unknown $\Rightarrow H_0, H_1$ composite

Whether we want a Bayes Risk or Neyman-Pearson detector, the optimal decision rule is the LRT:

$$\Lambda(\underline{x}) = \frac{f(\underline{x};\underline{\theta}_1)}{f(\underline{x};\underline{\theta}_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

In this form, the LRT requires knowledge of the unknown parameter, and hence is not useful.

Sometimes, however, if we write the test in a different form, the dependence on the unknown parameter goes away.

Ex | Signal Detection in AWGN, σ^2 unknown:

$$H_0 : \underline{x} = \underline{w} \qquad w(n) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$H_1 : \underline{x} = \underline{s} + \underline{w} \qquad \sigma^2 \text{ unknown}$$

LRT reduces to

$$\underline{s}^T \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2 \log(\eta) + \frac{\underline{s}^T \underline{s}}{2}$$

If the hypotheses are equally probable a priori, then $\eta = 1$ and we have

$$\underline{s}^T \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\underline{s}^T \underline{s}}{2}$$

← independent of σ^2

For $M > 2$ hypotheses

$$H_k : \underline{x} = \underline{s}_k + \underline{w}$$

$$\Rightarrow \text{ minimize } \frac{1}{\sigma^2} \| \underline{x} - \underline{s}_k \|^2$$

This situation is rare in practice.

A more common occurance is when the LRT can be reduced to a test statistic that does not depend on the unknown parameter under H_0. This allows us to set the false alarm rate.

Ex Unknown signal amplitude

$$H_0 : x(n) = w(n)$$
$$H_1 : x(n) = As(n) + w(n)$$

$w(n) \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$, σ^2 known, A unknown

The LRT reduces to

$$\underbrace{A \underline{s}^T \underline{x}}_{\text{test statistic}} \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2 \log(\eta) + A^2 \frac{\underline{s}^T \underline{s}}{2}$$

↳ test statistic depends on unknown amplitude $\Rightarrow$ don't know distribution under H_0 $\Rightarrow$ can't set P_F

What if we know $A > 0$? Then, dividing by A, we have

$$\underline{s}^T \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma^2}{A} \log(\eta) + A \frac{\underline{s}^T \underline{s}}{2} \equiv \gamma$$

Under H_0

$$\underline{s}^T \underline{x} \sim$$

← independent of A!

(a)

We can now set the threshold γ to achieve a certain P_F:

$$P_F =$$

$$\Longrightarrow \gamma =$$

So we can set γ to achieve P_F. So what? Is this detector optimal in any sense? Is P_D maximized?

Since the LRT is equivalent to

$$\underline{s}^T \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

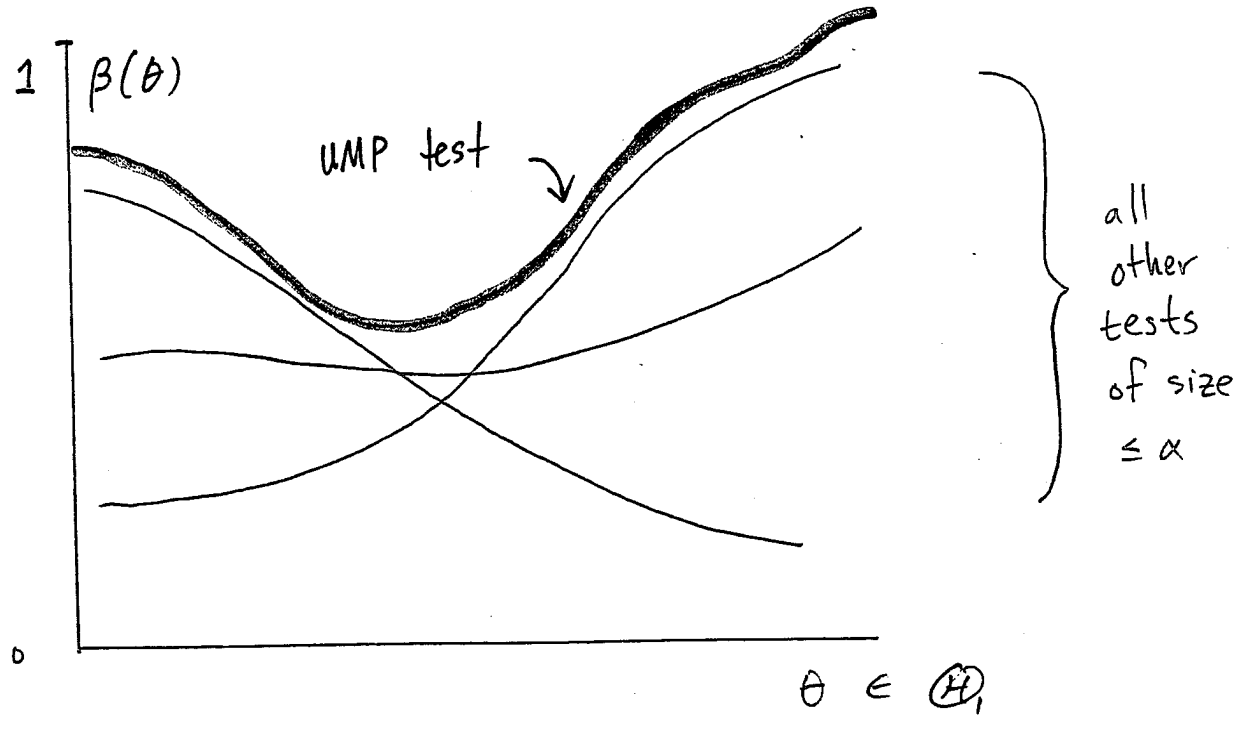
and γ can be selected to ensure $P_F = \alpha$, the NP lemma implies that this test is optimal regardless of the value of A!

UMP Tests

A test/detector is a uniformly most powerful (UMP) test of size α if it has the largest power

$$\beta(\underline{\theta}) := P(\text{declare } H_1 ; \underline{\theta}), \quad \underline{\theta} \in \Theta_1$$

among all tests of size $\leq \alpha$, $\forall \underline{\theta} \in \Theta_1$



Unfortunately, UMP tests rarely exist. However, there is a certain class of problems for which they do.

Monotone Likelihood Ratios

Suppose a measurement $\underline{x}$ has pdf/pmf determined by a scalar parameter θ and let θ_0 be fixed. Suppose we are interested in the one-sided problem

$$H_0: \theta = \theta_0$$
$$H_1: \theta > \theta_0$$

Further suppose that t is a scalar sufficient statistic for θ.

Observe that for any $\theta_1 > \theta_0$, the likelihood ratio is a function of t:

$$\Lambda(\underline{x}) = \frac{f(\underline{x}; \theta_1)}{f(\underline{x}; \theta_0)} = \frac{a(\underline{x})\, b_{\theta_1}(t)}{a(\underline{x})\, b_{\theta_0}(t)} = \frac{b_{\theta_1}(t)}{b_{\theta_0}(t)} = \tilde{\Lambda}(t)$$

by the Fisher-Neyman factorization.

Proposition If $\tilde{\Lambda}(t)$ is monotone increasing for all $\theta_1 > \theta_0$, then

$$t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

is UMP of size α, where α is determined by

$$P(T > \gamma \,;\, \theta = \theta_0) = \alpha.$$

A similar result holds if $\tilde{\Lambda}(t)$ is monotone decreasing for all $\theta_1 > \theta_0$.

Proof Suppose $\theta = \theta_1 > \theta_0$. We need to show that

$$P(T > \gamma \,;\, \theta = \theta_1)$$

is as large as possible among all tests with size α. By the NP Lemma, the most powerful test for $\theta = \theta_1$ vs. $\theta = \theta_0$ is

$$\tilde{\Lambda}(t) \gtrless \eta$$

where η is such that

$$P(\tilde{\Lambda}(T) > \eta \,;\, \theta = \theta_0) = \alpha.$$

Since $\tilde{\Lambda}(t)$ is monotone increasing, so is its inverse, and the most powerful test simplifies to

$$t \underset{H_0}{\overset{H_1}{\gtrless}} \tilde{\Lambda}^{-1}(\eta) \equiv \gamma.$$

Since the distribution of $T; \theta = \theta_0$ is independent of θ_1, we can set the value of γ without knowledge. In other words, our test has greatest power for all $\theta_1 > \theta_0$. ∎

In short, the monotone LR property implies that we can eliminate θ_1 from the test statistic, and since H_0 is simple, we can set the threshold to ensure the desired size.

Remarks ① In the case of discrete data, the thresholding test may have the form

$$t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

where if $t = \gamma$ we flip a "ρ-coin" such that

$$P\{T > \gamma ; \theta_0\} + \rho P\{T = \gamma ; \theta_0\} = \alpha.$$

② If $\tilde{\Lambda}(t)$ is monotone <u>decreasing</u>, then the inequalities in the thresholding test are reversed.

The Karlin-Rubin Theorem

The preceeding result can be generalized to the case where Θ_0 is also composite.

Now suppose θ_0 is fixed and consider the one-sided problem

$$H_0: \theta \leq \theta_0$$
$$H_1: \theta > \theta_0$$

To state the general result, we need to generalize our notion of "size."

Let $\phi(\underline{x}) \in \{0,1\}$ denote an arbitrary test. We define the <u>size</u> of ϕ by

$$\text{size}(\phi) = \sup_{\underline{\theta} \in \Theta_0} P\{\phi(\underline{X}) = 1 \mid \underline{\theta}\}$$
$$= \sup_{\underline{\theta} \in \Theta_0} E_{\underline{\theta}}\{\phi(\underline{X})\}.$$

This is essentially the maximum false alarm rate over all possible null hypotheses.

Theorem | Suppose t is a scalar suff. stat. for θ and that

$$\tilde{\Lambda}_{\theta_1,\theta_0}(t) = \frac{b_{\theta_1}(t)}{b_{\theta_0}(t)}$$

is monotone increasing for each $\theta_1 > \theta_0$.

Then a UMP test of size α is given by

$$\phi(t) = \begin{cases} 1 & \text{if } t > \gamma \\ \text{flip a } \rho\text{-coin} & \text{if } t = \gamma \\ 0 & \text{if } t < \gamma \end{cases}$$

where γ, ρ are chosen such that

$$P\{T > \gamma \mid \theta = \theta_0\} + \rho P\{T = \gamma \mid \theta = \theta_0\} = \alpha.$$

Remarks | (1) See Scharf, p 124, for a proof.

(2) Similar result holds if $\Lambda(t)$ is monotone decreasing.

(3) Similar result applies to the problem

$$H_0 : \theta_{min} \lesseqgtr \theta \leq \theta_0$$

$$H_1 : \theta_0 < \theta \lessgtr \theta_{max}$$

where $\theta_{min}, \theta_{max}$ are fixed and possibly unknown.

When is $\tilde{\lambda}(t)$ monotone increasing? When the LRT can be reduced to

$$t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

by a series of monotone increasing transformations.

Exercise Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, N$ and μ known, and consider testing

$$H_0 : 0 < \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 \geq \sigma_0^2$$

with σ_0^2 fixed, known. Find a UMP test.

Solution Let $\sigma_1^2 > \sigma_0^2$. Then

$$\Lambda(\underline{x}) = \frac{(2\pi\sigma_1^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}}{(2\pi\sigma_0^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}}$$

$$= \left(\frac{\sigma_0}{\sigma_1}\right)^N \exp\left\{\frac{1}{2}\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)\underbrace{\sum_{i=1}^{N}(x_i-\mu)^2}_{\vdots\,t}\right\}$$

Then

$$\Lambda(t) \gtrless \eta \iff \exp\left\{\frac{1}{2}\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)t\right\} \gtrless \eta\left(\frac{\sigma_1}{\sigma_0}\right)^N$$

$$\iff \left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)t \gtrless 2\log\left[\eta\left(\frac{\sigma_1}{\sigma_0}\right)^N\right]$$

$$\iff t \gtrless \frac{2\log\left[\eta\left(\frac{\sigma_1}{\sigma_0}\right)^N\right]}{\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\right)} =: \gamma$$

Since $\sigma_1^2 > \sigma_0^2$, all steps are monotone increasing $\Longrightarrow$ UMP test exists.

To set the threshold, recall

$$\frac{T}{\sigma_0^2} = \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma_0} \right)^2 \sim \chi_N^2 \qquad \text{if } \sigma^2 = \sigma_0^2$$

Define

$$Q_{\chi_N^2}(r) := P\{ \chi_N^2 > r \}.$$

Then the size of our test is

$$P\{ T > \gamma \; ; \sigma^2 = \sigma_0^2 \}$$

$$= P\left\{ \frac{T}{\sigma_0^2} > \frac{\gamma}{\sigma_0^2} \; ; \sigma^2 = \sigma_0^2 \right\}$$

$$= Q_{\chi_N^2}\left(\frac{\gamma}{\sigma_0^2} \right)$$

$$= \alpha$$

$$\Longrightarrow \quad \gamma = \sigma_0^2 \, Q_{\chi_N^2}^{-1}(\alpha).$$

Two-Sided Problems

What if we are interested in a slightly different problem?

$$H_0 : x(n) = w(n)$$

$$H_1 : x(n) = As(n) + w(n), \quad A \neq 0$$

That is

$$H_0 : A = 0$$

$$H_1 : A \neq 0$$

This is called a <u>two-sided</u> test, and UMPs <u>never</u> exist for such tests. We must be content with a suboptimal detector.

