Introduction to Detection and Estimation Theory

Aleksandar Dogandžić

2017-01-10

Contents

Reading: §1 in the textbook.

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.

✻ Taxonomy of detection and estimation is based on the parameter space:

We denote the parameter space of θ by sp_{Θ}.

- Estimation:

$$\boldsymbol{\theta} \in \mathbb{R}^n, \mathbb{C}^n$$

etc.

✻ EXAMPLE: Given the results of N independent flips of a coin, determine the probability p with which it lands on heads.

- Detection (simple hypothesis testing):

$$\theta \in \{0, 1\}$$

corresponding typically to signal absence and presence.

Example: Determine whether or not the coin is fair.

- Classification (multihypothesis testing):

$$\theta \in \{0, 1, \dots, M-1\}$$

e.g., symbols in an M-ary constellation.

Example: Distinguish digital images of handwritten digits.

✻ RELEVANT REFERENCES: (Bickel and Doksum 2001; Gelman et al. 2014; Hero 2015; Johnson 2013; Kay 1993, 1998; Poor 1994).

Basic Concepts

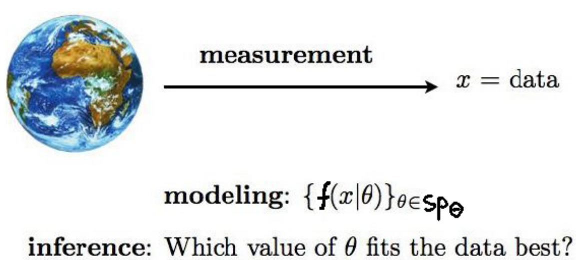


Figure 1: Statistical signal processing: Measurement, modeling and inference.

INGREDIENTS:

- x is measurement that we collect, modeled as a random variable X,
- θ is the true state of nature,
- *data model*, describing the probability distribution of X for a given θ. If, according to this model, X is a continuous random variable, then we specify the data model using $f_{X \mid \Theta}(x \mid \theta)$, the probability density function (pdf) of X given θ. If X is a discrete random variable, then we specify the data model using $p_{X \mid \Theta}(x \mid \theta)$, the probability mass function (pmf) of X given θ.

see the lectures by Prof. Tsitsiklis on discrete and continuous random variables

✻ EXAMPLES:

- continuous X: Gaussian;
- discrete X: Poisson, Bernoulli.

NOTE: If we decide to assign a probability distribution to $\boldsymbol{\theta}$, then we also need the prior pdf or pmf of θ (epistemic[1] probability): Prior pdf and pmf are denoted $f_\Theta(\theta)$ and $p_\Theta(\theta)$, respectively; we can also have a combined pdf–pmf prior.

[1] From the Greek words episteme (knowledge) and epistanai (to know or understand). Here, epistemic refers to our knowledge about the true state of nature.

☞ GOAL: Find the true state of nature θ, see Fig. 1.

COMMENTS

- `Detection`: Suppose that θ takes one of two possible values $\theta \in \{0, 1\}$ so that either $f_X(x \mid 1)$ or $f_X(x \mid 0)$ fit the data x the best. Hence, we need to decide if $f_{X \mid \Theta}(x \mid 1)$ or $f_{X \mid \Theta}(x \mid 0)$ is the better model.

- `Estimation`: Suppose that θ belongs to an infinite set. Then we must decide or choose among an infinite number of models. In this sense, estimation may be viewed as an extension of detection to infinite model classes. This extension presents many new challenges and issues and so it is given its own name.

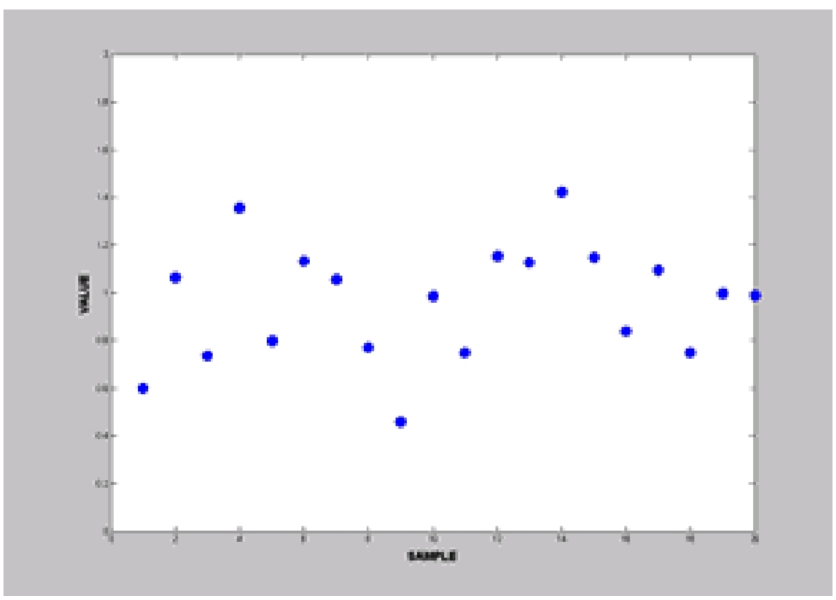


Figure 2: Measurements $x[n]$ as functions of the sample index n.

✻ ESTIMATION example: Assume that a finite data set $\boldsymbol{x} = (x[n])_{n=0}^{N-1}$ is available, depicted in Fig. 2. The measurements $\boldsymbol{x}$ depend on the parameter θ through a probabilistic model. An estimator of θ is a function of the data:

$$\hat{\theta}(\boldsymbol{x}).$$

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

Applications and Examples

APPLICATIONS:

- communications,
- radar and sonar (see Figs. 3 and 4),
- nondestructive evaluation (NDE),
- medicine,
- artifical intelligence and machine learning (Barber 2012; Bishop 2006; Murphy 2012), e.g., deep learning and neural networks[2] (Goodfellow et al. 2016).
- controls,
- seismology,
- econometrics (Hansen 2016).

