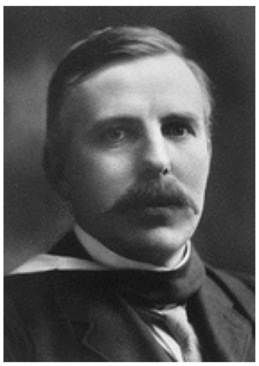
Detection and Estimation Theory

What is this class about?

- **Goal:** Extract useful information from noisy data.
- **Strategy:** Formulate probabilistic model of data x that depends on underlying parameter(s) θ.
- Terminology depends on the parameter space:
  - Estimation:
    - $\boldsymbol{\theta} \in \mathbb{R}^n, \mathbb{C}^n$ etc.
  - Detection (simple hypothesis testing):
    - $\theta \in \{0, 1\}$, i.e. $0 \equiv$ signal absent, $1 \equiv$ signal present.
  - Classification (multihypothesis testing):
    - $\theta \in \{0, 1, \ldots, M-1\}$, e.g. symbols in an M-ary constellation.

Statistics and Science

"If your experiment needs statistics, you ought to have done a better experiment"



Ernest Rutherford (1871-1937)

[0]Taken from http://www.stats.bris.ac.uk/~peter/slides/RS.pps. Please see the whole slide show.

Applications

- Communications,
- Radar and sonar,
- Nondestructive evaluation (NDE) of materials,
- Biomedicine,
- Controls,
- Seismology, etc.

Bibliography: We refer to the following books:

(Kay-I) S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1993, pt. I.

(Kay-II) S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1998, pt. II.

(B & D) P.J. Bickel and K.A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics,* 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2001.

(Gelman et al.) A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, 2nd ed. New York: Chapman & Hall, 2004.

(Wasserman) L. Wasserman, *All of Statistics*. New York: Wiley, 1987.

(Poor) H.V. Poor, *An Introduction to Signal Detection and Estimation,* 2nd ed. New York: Springer-Verlag, 1994.

(Liu) J.S. Liu, *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag, 2001.

(Ripley) B.D. Ripley, *Stochastic Simulation*. New York: Wiley, 1987.

(Van Trees) H.L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968, pt. I.

and a few others.

Basics of Estimation Theory

Basic Ingredients:

- $x \equiv$ observable random variable (measurement that we collect),
- $\theta \equiv$ "true state of nature" (parameter that we wish to estimate),
- $p(x|\theta) \equiv$ data model (probability density or mass of x for a given θ; tells us how likely a particular value of x is given the true state of nature), can be
  - continuous in $x \Longrightarrow$ probability density function (pdf)
    e.g. Gaussian,
  - discrete in $x \Longrightarrow$ probability mass function (pmf)
    e.g. Poisson.

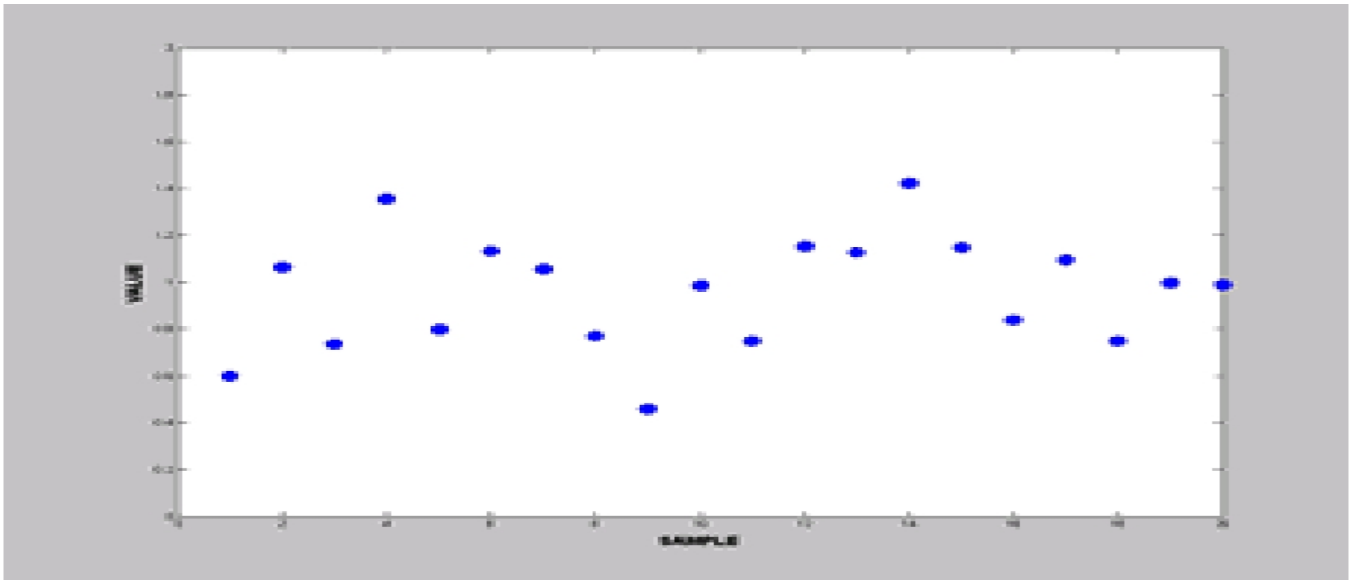
If we decide to assign a probability distribution to $\boldsymbol{\theta}$, then we also need