What can we do?

Consider the scalar case:

$$H_0: \; x \sim \mathcal{N}(0, \sigma^2) \qquad \sigma^2 \text{ known}$$

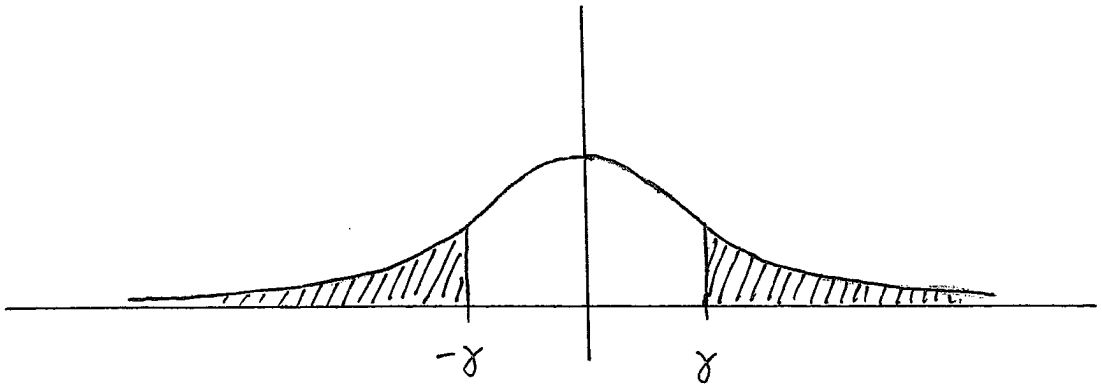
$$H_1: \; x \sim \mathcal{N}(A, \sigma^2) \qquad A \neq 0, \text{ unknown}$$

Intuitively, the decision rule

$$|x| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

comes to mind.

Large excursions of the observation x from 0 may indicate the signal is present.



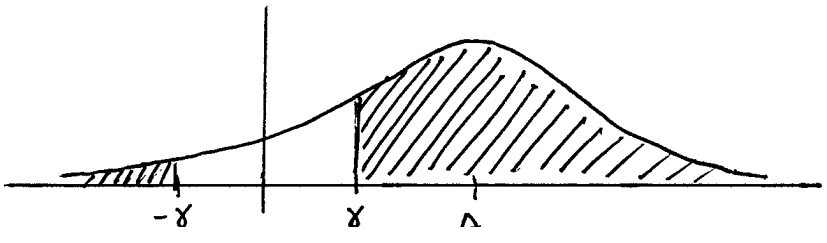
H_0 does not depend on A, so we may set the threshold γ by constraining P_F:

$$P_F = 2\int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx = 2Q\left(\frac{\gamma}{\sigma}\right)$$

$$\Rightarrow \quad \gamma = \sigma Q^{-1}\left(\frac{P_F}{2}\right)$$

In terms of A, the detection probability is

(b) $P_D =$



To evaluate our suboptimal detector, we can compare it to the clairvoyant detector, which assumes full knowledge of unknowns.

What is the clairvoyant detector for this problem (unknown amplitude)?

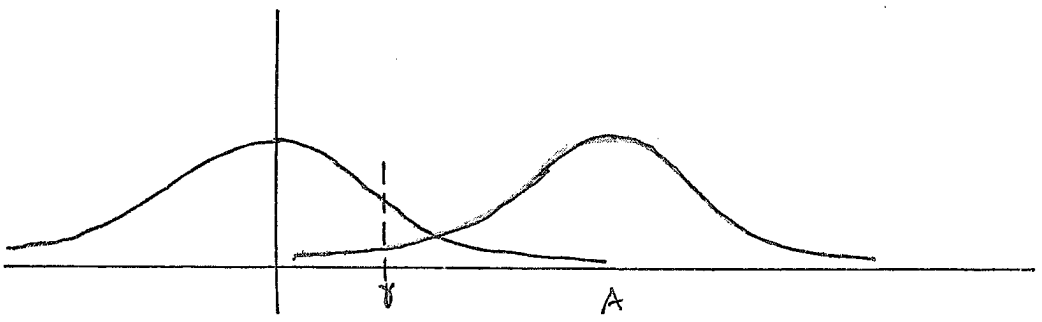
If $A > 0$,

$$x \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \equiv \frac{\sigma^2}{A} \log(\eta) + \frac{A}{2}$$

(c) $\Rightarrow P_F =$

$\gamma =$

$P_D =$



If $A < 0$

$$X \underset{H_1}{\overset{H_0}{\gtrless}} \frac{\sigma^2}{A}\log(\eta) + \frac{A}{2} \equiv \gamma$$

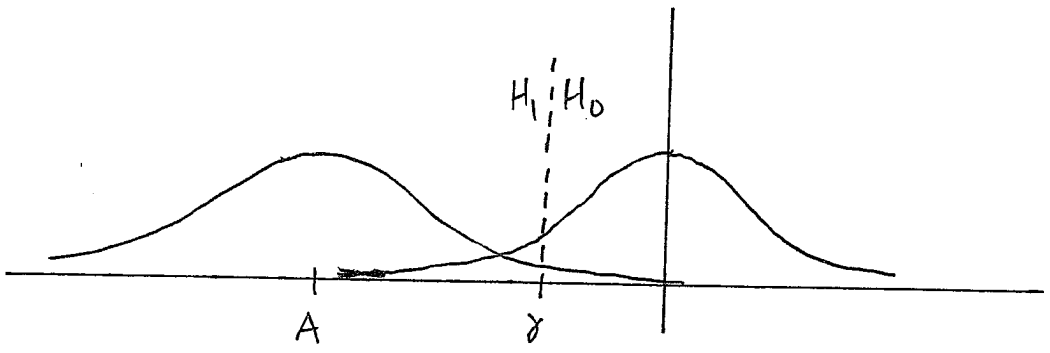
inequalities reversed

$$\Rightarrow \quad P_F = 1 - Q\left(\frac{\gamma}{\sigma}\right) = Q\left(\frac{-\gamma}{\sigma}\right)$$

$$\gamma = -\sigma Q^{-1}(P_F)$$

$$P_D = 1 - Q\left(\frac{\gamma - A}{\sigma}\right) = Q\left(\frac{A - \gamma}{\sigma}\right)$$

$$= Q\left(Q^{-1}(P_F) + \frac{A}{\sigma}\right)$$



Summary

$A > 0$: $P_D = Q\left(Q^{-1}(P_F) - \frac{A}{\sigma}\right)$

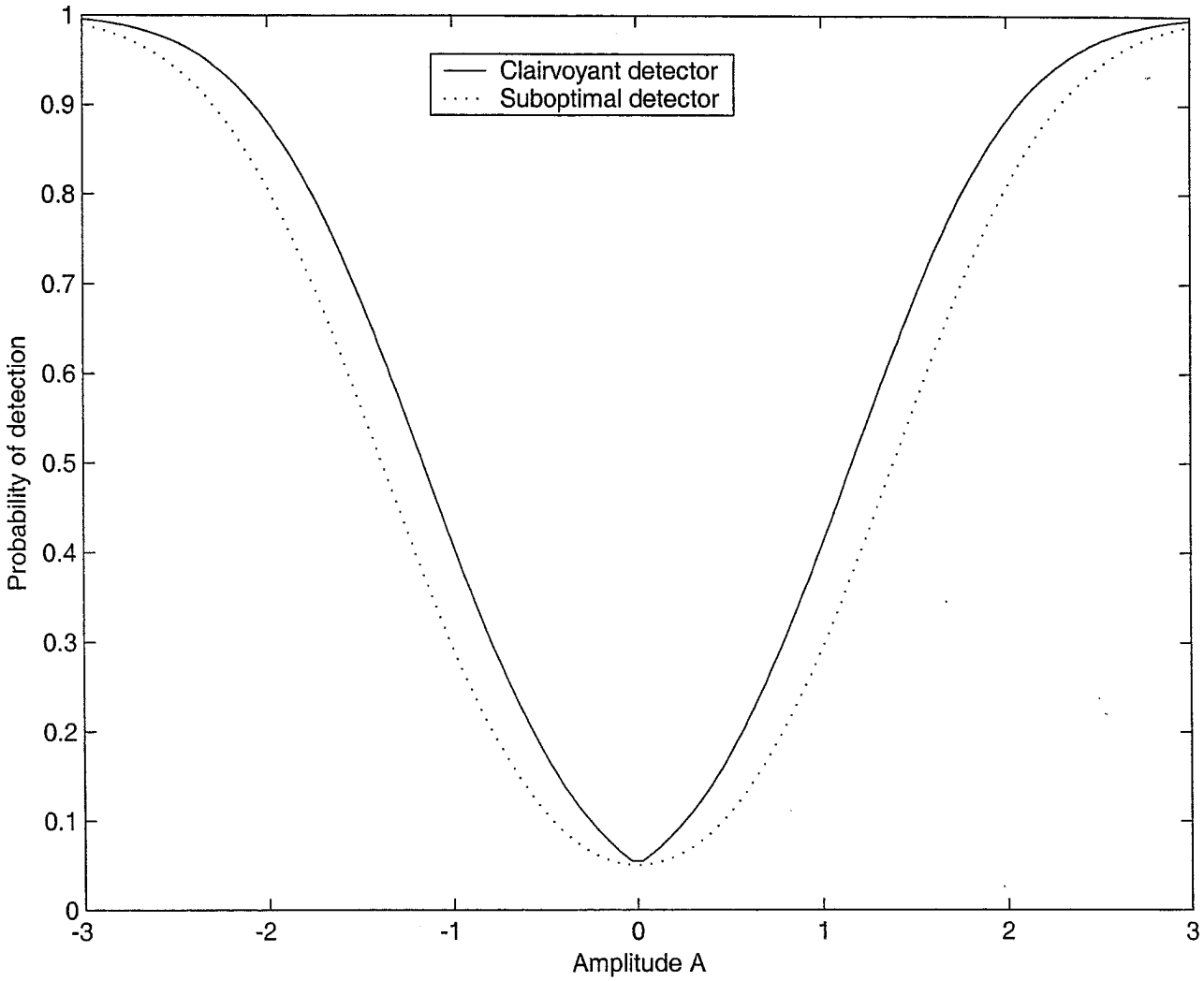
$A < 0$: $P_D = Q\left(Q^{-1}(P_F) + \frac{A}{\sigma}\right)$

We can combine these to obtain

$$P_D = Q\left(Q^{-1}(P_F) - \sqrt{\frac{A^2}{\sigma^2}}\right)$$

for all $A \neq 0$.

$P_F = 0.5, \quad \sigma^2 = 0.5$



Clairvoyant: $P_D = Q\left(Q^{-1}(P_F) - \sqrt{\frac{A^2}{\sigma^2}}\right)$

Suboptimal: $P_D = Q\left(Q^{-1}\left(\frac{P_F}{2}\right) - \sqrt{\frac{A^2}{\sigma^2}}\right)$

$$+ Q\left(Q^{-1}\left(\frac{P_F}{2}\right) + \sqrt{\frac{A^2}{\sigma^2}}\right)$$

Let's return to the vector case:

$$H_0: \underline{x} = \underline{w}$$

$$H_1: \underline{x} = A\underline{s} + \underline{w} \quad , \quad A \neq 0$$

How might we generalize our previous detector?

Recall the LRT reduces to

$$A\underline{s}^T\underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2 \log(\eta) + A^2 \frac{\underline{s}^T\underline{s}}{2}$$

For a suboptimal detector we could take

$$|\underline{s}^T\underline{x}| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

Exercise | Derive P_D as a function of A, P_F, $\underline{s}$, and σ^2, and compare to clairvoyant detector.

Summary

- Most real-world detection problems involve un
- In very special cases: LRT independent of unk
- One-sided problems: If LR is monotone, UM test exists.
- Two-sided problems: UMP tests never exist, but reasonable suboptimal detectors do.
- Next lecture: General strategies for devising suboptimal detectors when no UMP test exists.

Key a. $\underline{s}^T \underline{x} \sim N(0, \sigma^2 \underline{s}^T \underline{s})$ under H_0

$$P_F = P(\underline{s}^T \underline{x} > \gamma \mid H_0) = Q\left(\frac{\gamma}{\sigma\sqrt{\underline{s}^T \underline{s}}}\right) = \alpha$$

$$\Rightarrow \gamma = \sigma\sqrt{\underline{s}^T \underline{s}}\, Q^{-1}(\alpha)$$

b. $P_D = P(|X| > \gamma \mid H_1) = P(X > \gamma \mid H_1) + P(X < -\gamma \mid H_1)$

$$= Q\left(\frac{\gamma - A}{\sigma}\right) + Q\left(\frac{\gamma + A}{\sigma}\right)$$

c. $P_F = Q\left(\frac{\gamma}{\sigma}\right)$, $\gamma = \sigma Q^{-1}(P_F)$,

$$P_D = Q\left(\frac{\gamma - A}{\sigma}\right) = Q\left(Q^{-1}(P_F) - \frac{A}{\sigma}\right)$$

BAYES FACTORS AND GLRTS

Signal Detection in the presence of unknowns: Part II

We've seen that in special cases of parametric uncertainty (one-sided tests with monotonic likelihood ratios), the LRT reduces to a UMP thresholding test.