[2] "One of them is in deep learning. And there, each "neuron" is really a cartoon. It's a linear-weighted sum that's passed through a nonlinearity. Anyone in electrical engineering would recognize those kinds of nonlinear systems. Calling that a neuron is clearly, at best, a shorthand. It's really a cartoon. There is a procedure called logistic regression in statistics that dates from the 1950s, which had nothing to do with neurons but which is exactly the same little piece of architecture," see (Gomes 2014)

Radar

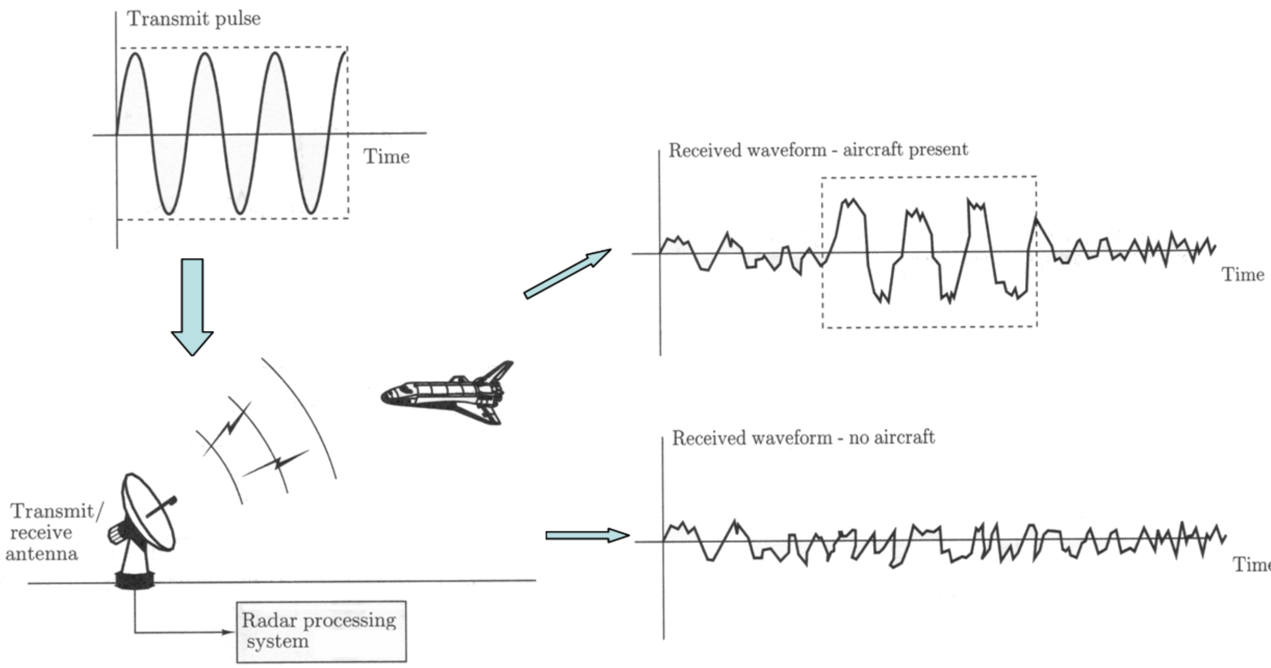


Figure 3: Radar system.

DETECTION: Decide on the presence or absence of an approaching target.

✻ ESTIMATION (after detection): Determine aircraft range as a function of time.

Sonar & ultrasound

ESTIMATION: Determine bearing of a target by processing signals collected by a hydrophone array, see Fig. 4. For more on array signal processing, see (Stoica and Moses 2005, §6).

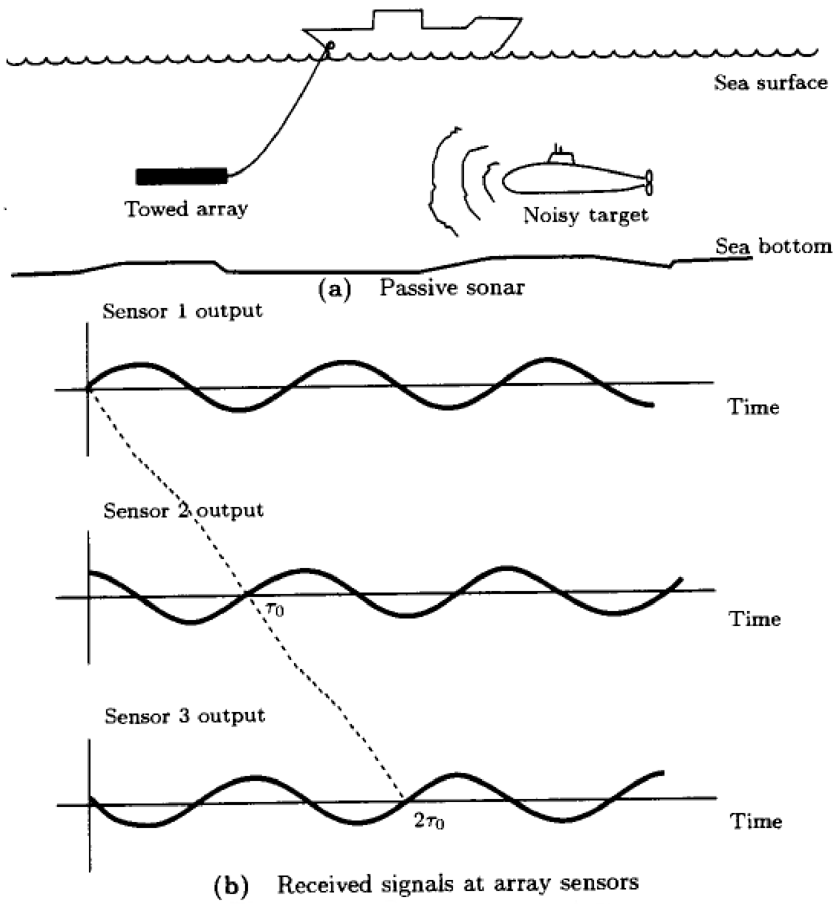


Figure 4: Passive sonar system.