- $\pi(\theta) \equiv$ prior pdf/pmf on θ (epistemic probability).
  (Epistemic[1] refers to our knowledge about the true state of nature.)

Goal: Find the true state of nature θ.

[1] From the Greek words episteme (knowledge) and epistanai (to know or understand).

Example: Discrete-time Data



Measurements $x[n]$ vs. sample index n, $n = 0, 1, \ldots, N-1$.

Assume that a finite data set is available:

$$\boldsymbol{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = [x[0], x[1], \ldots x[N-1]]^T.$$

The measurements $\boldsymbol{x}$ depend on the parameter θ through a probabilistic model. An estimator of θ is a function of the data:

$$\widehat{\theta}(x[0], x[1], \ldots x[N-1]) = \widehat{\theta}(\boldsymbol{x}).$$

Note: The estimator $\widehat{\theta}(\boldsymbol{x})$ depends *only* on the observed data, i.e. it must be *realizable*.

Types of Estimators:

- **Off-line (batch) estimator.** Example — linear in data:

$$\widehat{\theta} = \sum_{n=0}^{N-1} a_n x[n].$$

- **On-line (sequential) estimator.** Example ($\theta = s[n]$):

$$\widehat{s}[n] = a_0 x[n] + a_1 x[n-1] + \ldots + a_{N-1} x[n-N+1].$$

Maximum Likelihood (ML) Estimation

$p(x \,|\, \theta)$ viewed as a function of θ is the *likelihood function*.

Comments on the likelihood function:

- For a given θ and discrete case, $p(x \,|\, \theta)$ is the probability of observing the data point x. In the continuous case, it is approximately proportional to probability of observing a point in a small rectangle around x.

- However, when we think of $p(x \,|\, \theta)$ as a function of θ, it provides, for a given observed x, the "likelihood" or "plausibility" of various θ's.

ML estimation: Maximize the likelihood with respect to θ, i.e.

$$\widehat{\theta} = \arg\max_{\theta} p(x \,|\, \theta).$$

ML is one of the most popular methods in statistics, communications, and signal processing. We will see later that the mean-square error of ML estimators typically attains the best possible asymptotic performance (given by the Cramér-Rao bound).

Bayesian Inference

In Bayesian inference, parameters (θ, say) are assigned probability distributions and inference is based on the posterior distribution of θ

$$p(\theta \,|\, x) = \underbrace{\frac{p(x|\theta)\,\pi(\theta)}{\int p(x|\vartheta)\,\pi(\vartheta)\,d\vartheta}}_{\text{Bayes' rule (continuous-parameter version)}} . \qquad (1)$$

Note: $p(\theta \,|\, x)$ is an epistemic probability.

Common Bayesian estimators:

- Maximum *a posteriori* (MAP): $\widehat{\theta}_{\mathrm{MAP}} = \arg\max_\theta p(\theta \,|\, x)$ and
- Minimum mean-square error (MMSE): $\widehat{\theta}_{\mathrm{MMSE}} = \mathrm{E}\,[\theta \,|\, x]$.

Comments: MAP estimation is typically the most tractable as it does not require computing the denominator in (1), which is usually analytically intractable. (Note that the denominator is not a function of $\boldsymbol{\theta}$.)