Generally, UMP test do not exist. We will study two popular methods for devising (usually) sub-optimal detectors in these more challenging problems:

1. Bayes factors
2. Generalized LRTs

These two methods differ in how they model the unknown parameter $\underline{\theta}$:

1. $\underline{\theta}$ is itself a random quantity

   (Bayesian approach)

2. $\underline{\theta}$ is unknown, but fixed

   (Classical approach)

Bayes Factors

Consider

$$H_0 : \underline{X} \sim f(\underline{x} \mid \underline{\theta}_0)$$

$$H_1 : \underline{X} \sim f(\underline{x} \mid \underline{\theta}_1)$$

← not necessarily same parametric family

where $\underline{\theta}_0$ and $\underline{\theta}_1$ are unknown.

Assume $\underline{\theta}_0$ and $\underline{\theta}_1$ are realizations of the prior distributions

$$f(\underline{\theta}_k \mid H_k), \quad k = 0, 1$$

Then

$$f(\underline{x} \mid H_k) = \int f(\underline{x}, \underline{\theta}_k \mid H_k)\, d\underline{\theta}_k$$

$$= \int f(\underline{x} \mid H_k, \underline{\theta}_k)\, f(\underline{\theta}_k \mid H_k)\, d\underline{\theta}_k$$

Thus the LR statistic is

$$\Lambda(\underline{x}) = \frac{f(\underline{x} \mid H_1)}{f(\underline{x} \mid H_0)}$$

$$= \frac{\int f(\underline{x} \mid H_1, \underline{\theta}_1)\, f(\underline{\theta}_1 \mid H_1)\, d\underline{\theta}_1}{\int f(\underline{x} \mid H_0, \underline{\theta}_0)\, f(\underline{\theta}_0 \mid H_0)\, d\underline{\theta}_0}$$

This "integrated likelihood ratio" is called the Bayes factor for testing H_1 vs. H_0

Example: Geometric vs. Poisson

H_0 : $x_1, \ldots, x_N$ iid geometric

$$x_i \sim \theta_0 (1-\theta_0)^{x_i}, \quad 0 \le \theta_0 \le 1$$

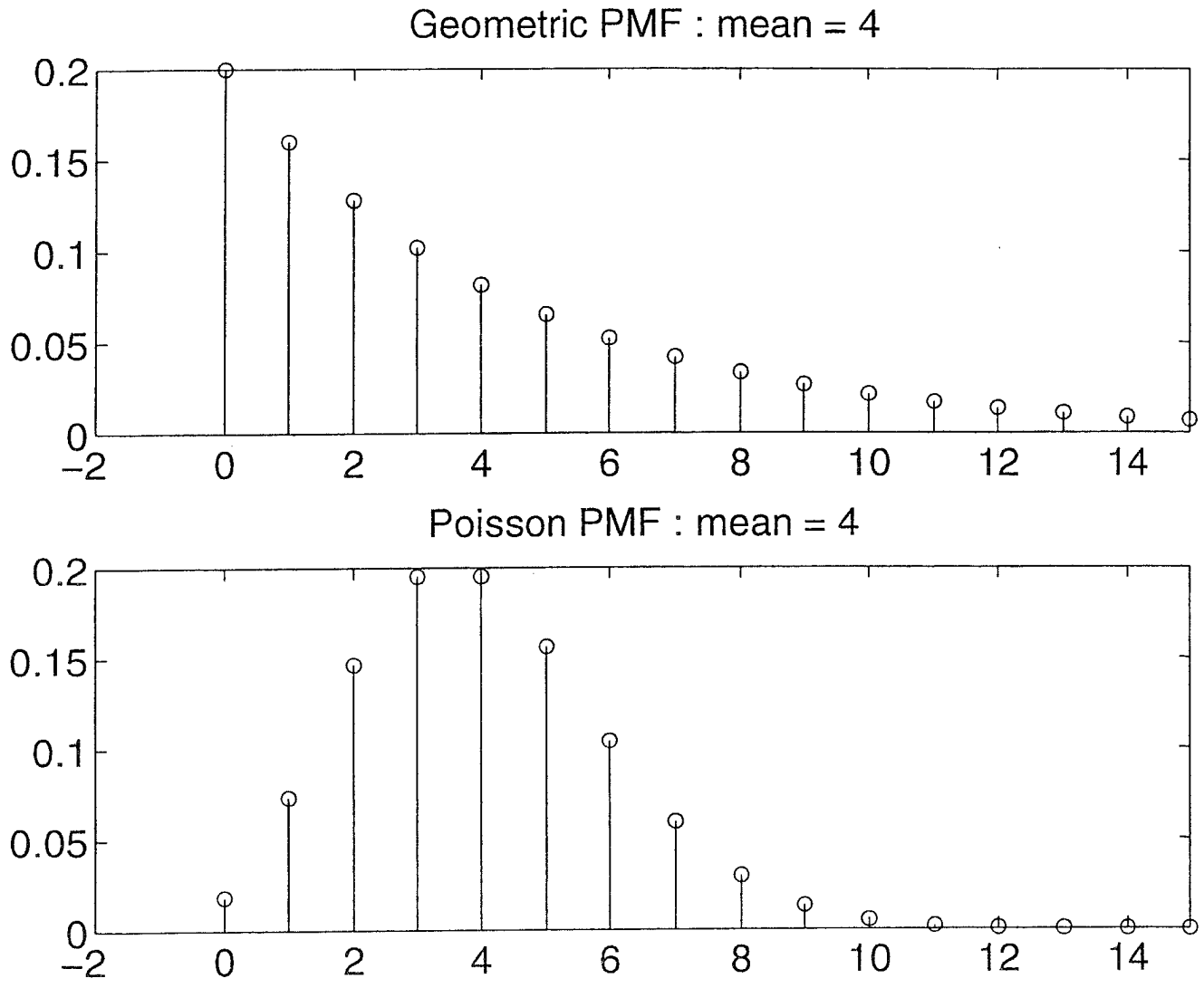
$$f(\underline{x} \mid H_0, \theta_0) = \theta_0^N (1-\theta_0)^{\sum_{i=1}^N x_i}$$

H_1 : $x_1, \ldots, x_N$ iid Poisson

$$x_i \sim e^{-\theta_1} \frac{\theta_1^{x_i}}{x_i!}, \quad \theta_1 \ge 0$$

$$f(\underline{x} \mid H_1, \theta_1) = e^{-N\theta_1} \frac{\theta_1^{\sum_{i=1}^N x_i}}{\prod_{i=1}^N x_i!}$$

Can we apply Karlin-Rubin?


Geometric PMF : mean = 4
0.2
0.15
0.1
0.05
0
−2
0
2
4
6
8
10
12
14
Poisson PMF : mean = 4
0.2
0.15
0.1
0.05
0
−2
0
2
4
6
8
10
12
14

Taking a Bayesian approach, let's assume prior distributions for θ_0 and θ_1:

Modeling uncertainty in θ_k

Under H_0: $\quad 0 \leq \theta_0 \leq 1$

$$f(\theta_0) = I_{[0,1]}(\theta_0)$$

uniform: no preference for any value

Under H_1: $\quad \theta_1 \geq 0$

$$f(\theta_1) = e^{-\theta_1}$$

exponential: favors smaller values

Integrated Likelihoods

$$f(\underline{x} \mid H_0) = \int_0^1 \theta^N (1-\theta)^t \, d\theta = \frac{N!\, t!}{(N+t+1)!}$$

$$t = \sum_{i=1}^{N} x_i$$

$$f(\underline{x} \mid H_1) = \int_0^\infty \frac{e^{-(N+1)\theta}\, \theta^t}{\prod_{i=1}^{N} x_i!} = \frac{t!}{(N+1)^{t+1} \prod_{i=1}^{N} x_i!}$$

Bayes Factor:

$$\Lambda(\underline{x}) = \frac{f(\underline{x} \mid H_1)}{f(\underline{x} \mid H_0)}$$

$$= \frac{(N+t+1)!}{N!\,(N+1)^{t+1} \prod_{i=1}^{N} x_i!} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Comments

1. We carefully chose our priors so that the integrals could be computed in <u>closed form.</u>
2. The "integrated LRT" is optimal only if we used the <u>correct</u> priors.
3. In general, the computationally convenient prior is not the correct prior, so we must
   (a) be content with a <u>suboptimal</u> detector, OR
   (b) resort to time consuming <u>numerical integration</u> techniques
4. Bayes factor has unknown distribution; must <u>set threshold experimentally.</u>

Generalized Likelihood Ratio Tests (GLRTs)

Consider two competing models

$$H_0: \underline{X} \sim f(\underline{x} \mid \underline{\theta}_0)$$

$$H_1: \underline{X} \sim f(\underline{x} \mid \underline{\theta}_1)$$

The models each have unknown parameters (not necessarily the same distributional family)

Idea: Use the data to estimate the unknown parameters and plug in to the LRT

$$\tilde{\Lambda}(\underline{x}) = \frac{f(\underline{x} \mid \hat{\underline{\theta}}_1)}{f(\underline{x} \mid \hat{\underline{\theta}}_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \leftarrow \text{GLRT}$$

$\hat{\theta}_k = \hat{\theta}_k(\underline{x})$ = data-based estimate

When estimating $\underline{\theta}$, we use the

maximum likelihood estimate (MLE)

$$\hat{\underline{\theta}}_{ML} = \arg \max_{\underline{\theta}} f(\underline{x} \mid \underline{\theta})$$

In summary, the GLRT is

$$\tilde{\Lambda}(\underline{x}) = \frac{\max_{\underline{\theta}_1} f(\underline{x} \mid H_1, \underline{\theta}_1)}{\max_{\underline{\theta}_0} f(\underline{x} \mid H_0, \underline{\theta}_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Notes

1. We maximize the numerator and denominator separately. We do not maximize their ratio.

2. It can be shown that under mild conditions the GLRT

$$\tilde{\Lambda}(\underline{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

is asymptotically (as $N \to \infty$) UMP among all decision rules that are invariant to the unknown parameters (i.e., that don't depend on the unknown parameters) See Kay, vol. II.

Ex | Geometric vs. Poisson

Geometric:

$$\hat{\theta}_{ML}|H_0 = \arg\max_{\theta}\left[\theta^N(1-\theta)^t\right], \quad t = \sum_{i=1}^{N} x_i$$

$$\frac{\partial}{\partial\theta}\left(\theta^N(1-\theta)^t\right) = N\theta^{N-1}(1-\theta)^t - \theta^N t(1-\theta)^{t-1}$$

set derivative to zero $\Rightarrow$

$$N\theta^{N-1}(1-\theta)^t = \theta^N t(1-\theta)^{t-1}$$

$$\Rightarrow N(1-\theta) = t\theta$$

$$\Rightarrow \hat{\theta}_{ML}|H_0 = \frac{1}{1+t/N} = \frac{1}{1+\frac{1}{N}\sum x_i}$$

Poisson

$$\hat{\theta}_{ML}|H_1 = \arg\max_{\theta}\left[e^{-N\theta}\frac{\theta^t}{\prod_{i=1}^{N} x_i!}\right]$$

$$\frac{\partial}{\partial\theta}\left(e^{-N\theta}\frac{\theta^t}{\prod x_i!}\right) = -Ne^{-N\theta}\frac{\theta^t}{\prod x_i!} + e^{-N\theta}\frac{t\theta^{t-1}}{\prod x_i!}$$

set derivative to zero $\Rightarrow$

$$Ne^{-N\theta}\frac{\theta^t}{\prod x_i!} = e^{-N\theta}\frac{t\theta^{t-1}}{\prod x_i!}$$

$$\Rightarrow N\theta = t$$

$$\Rightarrow \hat{\theta}_{ML}|H_1 = \frac{t}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

GLRT

$$\tilde{\Lambda}(\underline{x}) = \frac{\max_{\theta_1} f(\underline{x} \mid H_1, \theta_1)}{\max_{\theta_0} f(\underline{x} \mid H_0, \theta_0)}$$

$$= \frac{e^{-t}\left(t/N\right)^t}{\left(\prod_{i=1}^{N} x_i!\right)\left(\frac{1}{1+t/N}\right)^N\left(\frac{t/N}{1+t/N}\right)^t} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

The GLR statistic does not involve the unknown parameters, but it is still difficult to set the threshold because the distribution of $\tilde{\Lambda}(\underline{x})$ is unknown.

The threshold must be set experimentally or through numerical simulations.

Exercise Consider the detection problem

$$H_0: \quad X(n) = w(n)$$

$$H_1: \quad X(n) = A + w(n)$$

$n = 0, 1, \ldots, N-1$

where A is unknown and $w(n) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, σ^2 known.

Find the GLRT. Set the threshold to ensure a false alarm rate of α.