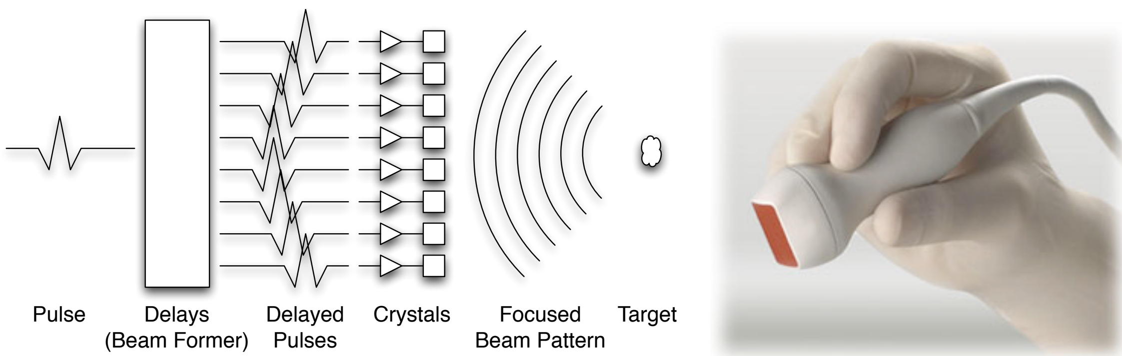


Figure 5: Ultrasound transducer.

Seismology

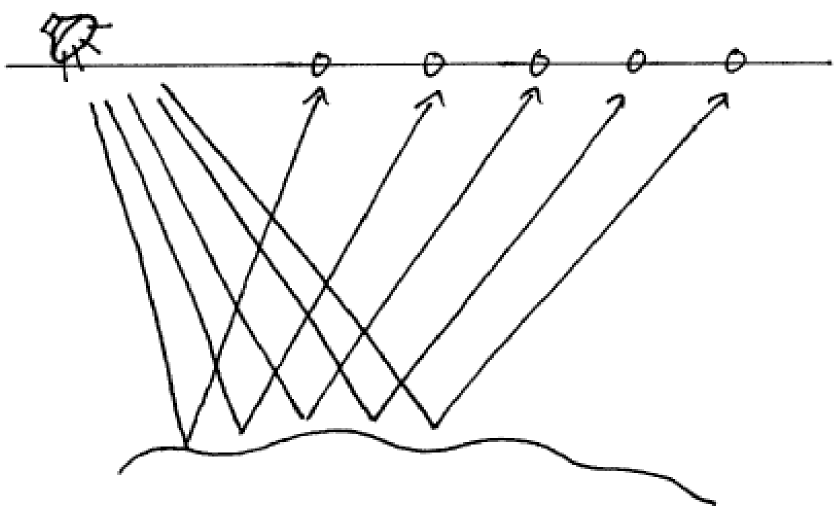


Figure 6: Seismic imaging.

Estimate depth below ground of an oil pool based on reflected acoustic waves, see Fig. 6.

Detection: Binary coherent communication

Consider a binary phase-shift keying (BPSK) communication system and denote by

$$\boldsymbol{s} = \big[s[0], s[1], \ldots, s[N-1]\big]^T$$

a digitized signal. A transmitter communicates a bit of information by sending $\boldsymbol{s}$ or $-\boldsymbol{s}$ for 1 or 0, respectively. The receiver measures a noisy version $\boldsymbol{x} = [x[0], x[1], \ldots, x[N-1]]^T$ of the transmitted signal, which we model as

$$(X[n])_{n=0}^{N-1} = \theta s[n] + W[n].$$

Notation:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = [a_1, a_2, \ldots a_N]^T$$

where "^{T}" denotes the transpose.

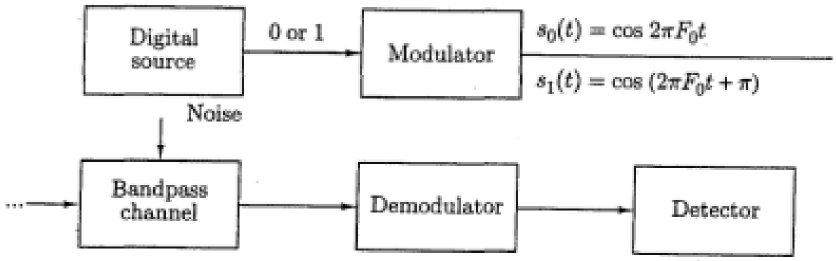


Figure 7: BPSK system.

The parameter θ is either $+1$ or -1, depending on which bit the transmitter is sending. Here, $W[n]$ represent errors incurred during the transmission process. So we have two models, or hypotheses, for the data:

$$\begin{aligned} H_0 : \quad & (X[n])_{n=0}^{N-1} = +s[n] + W[n] \\ H_1 : \quad & (X[n])_{n=0}^{N-1} = -s[n] + W[n]. \end{aligned}$$

Assume that all quantities are real-valued.