The MMSE estimator is derived by minimizing the Bayesian mean-square error (BMSE):

$$\mathrm{BMSE}(\widehat{\theta}) = \mathrm{E}_{\,x,\theta}[(\widehat{\theta} - \theta)^2].$$

Bayesian vs. Classical (Non-Bayesian) Analysis

- In classical (non-Bayesian) analysis, inference is made based *only* on the probabilistic model

$$p(x \,|\, \theta) = p(x\,;\,\theta)$$

which is called likelihood. When specifying the probabilistic model,

  – in Bayesian inference, we emphasize *conditioning on θ* (i.e. the fact that θ is treated as a random variable) and use $p(x \,|\, \theta)$ to denote the likelihood.
  – in classical inference, we denote the likelihood as $p(x;\theta)$.

Criticism against Bayesian approach:

- subjectivity,
- different inferences possible based on the same data.

Criticism against non-Bayesian approach:

- ignores prior information,
- data that have never been observed used for inference.

Model and Identifiability

A model is a parametrized pdf or pmf $p(\boldsymbol{x}; \boldsymbol{\theta})$.

Example: DC level in Gaussian noise

$$x = \underbrace{\theta}_{\text{parameter}} + \underbrace{w}_{\text{noise } \sim \mathcal{N}(0,\sigma^2)}$$

leading to $x \sim \mathcal{N}(\theta, \sigma^2)$:

$$p(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2\sigma^2}(x-\theta)^2\Big].$$

Note: Kay-I uses $\theta = A$ to denote the DC level, see e.g. Chapter 1.3.

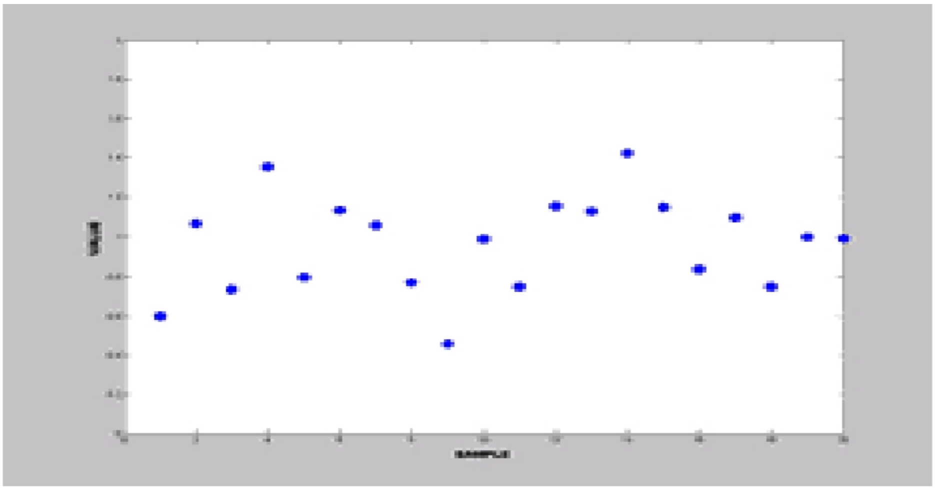
For a specific value of θ, the density defines a model. Our goal is to estimate the "best" model based on observation(s) x.

Identifiability: An important property of a model structure: *parameter identifiability* — for (almost) all values of $\boldsymbol{x}$ and θ we want the following to hold:

$$p(\boldsymbol{x}; \theta) = p(\boldsymbol{x}; \eta) \quad \Longleftrightarrow \quad \theta = \eta.$$

Note: we do not care much about identifiability when deriving estimation algorithms — there are many examples of deliberately fitting models that are not identifiable, some of which we will see in this class (e.g. PX-EM algorithm). But, this needs to be done carefully.

Example: DC Level in White Gaussian Noise



noise

DC-level

observation $\longrightarrow$ $x[n] = A + w[n]$

Choose white Gaussian noise model:

$$w[n] \sim \mathcal{N}(0, \sigma^2), \quad n = 0, 1, \ldots, N-1$$

and

$$p(w[0], w[1], \ldots, w[N-1]) = \prod_{n=0}^{N-1} p(w[n])$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\Big(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} w[n]^2\Big).$$

Hence, $x[n]$ are independent Gaussian $\mathcal{N}(A, \sigma^2)$.

For simplicity, assume first that the noise level σ^2 is known.