Solution

$$\hat{A}_{ML} = \arg\max \left[\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{N} (x_i - A)^2 \right\} \right]$$

$$= \arg\min \left[\sum_{i=1}^{N} (x_i - A)^2 \right]$$

$$\frac{\partial}{\partial A} \left(\sum (x_i - A)^2 \right) = -2 \sum_{i=1}^{N} (x_i - A) = -0$$

$$\Rightarrow \hat{A}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i \equiv t$$

$$\tilde{\Lambda}(\underline{x}) = \frac{\exp\left\{ \frac{-1}{2\sigma^2} \sum (x_i - \hat{A})^2 \right\}}{\exp\left\{ \frac{1}{2\sigma^2} \sum x_i^2 \right\}}$$

$$\Rightarrow \log \tilde{\Lambda}(\underline{x}) = \frac{1}{2\sigma^2} \left(-N\hat{A}^2 + 2\hat{A} \sum_{i=1}^{N} x_i \right)$$

$$= \frac{N}{2\sigma^2} \cdot \hat{A}^2 \underset{H_0}{\overset{H_1}{\gtrless}} \log(\eta)$$

$$\Rightarrow \boxed{|t| \gtrless \gamma \equiv \sqrt{\frac{2\sigma^2}{N} \log(\eta)}}$$

$$t = \hat{A} = \frac{1}{N} \sum x_i$$

Under H_0, $T \sim N(0, \frac{\sigma^2}{N})$ and therefore

$$P_F = P(|T| > \gamma \mid H_0)$$

$$= 2P(T > \gamma \mid H_0)$$

$$= 2Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right) = \alpha$$

$$\Rightarrow \quad \gamma = \frac{\sigma}{\sqrt{N}} Q^{-1}\left(\frac{\alpha}{2}\right)$$

Remarks: 1. This is basically the same suboptimal test we studied last time.

2. A Bayes factor test with $A \sim N(0, \tau^2)$ leads to the same detector.

Unknown Signal Delay

Consider a binary test for the presence or absence of a signal. Assume that the signal waveform is known, but its time origin is not.

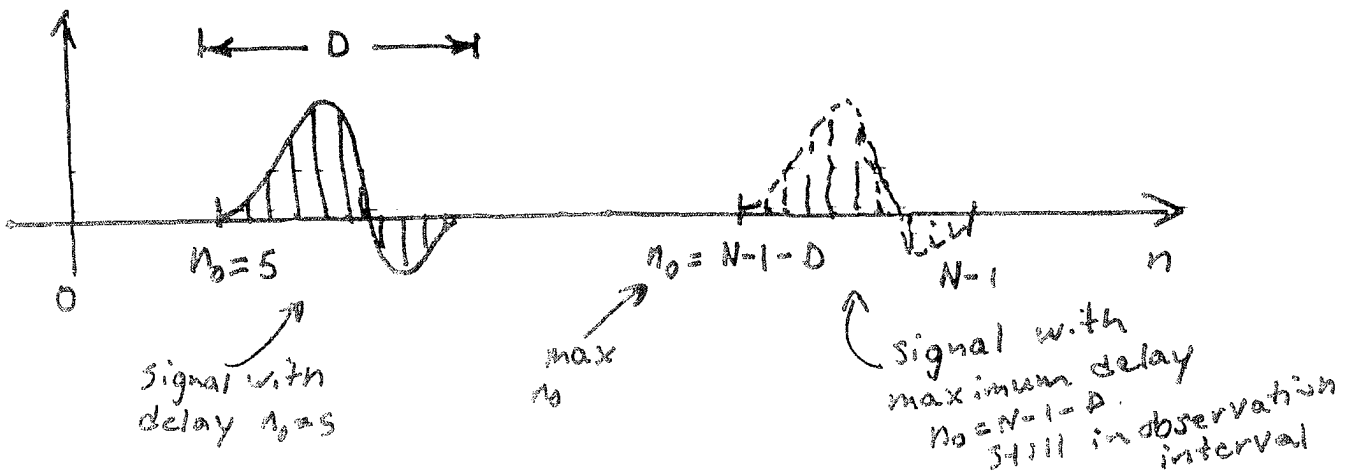
$$H_0: \quad x(n) = w(n) \qquad n = 0, 1, \ldots, N-1$$

$$H_1: \quad x(n) = s(n - n_0) + w(n)$$

n_0 is an unknown integer.

$\underline{W} \sim N(\underline{0}, \sigma^2 I)$.

We will also assume that for all possible values of n_0 the signal lies completely in the observation interval:



Under H_1

$$\underline{x} \sim f(\underline{x} \mid H_1, n_0)$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \left[x(n) - s(n-n_0) \right]^2 \right\}$$

If we form $\Lambda(\underline{x})$, the LR, it is easy to see that it depends on the unknown delay n_0 and a UMP test does not exist.

Therefore, we again will try the GLRT approach.

Maximizing $f(\underline{x} \mid H_1, n_0)$ with respect to n_0 is equivalent to maximizing

$$\sum_{n=0}^{N-1} \left[x(n)\, s(n-n_0) - \frac{1}{2} s^2(n-n_0) \right]$$

$$= \sum_{n=0}^{N-1} \left[x(n)\, s(n-n_0) \right] - \frac{1}{2} \| \underline{s} \|^2 -$$

The log GLRT is then

$$\log \tilde{\Lambda}(\underline{x}) = \log \frac{\max_{n_0} f(\underline{x} \mid H_1, n_0)}{f(\underline{x} \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \log \eta$$

or equivalently

$$\max_{n_0} \sum_{n=n_0}^{n_0+D-1} x(n)s(n-n_0) \underset{H_0}{\overset{H_1}{\gtrless}} \underbrace{\sigma^2 \log \eta + \frac{\|\underline{s}\|^2}{2}}_{\gamma}$$

Using the matched filter interpretation of the test statistic, the decision rule can be expressed as

$$\max_{n_0} \left[x(n) * s(D-1-n) \right] \Big|_{n=D-1+n_0} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

Hence, we simply compute each output of the matched filter (corresponding to each possible delay) and pick the largest value to evaluate the GLRT.

Note that the index at which the maximum occurs gives us the MLE of delay n_0.

$\Rightarrow$ detection and estimation problems are solved simultaneously.

Can we set the threshold to achieve a specified P_f?

Summary

- UMP tests rarely exist so we need methods for designing suboptimal detectors
  - ☐ Bayesian approach
    - → Bayes factor
  - ☐ Classical approach
    - → GLRT
- In practice GLRT is more widely used since integrated likelihoods are often difficult to compute
- Application: signal delay estimation / detection via matched filtering.

CFAR DETECTION

Signal Detection in Unknown Noise Level

When the noise variance or covariance are not known, the signal detection problem becomes significantly more difficult.

- Decision rules can be derived using the techniques we have been discussing, such as the GLRT
- Selecting a meaningful threshold, however, can be very difficult

Why? Both hypotheses depend on the noise! Therefore, we can't compute performance probabilities P_f and P_D.

How can we handle such situations?

Answer: Derive decision rules and thresholds that don't depend on P_F!

Unknown Noise Variance

Suppose we have the signal detection problem

$$H_0: \quad x(n) = w(n)$$

$$H_1: \quad x(n) = s(n) + w(n)$$

$n = 0, \ldots, N-1$

where $s(n)$ is a known signal waveform with a fixed amplitude

$$\underline{w} \sim N(\underline{0}, \sigma^2 R)$$

where R is a known covariance structure, with $\text{trace}(R) = N$. That is R is normalized.

σ^2 is unknown noise power

* Knowing R is equivalent to assuming a known correlation function for the noise, up to a constant which is the overall noise power σ^2.

Log Likelihood Ratio

$$\underline{s}^T R^{-1} \underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \sigma^2 \log \eta + \frac{1}{2} \underline{s}^T R^{-1} \underline{s}$$

γ

test statistic does not depend on σ^2

However, both hypotheses do involve σ^2 and consequently P_D and P_F are functions of σ^2

$\Rightarrow$ we can't determine a γ to guarantee a desired P_F

What can we do?

Perhaps the GLRT will provide a test statistic more amenable to analysis.

Exercise | Find the GLRT. Does the test statistic depend on σ^2 under H_0?

GLRT

$$\tilde{\Lambda}(\underline{x}) = \frac{\max_{\sigma^2} f(\underline{x} \mid H_1, \sigma^2)}{\max_{\sigma^2} f(\underline{x} \mid H_0, \sigma^2)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

$$f(\underline{x} \mid H_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2} |R|^{1/2}} \cdot \exp\left\{ -\frac{1}{2\sigma^2} (\underline{x}-\underline{s})^T R^{-1} (\underline{x}-\underline{s}) \right\}$$

$$f(\underline{x} \mid H_0, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2} |R|^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} \underline{x}^T R^{-1} \underline{x} \right\}$$

$$\hat{\sigma}^2_{ML} \mid H_1 = \frac{(\underline{x}-\underline{s})^T R^{-1} (\underline{x}-\underline{s})}{N}$$

$$\hat{\sigma}^2_{ML} \mid H_0 = \frac{\underline{x}^T R^{-1} \underline{x}}{N}$$

$\Rightarrow$

$$\tilde{\Lambda}(\underline{x}) = \frac{\dfrac{1}{(2\pi \hat{\sigma}^2_{ML}(H_1))^{N/2}}}{\dfrac{1}{(2\pi \hat{\sigma}^2_{ML}(H_0))^{N/2}}} = \left(\frac{\underline{x}^T R^{-1} \underline{x}}{(\underline{x}-\underline{s})^T R^{-1} (\underline{x}-\underline{s})} \right)^{\frac{N}{2}}$$

Derivation of ML noise estimate:

$$f(\underline{x} \mid H_1, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} |R|^{-\frac{1}{2}} \exp\left\{ \frac{-1}{2\sigma^2} \cdot Q \right\}$$

$$Q = (\underline{x} - \underline{s})^T R^{-1} (\underline{x} - \underline{s})$$

$$\Rightarrow \log f(\underline{x} \mid H_1, \sigma^2) = -\frac{N}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} Q$$

$$+ \text{ constant (not depending on } \sigma^2)$$

$$\Rightarrow \frac{\partial}{\partial \sigma^2} \log f(\underline{x} \mid H_1, \sigma^2) = \frac{-N}{2\sigma^2} + \frac{Q}{2(\sigma^2)^2}$$

$$\text{Set to } 0 \Rightarrow \frac{N}{2\sigma^2} = \frac{Q}{2(\sigma^2)^2}$$

$$\Rightarrow \hat{\sigma}^2_{ML} \mid H_1 = \frac{Q}{N} = \frac{(\underline{x} - \underline{s})^T R^{-1} (\underline{x} - \underline{s})}{N}$$

GLRT:

$$\underbrace{\left(\frac{\underline{x}^T R^{-1} \underline{x}}{(\underline{x}-\underline{s})^T R^{-1} (\underline{x}-\underline{s})} \right)^{\frac{N}{2}}}_{\text{test statistic } \tilde{\Lambda}(\underline{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \underbrace{\eta}_{\text{threshold}}$$

To set the threshold to achieve a desired P_F we need to consider the distribution of the test statistic under H_0.

Since both hypotheses depend on the unknown σ^2, we might suppose that $\tilde{\Lambda}(\underline{x})$ does as well.

But what if it didn't? What if $\tilde{\Lambda}(\underline{x})$ didn't depend on σ^2?

Definition | CFAR

If the distribution of the test statistic under H_0 is independent of the unknown parameter (noise variance), then the detector is a constant false-alarm rate (CFAR) detector.

That is, no matter what the unknown parameter is, for a given threshold level, P_F is constant.

Let's see if the GLRT has the CFAR property in this case.

Recall GLRT $\Rightarrow$

$$\tilde{\Lambda}(x) = \left(\frac{\underline{x}^T R^{-1} \underline{x}}{(\underline{x}-\underline{s})^T R^{-1} (\underline{x}-\underline{s})} \right)^{\frac{N}{2}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

Now since $a^b, b>0$ is a monotonic transformation, we have the following equivalent test:

$$\frac{\underline{x}^T R^{-1} \underline{x}}{(\underline{x}-\underline{s})^T R^{-1} (\underline{x}-\underline{s})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta^{2/N}$$

Under H_0, $\underline{x} = \underline{w} = \sigma \tilde{\underline{w}}$

where $\tilde{\underline{w}} \sim N(\underline{0}, R)$

and the test statistic is written

$$\frac{\sigma^2 \tilde{\underline{w}}^T R^{-1} \tilde{\underline{w}}}{(\sigma \tilde{\underline{w}} - \underline{s})^T R^{-1} (\sigma \tilde{\underline{w}} - \underline{s})}$$

In order for the test statistic to be invariant to the unknown noise variance, σ must cancel in the numerator and denominator.

$$\frac{\sigma^2 \tilde{\underline{w}}^T R^{-1} \tilde{\underline{w}}}{(\sigma\tilde{\underline{w}} - \underline{s})^T R^{-1} (\sigma\tilde{\underline{w}} - \underline{s})} \quad \leftarrow \text{test statistic}$$

Unfortunately, σ can't be eliminated and hence P_F will depend on σ.

$\Rightarrow$ GLRT is <u>not</u> CFAR

Remarkably, the situation is dramatically different if we assume that the signal amplitude is also unknown.

Consider the following scenario.

$$H_0: \underline{x} = \underline{w}$$

$$H_1: \underline{x} = A\underline{s} + \underline{w}$$

A unknown

$$\underline{w} \sim N(\underline{0}, \sigma^2 R)$$

$\text{trace}(R) = N$, R known

σ^2 unknown

Under H_0 we can express the observation as

$$\underline{x} = \sigma \underline{\tilde{w}}, \quad \underline{\tilde{w}} \sim N(\underline{0}, R)$$

Under H_1

$$\underline{x} = A\underline{s} + \underline{\sigma}\tilde{w}, \quad \tilde{w} \sim N(\underline{0}, R)$$

$$= \sigma(A'\underline{s} + \tilde{w}),$$ where again A' is just an unknown amplitude $A' = \frac{A}{\sigma}$

In both cases we can view σ as a scaling factor applied to the entire observation.