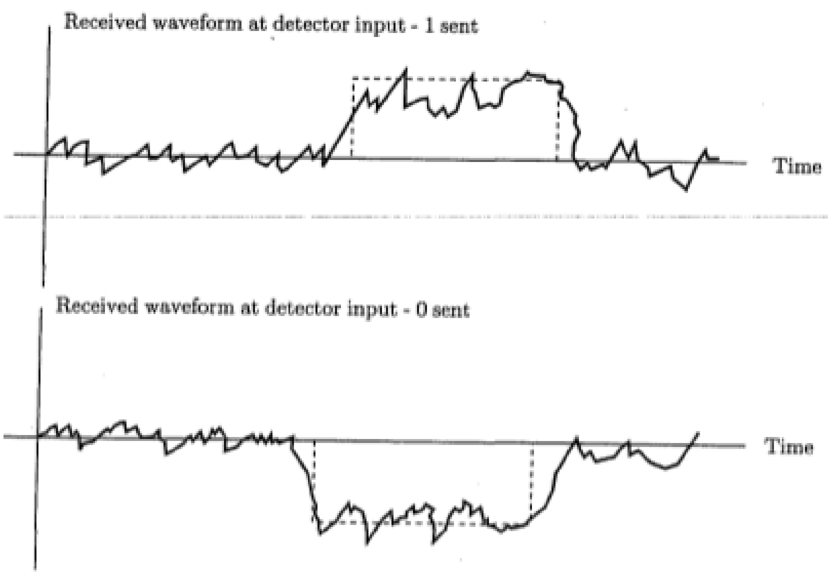


Figure 8: BPSK baseband waveforms.

Does $s[n]$ or $-s[n]$ match the measurements $x[n]$ and how well? This comparison can be made by computing a function of the data. Functions of data are called *statistics*. A natural statistic in this problem is the *correlation statistic*:

$$t(X) = \sum_{n=0}^{N-1} s[n]X[n] = \theta\Big(\sum_{n=0}^{N-1} s^2[n]\Big) + \Big(\sum_{n=0}^{N-1} s[n]W[n]\Big)$$

where $\boldsymbol{X} = (X[n])_{n=0}^{N-1}$.[3] If the errors are noise-like and do not resemble the signal $\boldsymbol{s}$, then $\sum_{n=0}^{N-1} s[n]W[n] \approx 0$ and a reasonable way to decide which value of the bit was sent is to decide that 0 was sent if $t(\boldsymbol{x}) < 0$ and that 1 was sent if $t(\boldsymbol{x}) > 0$. To quantify the performance of this test, we need a mathematical model for the errors $W[n]$.

[3] Note that $\boldsymbol{X}[n]$ are random variables introduced to model the measurements $x[n]$; hence, $x[n]$ are the realizations of these random variables. Later we will mix the two notations, but keep in mind the distinction.

Detection: DC signal absent or present

WE observe either $X = W$ (noise) or $X = \theta + W = 1 + W$ (signal+noise), where W is zero-mean Gaussian noise with variance σ^2:[4]

$$W \sim \mathcal{N}(0, \sigma^2).$$

[4] $\mathcal{N}(\mu, \sigma^2)$ is the shorthand notation for the Gaussian distribution with mean μ and variance σ^2.

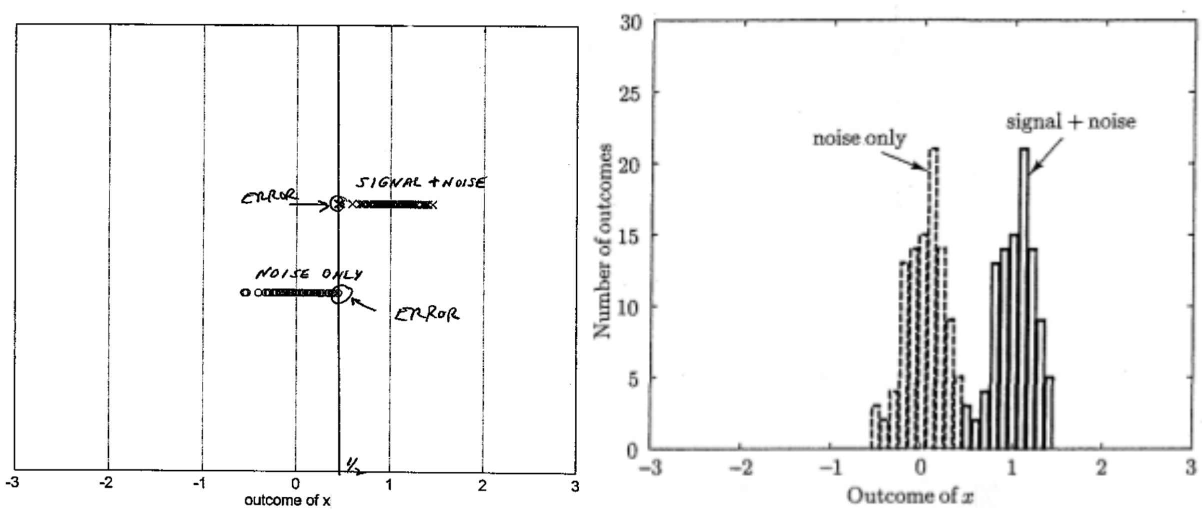


Figure 9: (Left) 100 realizations of X for signal present and absent (right) corresponding histograms for $\sigma^2 = 0.05$.

How to choose based on the observed value $X = x$?
Note:

$$\mathrm{E}(X) = \begin{cases} 0, & \text{if noise} \\ 1, & \text{if signal+noise} \end{cases}.$$

Detector chooses *signal present* if

$$x > 0.5.$$

Good detection? When will it be wrong? See Figs. 9–11.