Classical inference is based on the likelihood function:

$$p(\boldsymbol{x}; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\Big]. \qquad (2)$$

What if σ^2 is unknown? Then, the likelihood function is

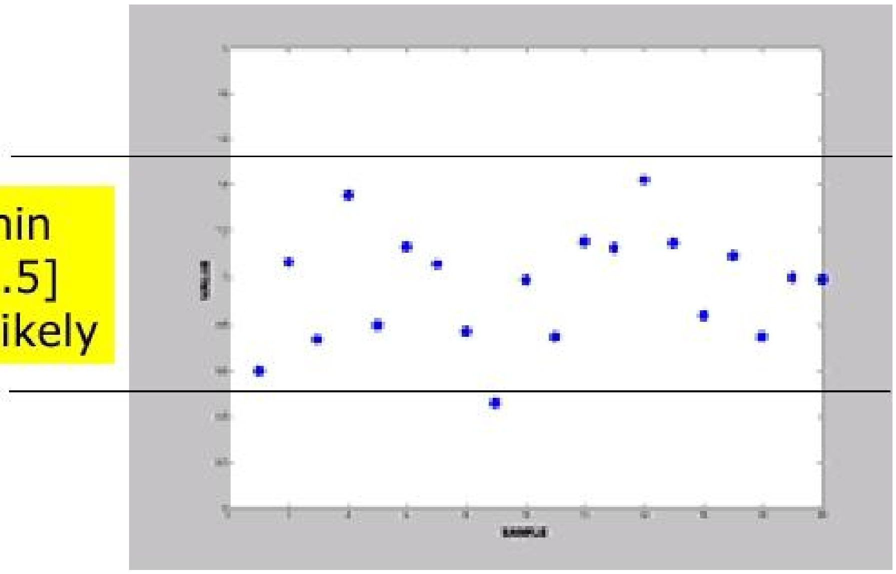
$$p(\boldsymbol{x}; A, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\Big]. \quad (3)$$

What is the difference between (2) and (3)?

Classical estimation theory from a signal processing point of view is covered in detail in Chapters 2–9 of Kay-I.

Example: Bayesian Estimation of DC Level

- A known to be within the interval [0.5 1.5]
- All values equally likely



Again, assume that σ^2 is known. Prior distribution on A:

$$\pi(A) = \begin{cases} 1, & 0.5 \le A \le 1.5 \\ 0, & \text{otherwise} \end{cases}.$$

Bayesian inference based on the posterior distribution of A:

$$\begin{aligned} p(A|x) &= \frac{\overbrace{p(x|A)}^{\text{likelihood function}} \qquad \pi(A)}{\int p(x|A)\,\pi(A)\,dA} \\ &= \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]\pi(A)}{\int_{0.5}^{1.5}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]dA} \end{aligned}$$

Recall common Bayesian estimators (see p. 10 of these notes):

- MAP:

$$\widehat{A}_{\mathrm{MAP}} = \arg\max_A p(A|x) = \arg\max_A [p(x|A)\pi(A)]$$

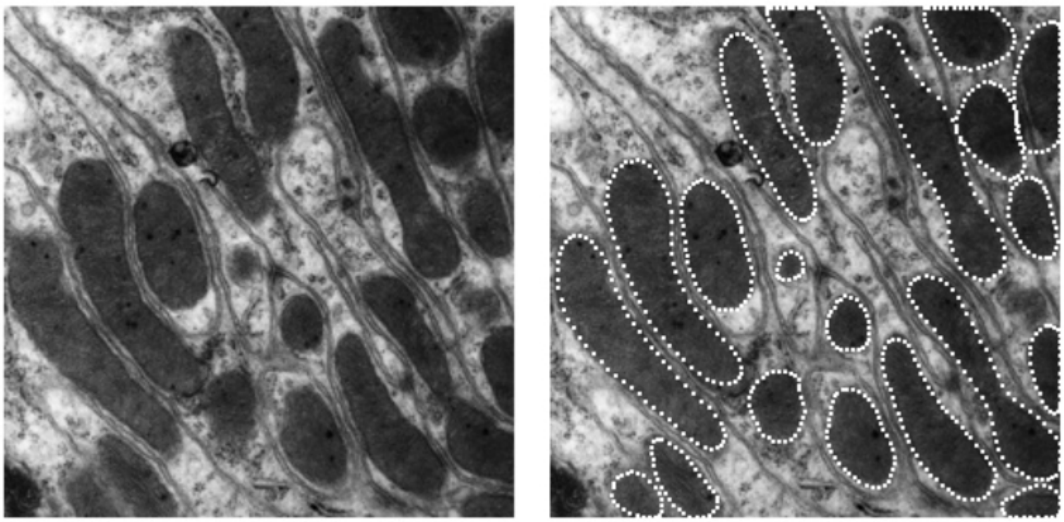
  and

- Minimum mean-square error (MMSE):

$$\widehat{A}_{\mathrm{MMSE}} = \mathrm{E}\,[A|x] = \int A\, p(A|x)\, dA.$$

Bayesian estimation from a signal processing point of view is covered in detail in Chapters 10–13 of Kay-I.