So, in the case where both the noise power and signal amplitude are unknown, we can view uncertainty in the noise variance as an arbitrary scaling of the data.
This interpretation will enable a GLRT with the CFAR property.

To form a GLRT for this case we also must estimate the unknown signal amplitude in hypothesis H_1.

$$f(\underline{x} \mid H_1, A, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2} |R|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\underline{x}-A\underline{s})^T R^{-1}(\underline{x}-A\underline{s})\right\}$$

So under H_1 we must jointly estimate A and σ^2.

$$(\hat{A}_{ML}, \hat{\sigma}^2_{ML}) = \arg \max_{A,\sigma^2} f(\underline{x} \mid H_1, A, \sigma^2)$$

Maximizing $f(\underline{x} \mid H_1, A, \sigma^2)$ is equivalent to maximizing $\log f(\underline{x} \mid H_1, A, \sigma^2)$.

$$\frac{\partial}{\partial A} \log f(\underline{x} \mid H_1, A, \sigma^2) = \frac{1}{\sigma^2} \underline{s}^T R^{-1} (\underline{x} - A\underline{s})$$

Setting this to zero we find

$$\hat{A}_{ML} = \frac{\underline{s}^T R^{-1} \underline{x}}{\underline{s}^T R^{-1} \underline{s}}$$, independent of σ^2

The joint maximum occurs at $\frac{\partial^2}{\partial \sigma^2 \partial A} \log f(\underline{x} \mid H_1, A, \sigma^2) = 0$

or

$$\frac{\partial}{\partial \sigma^2} \log f(\underline{x} \mid H_1, \sigma^2, \hat{A}_{ML}) = 0$$

And so

$$\hat{\sigma}^2_{ML} \,|\, H_1 = (\underline{x} - \hat{A}_{ML}\underline{s})^T R^{-1} (\underline{x} - \hat{A}_{ML}\underline{s})/N$$

$$= \left(\underline{x} - \frac{\underline{s}^T R^{-1}\underline{x}}{\underline{s}^T R^{-1}\underline{s}}\,\underline{s}\right)^T R^{-1} \left(\underline{x} - \frac{\underline{s}^T R^{-1}\underline{x}}{\underline{s}^T R^{-1}\underline{s}}\,\underline{s}\right)/N$$

Combining this with

$$\hat{\sigma}^2_{ML} \,|\, H_0 = \underline{x}^T R^{-1}\underline{x}/N \quad \text{(as before)}$$

we get a GLRT (in simplified form)

$$\frac{\underline{x}^T R^{-1}\underline{x}}{\left(\underline{x} - \frac{\underline{s}^T R^{-1}\underline{x}}{\underline{s}^T R^{-1}\underline{s}}\,\underline{s}\right)^T R^{-1} \left(\underline{x} - \frac{\underline{s}^T R^{-1}\underline{x}}{\underline{s}^T R^{-1}\underline{s}}\,\underline{s}\right)} \underset{H_0}{\overset{H_1}{\lessgtr}} \gamma$$

Under H_0, the test statistic can be written as

$$\frac{\sigma^2\,\tilde{\underline{w}}^T R^{-1}\tilde{\underline{w}}}{\left(\sigma\tilde{\underline{w}} - \sigma\frac{\underline{s}^T R^{-1}\tilde{\underline{w}}}{\underline{s}^T R^{-1}\underline{s}}\,\underline{s}\right)^T R^{-1} \left(\sigma\tilde{\underline{w}} - \sigma\frac{\underline{s}^T R^{-1}\tilde{\underline{w}}}{\underline{s}^T R^{-1}\underline{s}}\,\underline{s}\right)}$$

and the σ^2 factor on top and bottom cancel!

Therefore, the probability density of the test statistic does not depend on σ^2.

$\Rightarrow$ A threshold can be chosen to insure a specified P_F for any value of σ^2.

The GLRT has the CFAR property!

To actually set the threshold γ we need to relate γ to P_F which requires the test statistic's distribution under H_0.

Let's look at the test statistic more carefully, and to keep things simple set $R = I$ (white noise).

Test statistic: $(R = I)$

$$\frac{\underline{x}^T \underline{x}}{\left(\underline{x} - \frac{\underline{s}^T \underline{x}}{\underline{s}^T \underline{s}} \underline{s}\right)^T \left(\underline{x} - \frac{\underline{s}^T \underline{x}}{\underline{s}^T \underline{s}} \underline{s}\right)} \equiv t(\underline{x})$$

Note

$$\frac{\underline{s}^T \underline{x}}{\underline{s}^T \underline{s}} \underline{s} = \underline{s} \frac{\underline{s}^T \underline{x}}{\underline{s}^T \underline{s}} = \frac{\underline{s}\,\underline{s}^T}{\underline{s}^T \underline{s}} \cdot \underline{x}$$

does this look familiar?

The test statistic is written as

$$t(\underline{x}) = \frac{\underline{x}^T \underline{x}}{(\underline{x} - P_s \underline{x})^T (\underline{x} - P_s \underline{x})}$$

Look at the denominator:

$$(\underline{x} - P_s \underline{x})^T (\underline{x} - P_s \underline{x}) =$$

(a)

$$=$$

We now have

$$t(\underline{x}) = \frac{\underline{x}^T \underline{x}}{\underline{x}^T \underline{x} - \underline{x}^T P_s \underline{x}}$$

$$= \frac{\underline{x}^T (I - P_s)\underline{x} + \underline{x}^T P_s \underline{x}}{\underline{x}^T (I - P_s)\underline{x}}$$

$$= 1 + \frac{\underline{x}^T P_s \underline{x}}{\underline{x}^T (I - P_s)\underline{x}}$$

This gives us a modified test

$$t'(\underline{x}) = \frac{\underline{x}^T P_s \underline{x}}{\underline{x}^T (I - P_s)\underline{x}} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma - 1 \equiv \gamma'$$

Ok. So what? Well let's try to determine the distribution of the numerator and denominator of $t'(\underline{x})$.

Proposition | If P is a rank r projection matrix and $\underline{X} \sim \mathcal{N}(\underline{0}, \sigma^2 I)$, then $\frac{\underline{X}^T P \underline{X}}{\sigma^2} \sim \chi^2_r$

Proof:

$$P = U \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix} U^T \quad \text{(from 551)}$$

$$\Rightarrow \quad P = \sum_{i=1}^{r} \underline{u}_i \underline{u}_i^T$$

$$\Rightarrow \quad \underline{X}^T P \underline{X} = \sum_{i=1}^{r} \underline{X}^T \underline{u}_i \underline{u}_i^T \underline{X} = \sum_{i=1}^{r} (\underline{u}_i^T \underline{X})^2$$

Now $\underline{u}_i^T \underline{X} \sim \mathcal{N}(0, \sigma^2)$ since $\underline{u}_i^T \underline{u}_i = 1$.

In addition, if $i \neq j$, $\underline{u}_i^T X$ and $\underline{u}_j^T \underline{X}$ are uncorrelated (and hence independent) because

(b) $$E\left[(\underline{u}_i^T \underline{X})(\underline{u}_j^T \underline{X})\right] =$$

Def | If $U \sim \chi^2_p$, $V \sim \chi^2_q$ are independent, and

$Z = \dfrac{U/p}{V/q}$, we say Z has an F-distribution with

p, q degrees of freedom.

Recall our test statistic $t'(\underline{x}) = \dfrac{\underline{x}^T P \underline{x}}{\underline{x}^T (I-P)\underline{x}} = \dfrac{\underline{x}^T P \underline{x} / \sigma^2}{\underline{x}^T (I-P)\underline{x} / \sigma^2}$

Clearly the numerator and denominator are chi-squared RVs.
Are they independent? Writing

$$I - P = \sum_{i=r+1}^{N} \underline{u}_i \underline{u}_i^T$$

we can argue that $\underline{u}_i^T \underline{x}$, $i \le r$, and $\underline{u}_j^T \underline{x}$, $j \ge r+1$, are independent (as before), and therefore the two chi-squared RVs are independent.

In our case, $r = 1$ $\left(P = P_s = \dfrac{\underline{s}\,\underline{s}^T}{\underline{s}^T \underline{s}} \right)$.

Therefore, under H_0

$$(N-1)\, t'(\underline{x}) \sim F_{1, N-1}$$

$$\Rightarrow \quad \gamma' = \frac{1}{N-1} Q^{-1}_{F_{1,N-1}}(\alpha)$$

ensures $P_F = \alpha$ regardless of σ^2.

Summary:

CFAR: test statistic's distribution under H_0 is independent of unknown parameters $\Rightarrow$ constant false alarm rate

Most tests are not CFAR.

In special cases, like the unknown noise variance and unknown signal amplitude scenario, the GLRT is CFAR and F-distributed.

This allows us to design a detector and set a threshold to achieve a desired P_F even though σ^2 is unknown.

Key a. $(\underline{x} - P_s \underline{x})^T (\underline{x} - P_s \underline{x})$

$= \underline{x}^T \underline{x} - \underline{x}^T P_s \underline{x} - \underline{x}^T P_s^T \underline{x} + \underline{x}^T P_s^T P_s \underline{x}$

$= \underline{x}^T \underline{x} - \underline{x}^T P_s \underline{x}$

[$P = P^T$ and $P = P^2$ for projections]

$= \underline{x}^T (I - P_s) \underline{x}$

b. $E\left[(\underline{u}_i^T \underline{X})(\underline{u}_j^T \underline{X}) \right]$

$= E\left[\underline{u}_i^T \underline{X} \cdot \underline{X}^T \underline{u}_j \right]$

$= \underline{u}_i^T \cdot \sigma^2 I \cdot \underline{u}_j$

$= 0$

since $\underline{u}_i^T \underline{u}_j = \delta_{ij}$

APPLICATION: RAYLEIGH FADING CHANNEL

Goal: detect a sinusoid at a known frequency f_0, $0 < f_0 < 1/2$.

$$\left.\begin{aligned} H_0 &: x(n) = w(n) \\ H_1 &: x(n) = A\cos(2\pi f_0 n + \phi) + w(n) \end{aligned}\right\} n = 0, 1, \ldots, N-1$$

where $w(n) \overset{iid}{\sim} N(0, \sigma^2)$.

Assumptions

1. Time varying channel: A and ϕ are not fixed, but change with time due to moving transmitter/receiver or changing environmental conditions.
2. Multipath arrivals: Received signal is superposition of several versions of transmitted signal.

Examples: wireless comm, sonar

We have developed two general strategies for detection with unknown parameters, the Bayes Factor and GLRT.

Given our assumptions, the Bayesian approach makes more sense.

Observe

$$A\cos(2\pi f_0 n + \phi) = a\cos 2\pi f_0 n + b \sin 2\pi f_0 n$$

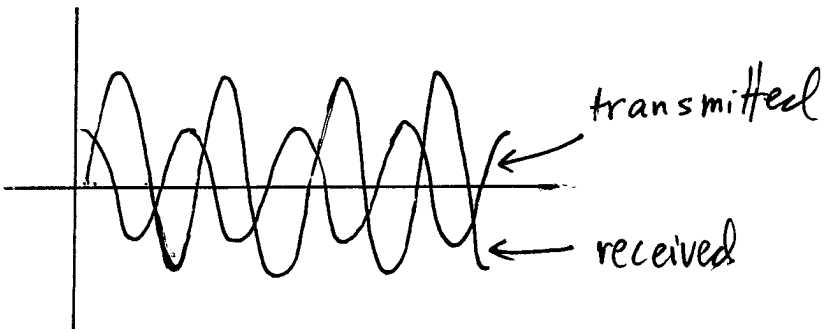
where

$$a = A\cos\phi, \qquad b = A\sin\phi$$

Let's specify the following prior:

$$\underline{\theta} = \begin{bmatrix} a \\ b \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix} \right)$$

↑ due to many multipath arrivals and appeal to central limit theorem



same frequency

different phase and amplitude

The phrase "Rayleigh fading" comes from the fact that

$$A = \sqrt{a^2 + b^2}$$

has a Rayleigh distribution.

$$f(A) = \begin{cases} \frac{A}{\tau^2} \exp\left(\frac{-A^2}{2\tau^2}\right), & A > 0 \\ 0 & A < 0 \end{cases}$$

Also note

$$\phi = \arctan\left(-\frac{b}{a}\right) \sim \text{Uniform}[0, 2\pi]$$

Bayes Factor

$$\frac{f(\underline{x} \mid H_1)}{f(\underline{x} \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad \text{where}$$

$$f(\underline{x} \mid H_1) = \int f(\underline{x} \mid H_1, \underline{\theta}) f(\underline{\theta} \mid H_1) \, d\underline{\theta}$$

Exercise | (a) Determine $f(\underline{x}|H_1)$ (b) Use the MIL and a suitable approximation to simplify the detector (c) Interpret the detector (d) Choose threshold to ensure $P_F = \alpha$.