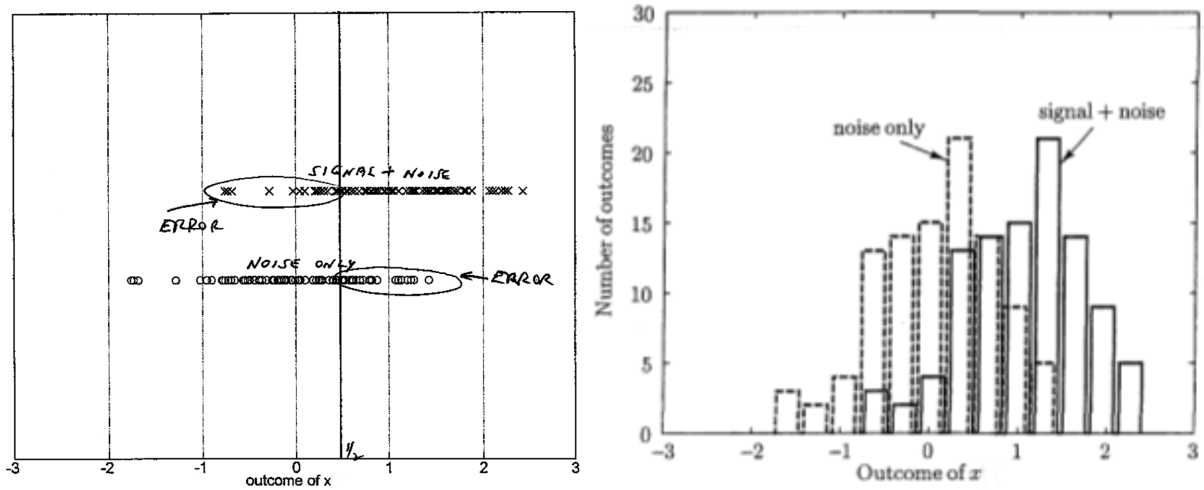


Figure 10: (Left) 100 realizations of X for signal present and absent (right) corresponding histograms for $\sigma^2 = 0.5$.

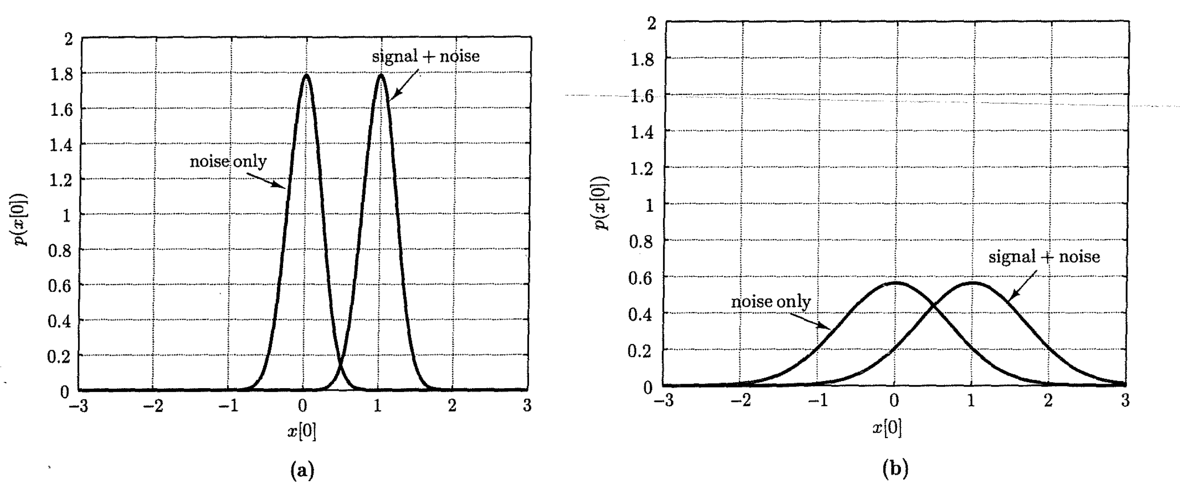


Figure 11: Pdfs of X for signal present and absent, for (left) $\sigma^2 = 0.05$ and (right) $\sigma^2 = 0.5$.

Parameter estimation: Sinusoids

SUPPOSE that we measure a sinusoidal signal in noise, modeled as

$$(X[n])_{n=0}^{N-1} = A\sin(\omega n + \varphi) + W[n]$$

where the amplitude A, frequency ω, and phase φ are unknown parameters to be estimated from the measurements $x[n]$ and $W[n]$ are noise/errors in the measurements. Note that here we assume that the signal has been sampled, which is most often the case in modern systems, but one could also pose a continuous-time version of this problem. For the most part we will focus on the discrete-time, sampled-data models in this course. In this case we have the set of parameters $\boldsymbol{\theta} = (A, \omega, \varphi)$. If we choose a probability distribution for the noise (say Gaussian), then we can write the probability distribution for our data $f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x} \mid \boldsymbol{\theta})$. Given the measurements $\boldsymbol{x}$, how do we estimate the values of the parameters?

Signal estimation

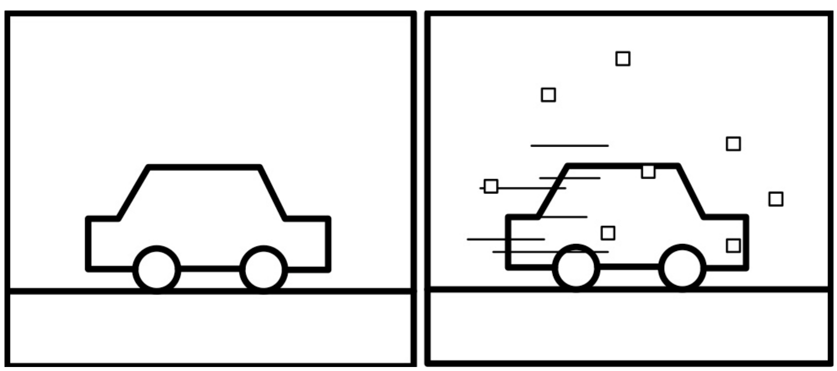


Figure 12: Image processing and reconstruction can involve complex estimation problems. We may have motion artifacts and noise, leading to a blurry and noisy image. Then, our goal is to "restore" the original image by deblurring and denoising.