A (Much) More Sophisticated Example: Mitochondria Segmentation



The data $\boldsymbol{x} \equiv$ electron micrograph of a cardiac muscle cell.

The parameter vector $\boldsymbol{\theta}$ contains:

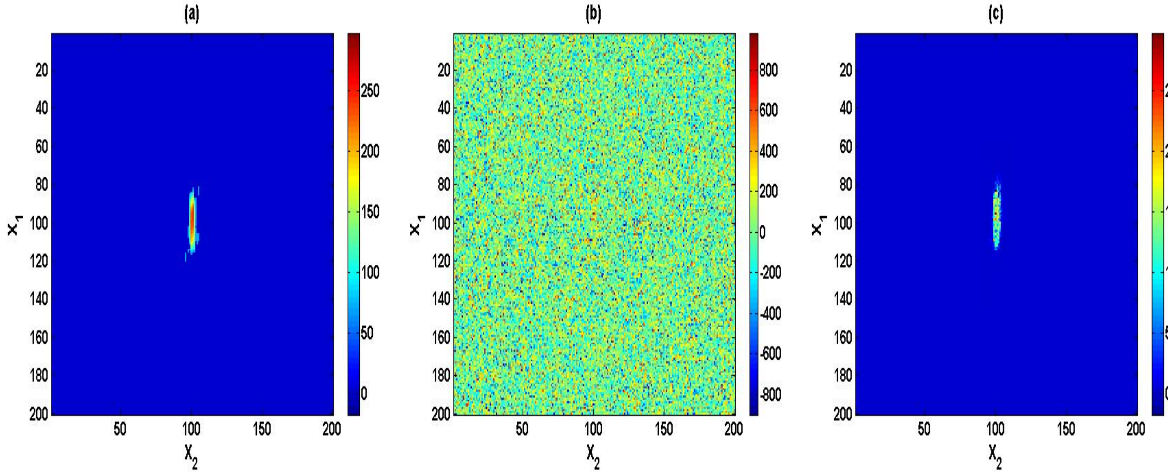
- number of mitochondria and
- Fourier parameters describing mitochondria shapes.

Prior $\pi(\boldsymbol{\theta})$ learned from hand-segmented training data, using several hundred hand-selected electron micrographs.

Idea: Mitochondria and cytoplasm have different textures $\Longrightarrow$ utilize Markov-random-field (MRF) models to learn the texture from the training data. Then use Markov chain Monte Carlo (MCMC) to draw samples from the posterior distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$.

Reference: U. Grenander and M.I. Miller, "Representations of knowledge in complex systems," *J. R. Stat. Soc., Ser. B,* vol. 56, pp. 549–603, 1994.

Another Bayesian MCMC Example



See

A. Dogandžić and B. Zhang, "Markov chain Monte Carlo defect identification in NDE images," to appear in *Proc. Annu. Rev. Progress Quantitative Nondestructive Evaluation (QNDE 2006),* Portland, OR, Aug. 2006.

and

A. Dogandžić and B. Zhang, "Bayesian NDE defect signal analysis," *IEEE Trans. Signal Processing*, vol. 55, pp. 372–378, Jan. 2007.

Some Challenges/ Interesting Topics

Can MCMC methods be used for decoding? (Wainwright & Jordan 03)[2] say that this hasn't been done successfully (yet perhaps?).

Distributed inference on sensor networks, e.g. discovering anomalous activity, fusing different modalities, doing all these tasks in a distributed manner and with limited power budget for communication and computation.

There are many biomedical applications, e.g. estimating and detecting tumors.

[2](**Wainwright & Jordan 03**) M.J. Wainwright and M.I. Jordan, "Graphical models, exponential families, and variational inference," Report no. 649, Department of Statistics, University of California, Berkeley, CA, 2003.