MIL: $(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(DA^{-1}B + C^{-1}\right)^{-1}DA$

Solution (a) Writing $\underline{\theta} = \begin{bmatrix} a \\ b \end{bmatrix}$ and

$$H = \begin{bmatrix} 1 & 0 \\ \cos(2\pi f_0) & \sin(2\pi f_0) \\ \vdots & \vdots \\ \cos(2\pi f_0 (N-1)) & \sin(2\pi f_0 (N-1)) \end{bmatrix}$$

our model for H_1 is

$$H_1: \quad \underline{X} = H\underline{\theta} + \underline{W}$$

where

$$\underline{\theta} \sim \mathcal{N}(\underline{0}, \tau^2 I_{2\times 2}), \quad \underline{W} \sim \mathcal{N}(\underline{0}, \sigma^2 I_{N\times N})$$

$\underline{\theta}$ and $\underline{W}$ independent

Therefore

$$\underline{X} = \begin{bmatrix} H & I \end{bmatrix} \begin{bmatrix} \underline{\theta} \\ \underline{W} \end{bmatrix}$$

$$\sim \mathcal{N}(\underline{0}, \sigma^2 I_{N\times N} + \tau^2 HH^T)$$

(b) log LRT:

$$-\frac{1}{2}\left[\underline{x}^T\left(\sigma^2 I + \tau^2 H H^T\right)^{-1}\underline{x} - \underline{x}^T(\sigma^2 I)^{-1}\underline{x}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \log\eta$$

Let's apply the matrix inversion lemma:

$$\left(\sigma^2 I + \tau^2 H H^T\right)^{-1} = \frac{1}{\sigma^2}\underline{I} -$$

$$\frac{1}{\sigma^2} I H \left(H^T \frac{1}{\sigma^2} I H + \frac{1}{\tau^2} I_{2\times 2}\right)^{-1} \underline{H}^T \cdot \left(\frac{1}{\sigma^2}\underline{I}\right)$$

Now $H^T H \approx \begin{bmatrix} N/2 & 0 \\ 0 & N/2 \end{bmatrix}$ ← exact when $f_0 = m/N$ approximate for large N, otherwise

So $\frac{1}{\sigma^2} H^T H + \frac{1}{\tau^2} I_{2\times 2} = \begin{pmatrix} \frac{N}{2\sigma^2} + \frac{1}{\tau^2} & 0 \\ 0 & \frac{N}{2\sigma^2} + \frac{1}{\tau^2} \end{pmatrix}$

$$\Rightarrow \left(\ \downarrow\ \right)^{-1} = \frac{1}{\frac{N}{2\sigma^2} + \frac{1}{\tau^2}} \underline{I}_{2\times 2}$$

$\Longrightarrow$: Log LRT =

$$\frac{1}{2}\underline{x}^T\left(\frac{1}{\sigma^4}\cdot\frac{1}{\frac{N}{2\sigma^2}+\frac{1}{\tau^2}}HH^T\right)\underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \log\eta$$

or

$$\underline{x}^T(HH^T)\underline{x} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

Now $\underline{x}^T(HH^T)\underline{x} = (H^Tx)^T(H^T\underline{x})$

$$= \|H^T\underline{x}\|^2$$

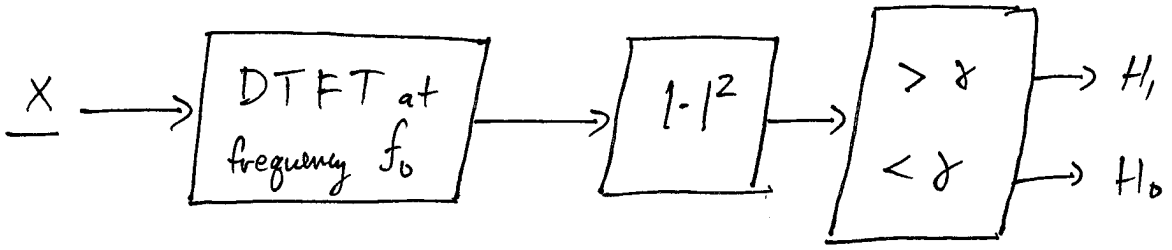
$$= \left\| \begin{bmatrix} \sum x(n)\cos(2\pi f_o n) \\ \sum x(n)\sin(2\pi f_o n) \end{bmatrix} \right\|^2$$

$$= \left(\sum x(n)\cos(2\pi n f_o)\right)^2 + \left(\sum x(n)\sin(2\pi n f_o)\right)^2$$

$$= \left|\sum x(n)e^{-i2\pi f_o n}\right|^2$$

$$\equiv P(\underline{x})$$

(c) Interpretation: Optimal detector given by



Fast implementation possible with FFT.

(d) How should we set γ so that $P_F = \alpha$?

Observe $\sum x(n) \cos(2\pi f_0 n) \sim N\left(0, \frac{N\sigma^2}{2}\right)$

under H_0.

Similarly, $\sum x(n) \sin(2\pi f_0 n) \sim N\left(0, \frac{N\sigma^2}{2}\right)$

under H_0.

$$\Rightarrow \frac{2}{N\sigma^2} \Gamma(\underline{x}) \sim \chi^2_2$$

$$\Rightarrow P_F = Pr\{\Gamma(\underline{x}) > \gamma\} = Pr\left\{\frac{2}{N\sigma^2}\Gamma(\underline{x}) > \frac{2}{N\sigma^2}\gamma\right\}$$

$$= Q_{\chi^2_2}\left(\frac{2}{N\sigma^2}\gamma\right) \Rightarrow \gamma = \frac{\sigma^2 N}{2} Q^{-1}_{\chi^2_2}(\alpha)$$

Summary

- To detect a sinusoid with unknown phase/amplitude, just look at the magnitude (squared) of the frequency component.
- Test statistic and threshold independent of prior variance τ^2.

A STATISTICIAN'S PERSPECTIVE

Cultural Differences

The subject matter of this course is very important in the field of statistics. The subjects of Estimation and Detection are likely to be covered in a course on "statistical inference," while Filtering and Spectral Estimation will be taught in the context of "time series analysis."

A statistician will emphasize certain topics moreso than an engineer (and vice versa). A sucessfull statistician/engineer should have command of the material/terminology in both fields. This set of notes is intended to help the engineer bridge that gap.

The two fields often use different terminology. Here are some examples. Often both terms will be used but one is usually preferred.

Engineering	**Statistics**
Statistical signal processing	Statistical data analysis
Estimation theory	Point estimation
Detection theory	Hypothesis testing
Filtering	Time series analysis
Gaussian distribution	Normal distribution
False alarm	Type I error, false positive
Miss	Type II error, false negative
Detector	Test
False alarm rate	Size, significance level
Detection rate	Power
GLRT	LRT

The Multivariate Gaussian: Composite Testing and Sampling Distributions

One sample problems

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

Two sample problems

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$$

$$Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$$

COMPOSITE HYPOTHESES IN THE UNIVARIATE GAUSSIAN MODEL	GLRT
TESTS ON THE MEAN: σ^2 KNOWN	
CASE III: $H_0 : \mu = \mu_o$, $H_1 : \mu \neq \mu_o$	$\mathcal{N}$
TESTS ON THE MEAN: σ^2 UNKNOWN	
CASE I: $H_0 : \mu = \mu_o$, $\sigma^2 > 0$, $H_1 : \mu > \mu_o$, $\sigma^2 > 0$	t
CASE II: $H_0 : \mu \leq \mu_o$, $\sigma^2 > 0$, $H_1 : \mu > \mu_o$, $\sigma^2 > 0$	t
CASE III: $H_0 : \mu = \mu_o$, $\sigma^2 > 0$, $H_1 : \mu \neq \mu_o$, $\sigma^2 > 0$	t
TESTS ON VARIANCE: KNOWN MEAN	
CASE I: $H_0 : \sigma^2 = \sigma_o^2$, $H_1 : \sigma^2 > \sigma_o^2$	χ^2
CASE II: $H_0 : \sigma^2 \leq \sigma_o^2$, $H_1 : \sigma^2 > \sigma_o^2$	χ^2
CASE III: $H_0 : \sigma^2 = \sigma_o^2$, $H_1 : \sigma^2 \neq \sigma_o^2$	χ^2
TESTS ON VARIANCE: UNKNOWN MEAN	
CASE I: $H_0 : \sigma^2 = \sigma_o^2$, $H_1 : \sigma^2 > \sigma_o^2$	χ^2
CASE II: $H_0 : \sigma^2 < \sigma_o^2$, $\mu \in \mathbb{R}$, $H_1 : \sigma^2 > \sigma_o^2$, $\mu \in \mathbb{R}$	χ^2
CASE III: $H_0 : \sigma^2 = \sigma_o^2$, $\mu \in \mathbb{R}$, $H_1 : \sigma^2 \neq \sigma_o^2$ $\mu \in \mathbb{R}$	χ^2
TESTS ON EQUALITY OF MEANS: UNKNOWN VARIANCE	
CASE I: $H_0 : \mu_x = \mu_y$, $\sigma^2 > 0$, $H_1 : \mu_x \neq \mu_y$, $\sigma^2 > 0$	t
CASE II: $H_0 : \mu_y \leq \mu_x$, $\sigma^2 > 0$, $H_1 : \mu_y > \mu_x$, $\sigma^2 > 0$	t
TESTS ON EQUALITY OF VARIANCES	
CASE I: $H_0 : \sigma_x^2 = \sigma_y^2$, $H_1 : \sigma_x^2 \neq \sigma_y^2$	F
CASE II: $H_0 : \sigma_x^2 = \sigma_y^2$, $H_1 : \sigma_y^2 > \sigma_x^2$	F
TESTS ON CORRELATION	
CASE I: $H_0 : \rho = \rho_o$, $H_1 : \rho \neq \rho_o$	t
CASE II: $H_0 : \rho = 0$, $H_1 : \rho > 0$	t

These basic tests are widely used in applied statistics, and therefore it is customary to catalogue them and study them together. The tests are usually named according to the sampling distribution, that is, the distribution of the test statistic.

Gaussian Sampling Distributions

The GLRT test statistics have one of four different sampling distributions.

Definitions

1. **Chi-square**: If $Z_i \overset{iid}{\sim} N(0,1)$, $i = 1, \ldots, r$ and $Y = \sum_{i=1}^{r} Z_i^2$ then $Y \sim \chi^2_r$

2. **Student t**: If $Z \sim N(0,1)$ and $Y \sim \chi^2_r$ (independent), and $X = \dfrac{Z}{\sqrt{Y/r}}$ then $X \sim t_r$.

3. **Fisher F**: If $U \sim \chi^2_p$ and $V \sim \chi^2_q$ (independent) and $W = \dfrac{U/p}{V/q}$, then $W \sim F_{p,q}$

It turns out that the GLRTs are very intuitive, as is usually the case when Gaussianity is assumed. In particular, they tend to involve the sample mean

$$\bar{x} = \frac{1}{n} \sum x_i$$

the sample variance

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

or the sample correlation coefficient $\hat{\rho}$.

This is why the distributions just discussed are important, because

$$\bar{X} \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and $\bar{X}$, S^2 are independent.

Example | One-sample t-test

Suppose a company produces 5 pound bags of sugar for retail. They have a new packaging process and want to test whether their bags have the correct weight. So they measure the weights of n bags, $X_1, \ldots, X_n$, selected at random. Assume $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ with μ, σ^2 unknown.

$$H_0 : \mu = \mu_0 \qquad (\mu_0 = 5)$$
$$H_1 : \mu \neq \mu_0$$

It can be shown (through routine steps) that the GLRT reduces to

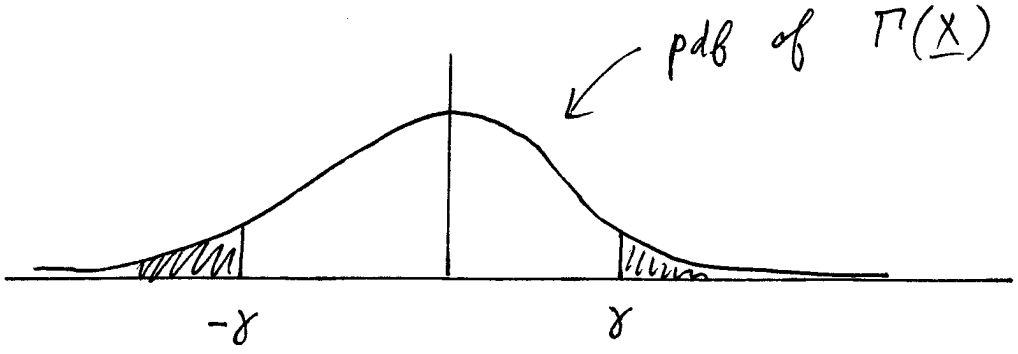
$$|T(\underline{x})| = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

Claim: $T(\underline{X}) \sim t_{n-1}$ under H_0.

To see this, note

$$T(\underline{x}) = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} = \frac{\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}}$$

$$= \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}$$



$$Q_{t_{n-1}}(\gamma) = \frac{\alpha}{2} \Rightarrow \gamma = Q^{-1}_{t_{n-1}}\left(\frac{\alpha}{2}\right)$$

Example / Two sample unpaired t-test.