AN imaging system does not produce perfect images: the measurements are distorted and noisy, see Fig. 12. You wish to restore the image. Assume that the distortion is a linear operation; then we can model the collected data by the following equation:

To save space and effort, the capital (random model) and lowercase (measurement) versions are merged: we use only the measurement version.

$$\boldsymbol{x} = H\boldsymbol{s} + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{s}$ is the ideal image we wish to recover (represented as a vector, each element of which is a pixel), H is model of the distortion (represented as a matrix), and $\boldsymbol{w}$ is a vector of noise.

For example, $\boldsymbol{s}$ and $\boldsymbol{x}$ may be the left and right images in Fig. 13.

Figure 13: An image and its distorted version obtained by spatial filtering with a Gaussian function.

Signal denoising

A SPECIAL CASE of signal estimation:

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{w}.$$

A special case of (1) with $\boldsymbol{H} = \boldsymbol{I}$.

Likelihood function and ML estimation

AN IMPORTANT CONCEPT: $f_{X \mid \Theta}(x \mid \theta)$ viewed as a function of θ is the *likelihood function*.

Comments on the likelihood function:

- For a given θ and discrete model distribution, the pmf $p_{X \mid \Theta}(x \mid \theta)$ is the probability of observing the data point x. In the continuous model distribution, the pdf $f_{X \mid \Theta}(x \mid \theta)$ is approximately proportional to probability of observing a point in a small rectangle around x.

- However, when we think of $f_{X \mid \Theta}(x \mid \theta)$ or $p_{X \mid \Theta}(x \mid \theta)$ as a function of θ, it provides, for a given observed x, the *likelihood* or *plausibility* of various θs.

Maximum-likelihood (ML) estimation: Maximize the likelihood with respect to θ, i.e.,

$$\widehat{\theta} = \arg\max_{\theta} f_{X \mid \Theta}(x \mid \theta).$$

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

Bayesian inference

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

$$f_{\Theta \mid X}(\theta \mid x) = \frac{f_{X \mid \Theta}(x \mid \theta) f_{\Theta}(\theta)}{\int f_{X \mid \Theta}(x \mid \vartheta) f_{\Theta}(\vartheta) \, d\vartheta}. \tag{2}$$

Bayes' rule, continuous-parameter version. See the lecture "Bayes rule variations" by Prof. Tsitsiklis.

☞ NOTE: $f_{\Theta \mid X}(\theta \mid x)$ is an epistemic probability.

✻ COMMON Bayesian estimators:

- maximum *a posteriori* (MAP):

$$\widehat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} f_{\Theta \mid X}(\theta \mid x); \tag{3}$$

- minimum mean-square error (MMSE):

$$\hat{\theta}_{\mathrm{MMSE}}(x) = \mathrm{E}_{\Theta \mid X}\big(\Theta \mid X = x\big). \tag{4}$$

COMMENTS: MAP estimation is typically the most tractable because it does not require computing the denominator in (2), which is usually analytically intractable. (Note that the denominator is not a function of $\boldsymbol{\theta}$.) So,

$$\widehat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} f_{X \mid \Theta}(x \mid \theta) f_{\Theta}(\theta). \tag{5}$$

The MMSE estimator is derived by minimizing the Bayesian mean-square error (BMSE):[5]

[5] for real-valued random variables

$$\mathrm{BMSE}\{\hat{\theta}\} = \mathrm{E}_{X,\Theta}\big(\big[\hat{\theta}(X) - \Theta\big]^2\big) \tag{6}$$

with respect to the estimator function $\hat{\theta}(\cdot)$.

✻ TIKHONOV-TYPE regularization versus MAP. In machine learning and sparse and low-rank signal processing, we often encoounter penalized minimization problems of the following form: minimize

$$f_u(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + u\, r(\boldsymbol{\theta}) \tag{7}$$

with respect to $\boldsymbol{\theta}$. Here, $\mathcal{L}(\boldsymbol{\theta})$ is the data-fidelity term or, using detection and estimation terminology, the negative log-likelihood (NLL):

$$\mathcal{L}(\boldsymbol{\theta}) = -\ln f_{\boldsymbol{X} \mid \boldsymbol{\Theta}}(\boldsymbol{x} \mid \boldsymbol{\theta}) \tag{8}$$

where we often add a constant to $\ln f_{X \mid \Theta}(x \mid \theta)$ for numerical stability, e.g., so that $\mathcal{L}(\boldsymbol{\theta}) \geq 0$, with minimum at zero.

Here, $u > 0$ is a *regularization tuning constant* that quantifies the strength of the *regularization term* $r(\boldsymbol{\theta})$. In many cases, we may interpret $u\, r(\boldsymbol{\theta})$ as the negative logarithm of the prior pdf[6] or pmf or, more

[6] $-\ln f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$, see (5)

precisely, a family of priors parameterized by u. Selecting u is a challenging problem; see (Pereyra et al. 2015) for an interesting solution to this problem that takes advantage of the MAP interpretation of the penalized NLL minimization in (7).

☞ POPULAR assumption: $f_u(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$, which requires both $\mathcal{L}(\boldsymbol{\theta})$ and $r(\boldsymbol{\theta})$ to be convex functions of $\boldsymbol{\theta}$. See (Bertsekas 2015) and references therein for descriptions of algorithms for solving convex optimization problems.

Bayesian versus classical inference

IN classical (non-Bayesian) analysis, inference is made based *only* on the likelihood function (probabilistic model):

$$f_{X\,|\,\Theta}(x\,|\,\theta) \quad \text{or} \quad p_{X\,|\,\Theta}(x\,|\,\theta).$$

☞ CRITICISM against Bayesian approach:

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

☞ CRITICISM against classical approach:

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

Model and Identifiability

Model

A model is a parametrized pdf or pmf $f_{X\,|\,\Theta}(x\,|\,\theta)$.

Example: DC level in Gaussian noise

$$X = \underbrace{\theta}_{\text{parameter}} + \underbrace{W}_{\text{noise}} \tag{9}$$

$W \sim \mathcal{N}(0, \sigma^2)$

leading to

$$X \sim \mathcal{N}(\theta, \sigma^2)$$

or

$$f_{X\,|\,\Theta}(x\,|\,\theta) = \mathcal{N}(x \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x-\theta)^2\Big].$$

see handout `probability distributions` on cybox and lectures by Prof. Tsitsiklis on Gaussian random variables

✻ Note: The textbook uses $\theta = A$ to denote the DC level, see (Kay 1993, §1.3).

Identifiability

An important property of a model structure is *parameter identifiability*:

$$f_{X\,|\,\Theta}(\cdot\,|\,\theta_1) = f_{X\,|\,\Theta}(\cdot\,|\,\theta_2) \quad \Longleftrightarrow \quad \theta_1 = \theta_2. \tag{10}$$

Here, $f_{X\,|\,\Theta}(\cdot\,|\,\theta_1) = f_{X\,|\,\Theta}(\cdot\,|\,\theta_2)$ means that the left and right sides are the same functions of x.

✻ 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 mention in this class[7]. But, this needs to be done carefully.

[7] The parameter-expanded EM (PX-EM) algorithm is one such example. We will introduce expectation-maximization (EM) algorithms later in this class.

Examples

Estimation: DC level in white Gaussian noise

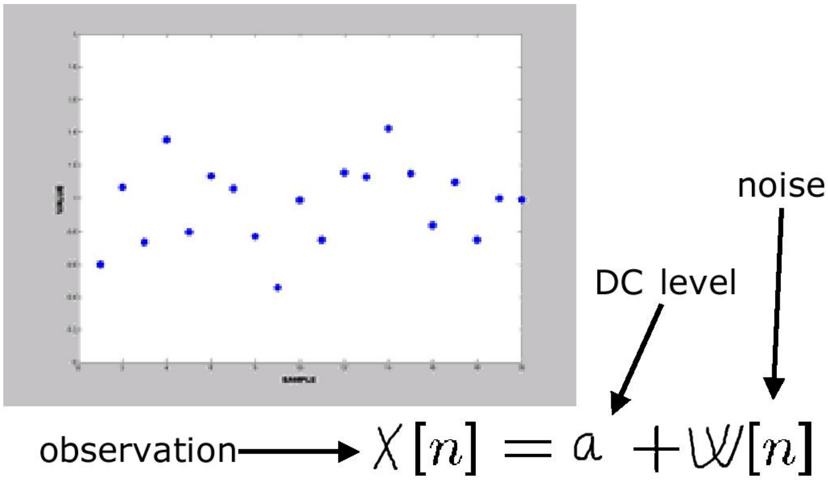
Choose white Gaussian noise model:

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

where $\boldsymbol{w} = (W[n])_{n=0}^{N-1}$ are independent, identically distributed (i.i.d.), and, therefore,

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

Hence, $X[n]$ are i.i.d. Gaussian $\mathcal{N}(a, \sigma^2)$. Why? We will review in handout `revprob`.



For simplicity, assume first that the noise level σ^2 is known. Statistical inference is based on the likelihood function:

$$f_{\boldsymbol{X}\,|\,A}(\boldsymbol{x}\,|\,a) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-a)^2\Big]. \tag{11a}$$

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

$$f_{\boldsymbol{X}\,|\,\boldsymbol{\Theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-a)^2\Big] \tag{11b}$$

where $\boldsymbol{\theta} = (a, \sigma^2)$. What is the difference between (11a) and (11b)?

Classical estimation theory from a signal-processing point of view is covered in detail in (Kay 1993, §2–9).

Bayesian inference of DC level in white Gaussian noise

SUPPOSE that the DC level A is known to be within the interval $[0.5, 1.5]$ with all values equally likely in this interval, see Fig. 14. Therefore, we set a prior distribution for A as

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

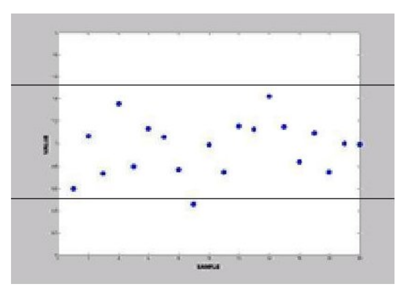


Figure 14: Prior information about the DC level.

Again, assume that σ^2 is known. Then, Bayesian inference is based on the posterior distribution of A:

$$\begin{aligned} f_{A\,|\,X}(a\,|\,x) &= \frac{f_{X\,|\,A}(x\,|\,a)\, f_A(a)}{\displaystyle\int f_{X,A}(x,\alpha)\,\mathrm{d}\alpha} \\ &= \frac{f_{X\,|\,A}(x\,|\,a)\, f_A(a)}{\displaystyle\int f_{X\,|\,A}(x\,|\,\alpha)\, f_A(\alpha)\,\mathrm{d}\alpha} \end{aligned}$$

$$= \frac{\exp\left[-\sum_{n=0}^{N-1}(x[n]-a)^2/(2\sigma^2)\right] f_A(a)}{\int_{0.5}^{1.5} \exp\left[-\sum_{n=0}^{N-1}(x[n]-\alpha)^2/(2\sigma^2)\right] \mathrm{d}\alpha}.$$

Recall the common Bayesian estimators: MAP in (3) and MMSE in (4). Bayesian estimation from a signal-processing point of view is covered in detail in (Kay 1993, §10–13).