To test the affect of two treatments for blood pressure, n_1 patients are given one treatment and n_2 patients are given another. Their blood pressures are measured

$$X_1, \ldots, X_{n_1} \sim N(\mu_X, \sigma^2)$$

$$Y_1, \ldots, Y_{n_2} \sim N(\mu_Y, \sigma^2)$$

$$H_0 : \mu_X = \mu_Y$$

$$H_1 : \mu_X \neq \mu_Y$$

The GLRT can be reduced to

$$|T(\underline{x}, \underline{y})| = \left| \frac{\bar{y} - \bar{x}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

where

$$s_p^2 = \frac{1}{n-2} \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2 \right)$$

$$= \frac{(n_1 - 1) s_x^2 + (n_2 - 2) s_y^2}{n-2}$$

Exercise] Use the fact that $\frac{n-2}{\sigma^2} S_p^2 \sim \chi^2_{n-2}$

to show $T(\underline{X}, \underline{Y}) \sim t_{n-2}$ under H_0.

Solution $\quad \bar{X} \sim N\left(\mu_x, \frac{\sigma^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_y, \frac{\sigma^2}{n_2}\right)$

$$\Rightarrow \quad \bar{Y} - \bar{X} \sim N\left(\mu_y - \mu_x, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

So under H_0

$$\frac{\bar{Y} - \bar{X}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1).$$

Thus, under H_0,

$$T(\underline{X}, \underline{Y}) = \frac{\bar{Y} - \bar{X}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{\bar{Y} - \bar{X} \Big/ \left(\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}{\sqrt{\frac{(n-2)S_p^2}{\sigma^2} \Big/ (n-2)}}$$

$$= \frac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \sim t_{n-2}$$

$$\Rightarrow \quad \gamma = Q^{-1}_{t_{n-2}}\left(\frac{\alpha}{2}\right)$$

Example | Two-sample paired t-test.

Suppose we measure a patient's blood pressure before and after a treatment

$$\left.\begin{aligned} X_1, \dots, X_n &\sim N(\mu_x, \sigma^2) \\ Y_1, \dots, Y_n &\sim N(\mu_y, \sigma^2) \end{aligned}\right\} \text{dependent!}$$

We may be able to gain information from the natural pairing of the measurements.

$$H_0: \mu_x = \mu_y$$
$$H_1: \mu_x \neq \mu_y$$

This leads to the paired t-test

$$|T(\underline{x}, \underline{y})| = \left| \frac{(\bar{y} - \bar{x})}{s_d/\sqrt{n}} \right| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

where

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - x_i - (\bar{y} - \bar{x}) \right)^2$$

is the sample variance of the pairwise differences.

It can be shown that $T \sim t_{n-1}$ under H_0.

In essence, the paired t-test is a one-sample t-test with $\mu_0 = 0$ for the differences $Z_i := Y_i - X_i$.

Remark | Both two sample problems assumed $\sigma_x^2 = \sigma_y^2$. If this cannot be assumed, the problem becomes more challenging.

For the unpaired problem (independent samples), the problem is called the Behrens-Fisher problem.

The natural statistic is

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}$$

but it's distribution depends on σ_x / σ_y under H_0!

The most common solution is Welch's approximation,

$$T \sim t_\nu$$

where

$$\nu = \frac{\left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2}{\left(\frac{s_x^2}{n_1}\right)^2/(n_1-1) + \left(\frac{s_y^2}{n_2}\right)^2/(n_2-1)}$$

Hypothesis Falsification

In our discussion of hypothesis testing thus far, we have said we need to either choose H_0 or choose H_1.

However, in many situations, our real objective is to prove the alternative hypothesis H_1 to be true. For this purpose we introduce the null hypothesis H_0, and our goal is to falsify H_0.

This philosophical distinction reflects the scientific viewpoint that it is easier to falsify a hypothesis (H_0) than to prove another (H_1).

H_0: coin is fair

H_1: coin is unfair

If we toss the coin 100 times and observe only 7 heads, we choose H_1 because the data falsifies H_0

For this reason, we can say that the outcome of a test is either to accept H_0 or to reject H_0.

There is a distinction between accepting H_0 and proving H_0 / rejecting H.

If we observe 49 heads of 100 it doesn't prove H_0, we just allow that H_0 might be true.

Said another way, the possible outcomes of a test are

- There is enough evidence to reject H_0
- There is insufficient evidence to reject H_0.

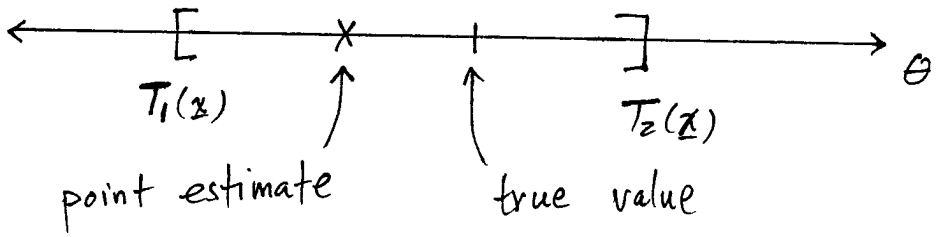
In other words, the null hypothesis is "innocent until proven guilty."

Confidence Intervals

Our study of estimation theory thus far has centered on "point estimation." An alternative is interval estimation.

Rather that outputing a single "point" in the parameter space, an interval estimator outputs an entire interval, called a confidence interval.

$$\underline{x} \longmapsto [T_1(\underline{x}), T_2(\underline{x})]$$



Definition A $100(1-\alpha)\%$ confidence interval for a scalar parameter θ is defined by endpoints $T_1(\underline{x})$, $T_2(\underline{x})$ such that

$$P\{\theta \in [T_1(\underline{X}), T_2(\underline{X})]\} = 1-\alpha.$$

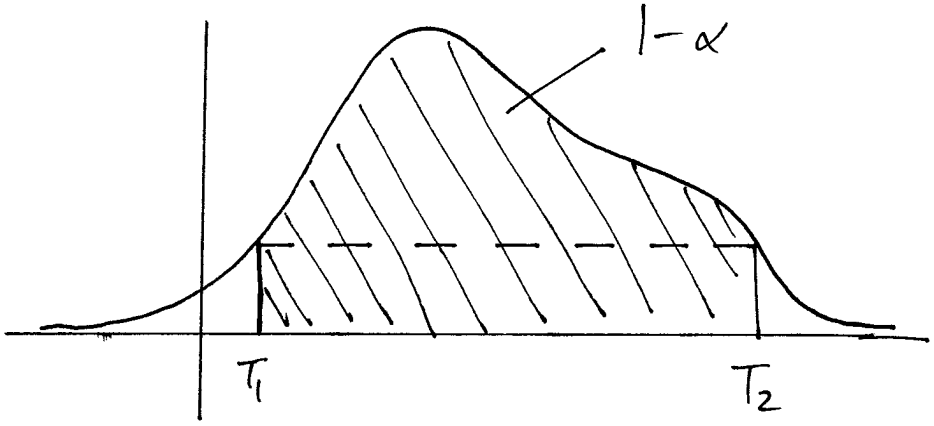
Bayesian confidence intervals

Bayesian estimation specifies a prior and likelihood and returns a posterior distribution.

A Bayesian confidence interval should satisfy

$$\int_{T_1(\underline{x})}^{T_2(\underline{x})} f(\theta|\underline{x})\, d\theta = 1-\alpha.$$

However, there are many such intervals. Therefore, we can impose an additional restriction, for example requiring the confidence interval to have minimal length.



The corresponding interval is a level set of the posterior:

$$[T_1, T_2] = \{\theta : f(\theta|\underline{x}) \geq \lambda\} \quad \text{for some } \lambda.$$

Classical confidence intervals

In classical estimation, the parameter is nonrandom, so what does it even mean to say

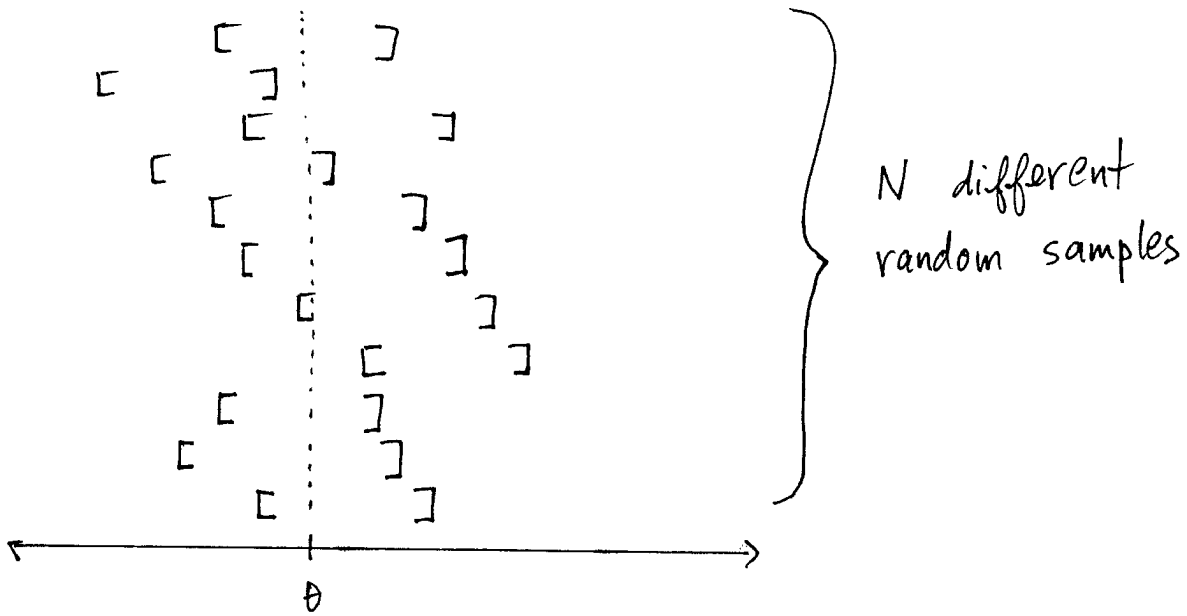
"I'm 95% sure that $\theta \in [T_1, T_2]$"?

Well, we must adjust our thinking. We view the measurement $\underline{X}$ as random so that

$$P\left\{ \theta \in [T_1(\underline{X}), T_2(\underline{X})] \right\}$$

is defined with respect to the randomness of $\underline{X}$.

What does this even mean?



For large N, we expect that at least $N(1-\alpha)$ of the interval estimates contain the true θ.

It turns out that we may derive confidence intervals using results about hypothesis tests.

Example Suppose $X_1, \dots, X_n \overset{iid}{\sim} N(\mu_0, \sigma^2)$ with μ_0, σ^2 unknown. Find a $100(1-\alpha)\%$ confidence interval for μ_0.

Recall the testing problem

$$H_0: \mu = \mu_0 \,, \sigma^2 > 0$$
$$H_1: \mu \neq \mu_0 \,, \sigma^2 > 0$$

We saw that the test

$$\left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_\alpha := Q_{t_{n-1}}^{-1}\left(\frac{\alpha}{2}\right)$$

has size α.

In other words

$$P\left\{ -\gamma_\alpha \leq \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq \gamma_\alpha \right\} = 1-\alpha.$$

when $\mu = \mu_0$.

That is,

$$1-\alpha = P\left\{-\gamma_\alpha \le \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \le \gamma_\alpha\right\}$$

$$= P\left\{-\gamma_\alpha \frac{S}{\sqrt{n}} \le \bar{X}-\mu_0 \le \gamma_\alpha \frac{S}{\sqrt{n}}\right\}$$

$$= P\left\{-\gamma_\alpha \frac{S}{\sqrt{n}} \le \mu_0-\bar{X} \le \gamma_\alpha \frac{S}{\sqrt{n}}\right\}$$

$$= P\left\{\bar{X}-\gamma_\alpha \frac{S}{\sqrt{n}} \le \mu_0 \le \bar{X}+\gamma_\alpha \frac{S}{\sqrt{n}}\right\}$$

Therefore,

$$\left[\bar{X}-\gamma_\alpha \frac{S}{\sqrt{n}},\ \bar{X}+\gamma_\alpha \frac{S}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for the mean.

Example | $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

Find $100(1-\alpha)\%$ CI for σ^2.

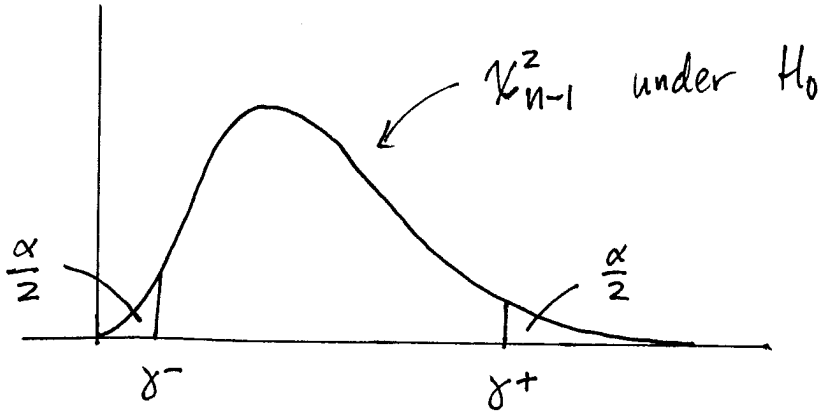
For the testing problem

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

the GLRT reduces to

$$\text{declare } H_0 \iff \gamma^- \leq \frac{(n-1)s^2}{\sigma_0^2} \leq \gamma^+$$



$$\gamma_\alpha^+ = Q^{-1}_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right)$$

$$\gamma_\alpha^- = Q^{-1}_{\chi^2_{n-1}}\left(1-\frac{\alpha}{2}\right)$$

↑ "equal tail" thresholds; other choices are possible, e.g. minimal length

Exercise | Find a $100(1-\alpha)\%$ CI for σ^2.