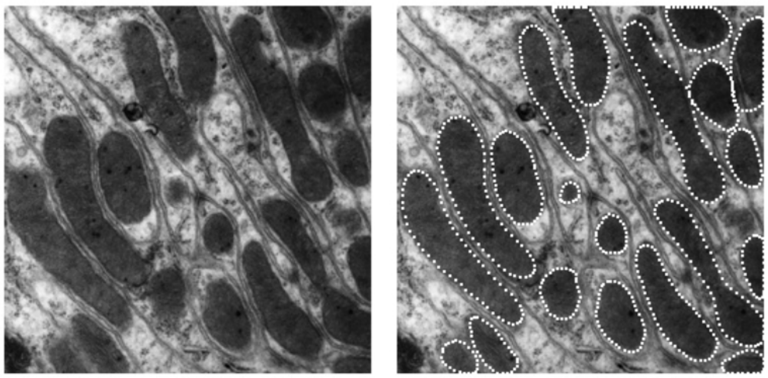


Figure 15: Mitochondria segmentation.

Examples of MCMC Bayesian inference

MITOCHONDRIA segmentation. Here, the data $\boldsymbol{x}$ are the electron micrograph of a cardiac muscle cell (Grenander and Miller 1994), see Fig. 15. The parameter vector $\boldsymbol{\theta}$ contains:

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

The prior distribution $f_\Theta(\theta)$ is learned from hand-segmented training data, using several hundred hand-selected electron micrographs.

IDEA: Mitochondria and cytoplasm have different textures; we can use Markov random field (MRF) models to learn the texture from training data. Then, we use Markov chain Monte Carlo (MCMC) techniques to draw samples from the posterior distribution $f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$.

✻ ANOTHER Bayesian MCMC example. See Fig. 16 and (Dogandžić and Zhang 2007).

Acronyms

BMSE Bayesian mean-square error. 12

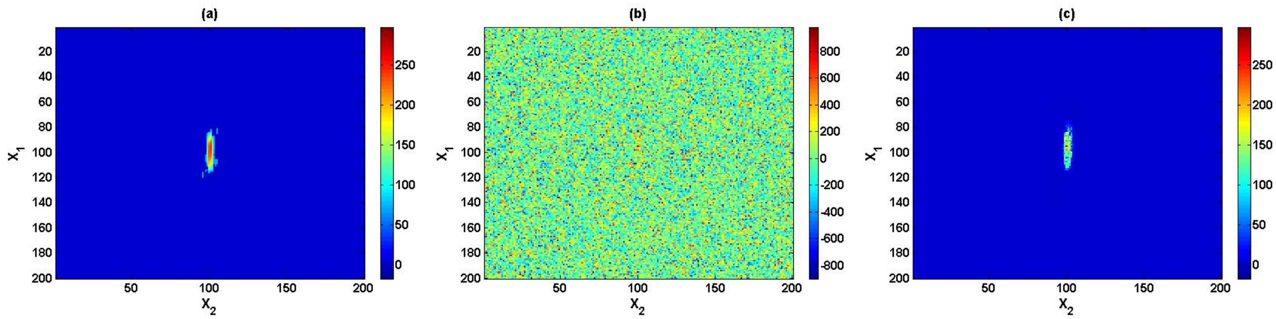


Figure 16: MCMC defect identification in NDE images.

BPSK binary phase-shift keying. 6, 7

CRB Cramér-Rao bound. 11

EM expectation-maximization. 14

i.i.d. independent, identically distributed. 14

MAP maximum *a posteriori*. 12, 13, 15

MCMC Markov chain Monte Carlo. 16

ML maximum-likelihood. 11

MMSE minimum mean-square error. 12, 15

MRF Markov random field. 16

NDE nondestructive evaluation. 4, 16

NLL negative log-likelihood. 12, 13

pdf probability density function. 2, 3, 9, 11–13

pmf probability mass function. 2, 3, 11–13

PX-EM parameter-expanded EM. 14

References

Barber, David (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press. URL: `http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/171216.pdf` (visited on 01/07/2017).

Bertsekas, Dimitri P. (2015). *Convex Optimization Algorithms*. Belmont, MA: Athena Scientific.

Bickel, Peter J. and Kjell A. Doksum (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Dogandžić, Aleksandar and Benhong Zhang (2007). "Markov chain Monte Carlo defect identification in NDE images". In: *Rev. Prog. Quant. Nondestr. Eval.* Ed. by Don O. Thompson and Dale E. Chimenti. Vol. 26. AIP Conf. Proc. Melville, NY, pp. 709–716.

Gelman, A., J. B. Carlin, H. S. Stern, David B. Dunson, Aki Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Taylor & Francis.

Gomes, Lee (2014). *Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts*. URL: `http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts` (visited on 01/07/2017).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press. URL: `http://www.deeplearningbook.org/`.

Grenander, Ulf and Michael I. Miller (1994). "Representations of knowledge in complex systems". In: *J. Roy. Stat. Soc. B* 56.4, pp. 549–603.

Hansen, Bruce E. (2016). *Econometrics*. URL: `http://www.ssc.wisc.edu/~bhansen/econometrics/`.

Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI.

Johnson, Don H. (2013). *Statistical Signal Processing*. Lecture notes. Rice Univ., Houston, TX.

Kay, Steven M. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall.

– (1998). *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ: Prentice Hall.

Murphy, Kevin Patrick (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Pereyra, Marcelo, José M Bioucas-Dias, and Mário A. T. Figueiredo (2015). "Maximum-a-posteriori estimation with unknown regularisation parameters". In: *Proc. European Signal Process. Conf.* Nice, France, pp. 230–234.

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

Stoica, Petre and Randolph Moses (2005). *Spectral Analysis of Signals*. Upper Saddle River, NJ: Pearson. URL: `http://user.it.uu.se/~ps/SASbook.zip`.