Solution

$$1-\alpha = P\left\{ \gamma_\alpha^- \le \frac{(n-1)S^2}{\sigma_0^2} \le \gamma_\alpha^+ \right\}$$

$$= P\left\{ \frac{(n-1)S^2}{\gamma_\alpha^+} \le \sigma_0^2 \le \frac{(n-1)S^2}{\gamma_\alpha^-} \right\}$$

$$\Rightarrow \left[\frac{(n-1)S^2}{\gamma_\alpha^+}, \frac{(n-1)S^2}{\gamma_\alpha^-} \right] \text{ is a } 100(1-\alpha)\% \text{ CI.}$$

The basic mechanism for constructing CI's presented here can be generalized using the concept of <u>pivots</u>.

p-values

The size or false alarm rate of a test is sometimes called the significance level.

Our general approach has been to set the significance level α in advance, and to make a hard binary decision (H_0 or H_1) depending on α.

Such "hard decisions" do not convey how close the observation was to the opposite decision.

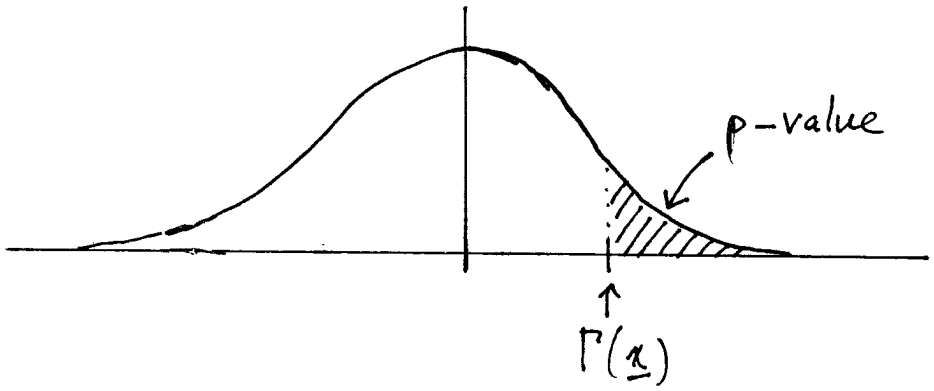
Definition Consider testing a simple null hypothesis against some alternative. The p-value of a measurement x is the probability, under the null hypothesis, of observing a measurement at least as extreme as x

Example $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known

$$\left.\begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{array}\right\} \xrightarrow{\text{GLRT}} T(\underline{x}) = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

Ignore the threshold for the moment.

The distribution of $T(\underline{X})$ is $N(0,1)$ under H_0.



The probability (under H_0) of $T(\underline{X})$ being more extreme than $T(\underline{x})$ is

(a)

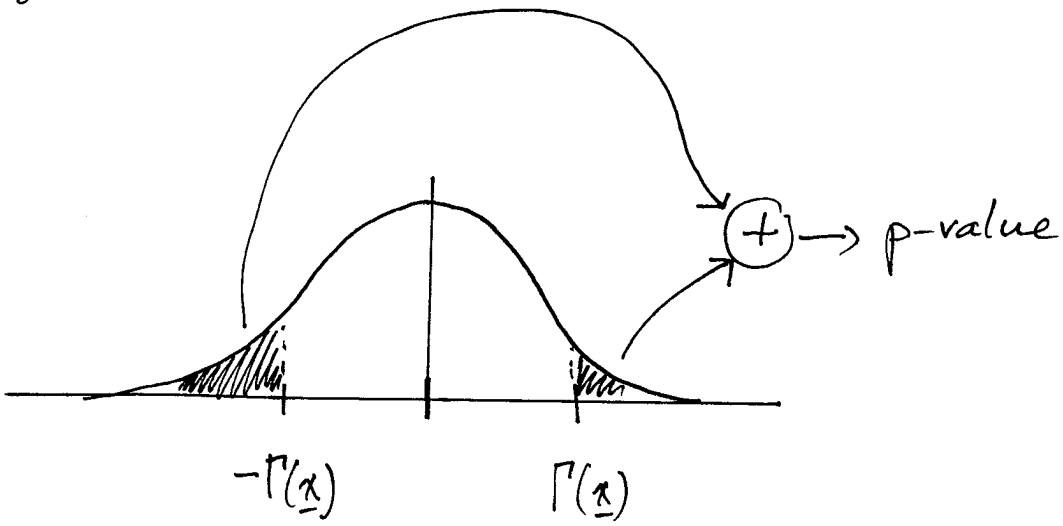
If we want a hard decision at level α we can express this in terms of the p-value as

(b)

Now consider the two-sided problem

$$\left.\begin{array}{l} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{array}\right\} \xrightarrow{\text{GLRT}} |T(\underline{x})| = \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

Again, $T(\underline{X}) \sim N(0,1)$ under H_0, but now "extreme" takes on a new meaning



(c) $\Longrightarrow$ p-value $=$

Typically a p-value $\leq .05$ is considered grounds for rejecting the null hypothesis. p-values are especially useful for addressing the multiple testing problem.

Multiple Testing

Suppose you want to decide whether a certain coin is fair.

$$H_0 : \theta = \frac{1}{2}$$

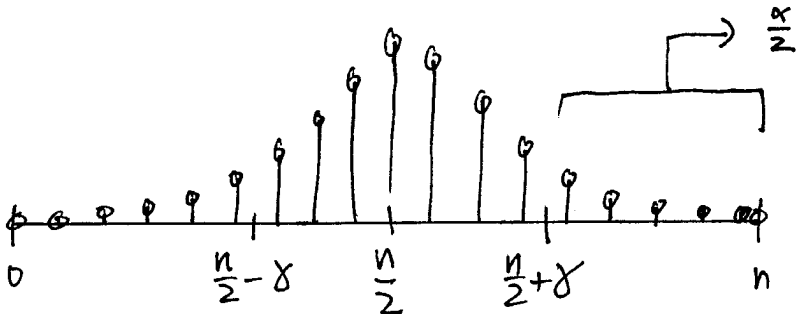
$$H_1 : \theta \neq \frac{1}{2}.$$

So you toss the coin n times and observe the number x of heads. A natural test is

$$\left| x - \frac{n}{2} \right| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

To ensure a false alarm rate of α, we know to choose γ such that

$$2 \cdot \sum_{k = \lceil \frac{n}{2} + \gamma \rceil}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n \approx \alpha$$



Now suppose you are presented with N different coins, and you must determine which of them are fair.

If you perform the test we just discussed, even if all the coins are fair, we expect to "discover" about $N \cdot \alpha$ unfair coins.

Example | If $N = 1000$, $\alpha = .05$, and we discover 50 unfair coins, would you really believe those coins are unfair?

A solution to this conundrum is to adopt an alternative notion of "size." A common choice is the family-wise error rate

$$FWER = P(\geq 1 \text{ } H_0 \text{ rejected} \mid \text{all } H_0 \text{ true})$$

Sidak correction

Suppose all measurements are independent.

Denote Ω_i = event that i^{th}

$$\Omega = \bigcup_{i=1}^{N} \Omega_i$$

Then

$$\begin{aligned} FWER &= P_{\text{all } H_0}(\Omega) \\ &= 1 - P_{\text{all } H_0}(\Omega^c) \\ &= 1 - P_{\text{all } H_0}\left(\bigcap_{i=1}^{N} \Omega_i^c\right) \qquad \text{by independence} \\ &= 1 - \prod_{i=1}^{N} P_{H_0^i}(\Omega_i^c) \\ &= 1 - (1-\alpha)^N \end{aligned}$$

α = size of individual test

Thus, if we desire $FWER \le \alpha'$, it suffices to set $\alpha = 1 - (1-\alpha')^{1/N}$ in each individual test.

Bonferroni Correction

If $\{\Omega_i\}_{i=1}^N$ are not independent, the union bound implies

$$\begin{aligned} FWER &= P_{\text{all } H_0}(\Omega) \\ &= P_{\text{all } H_0}\left(\bigcup_{i=1}^N \Omega_i\right) \\ &\leq \sum_{i=1}^N P_{H_0^i}(\Omega_i) \\ &= N \cdot \alpha \end{aligned}$$

So $\alpha = \frac{\alpha'}{N} \implies FWER \leq \alpha'$.

This "adjustment" is more conservative than the Sidak correction, but also more general.

Equivalently, if $p_1, \ldots, p_N$ are the p-values of the N tests, we may define the "adjusted p-values" $p_i' = N \cdot p_i$ and decide by comparing p_i' to α'.

False Discovery Rate

Many think the FWER is too conservative. It is often worthwhile to allow a few false alarms if the number of correct detections increases significantly.

This led Benjamini and Hochberg to study the false discovery rate (FDR)

$$FDR = E\left[\frac{FD}{D}\right]$$

where

D = # of "discoveries", i.e. H_0 rejected

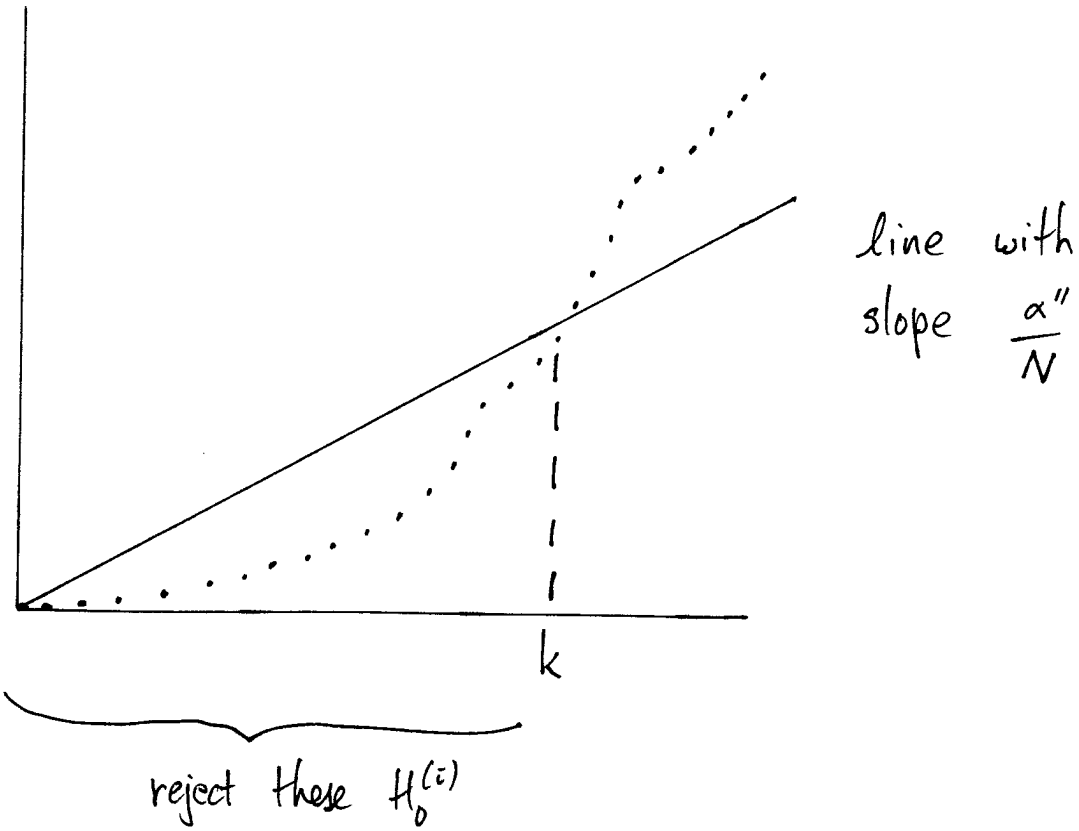
FD = # of "false discoveries", i.e. H_0 incorrectly rejected.

B & H showed how to ensure $FDR \leq \alpha$"

Let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(N)}$ be the ordered p-values, and let $H_0^{(i)}$ be the hypothesis corresponding to $p_{(i)}$. Let k be the largest i such that

$$p_{(i)} \le \frac{i}{N} \alpha''.$$

Reject all $H_0^{(i)}$, $i = 1, 2, \ldots, k$.



$\Rightarrow$ FDR $\le \alpha''$

Summary

- Statisticians study similar problems to those encountered in statistical signal processing, but often with different terminology and emphases.
- Composite testing with multivariate normal data:
  - GLRT yields intuitive tests
  - tests named after distribution of test stat. under H_0
- Classical confidence intervals: equivalence with classical hypothesis testing
- p-values: hypothesis testing with "soft" decisions, convenient for addressing multiple testing problem

Key

a. $Q(\Gamma(\underline{x}))$

b. $p\text{-value}(\underline{x}) = Q(\Gamma(\underline{x})) \underset{H_1}{\overset{H_0}{\gtrless}} \alpha$

c. $2Q(\Gamma(\underline{x}))$