13. Parameter Estimation

ECE 830, Spring 2014

Primary Goal

General problem statement:

We observe

$$X \sim p(x|\theta), \theta \in \Theta$$

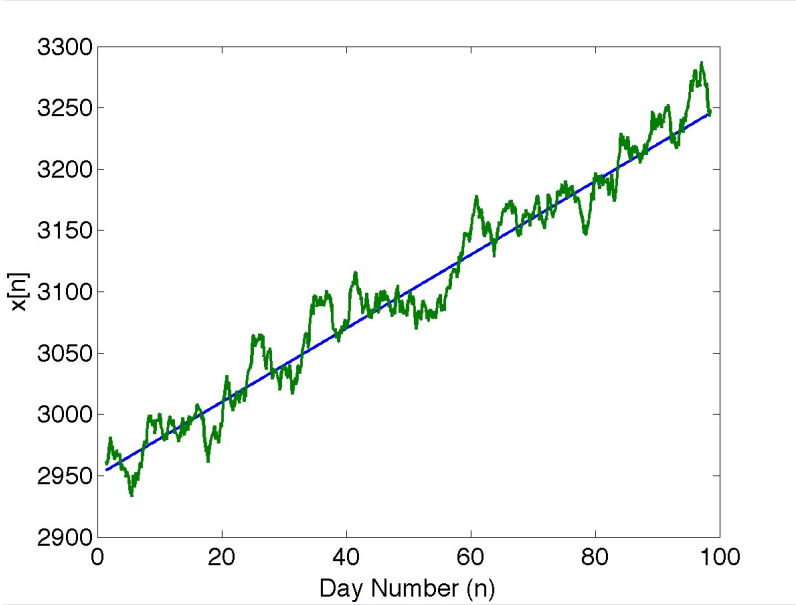
and the goal is to determine the θ that produced X.

Given a collection of observations $x_1, ..., x_n$ and a probability model

$$p(x_1, ..., x_n|\theta)$$

parameterized by the parameter θ, determine the value of θ that best matches the observations.

Example: Stock Market (Dow-Jones Industrial Avg.)



Based on this plot we might conjecture that the data is "on average" increasing. Probability model:

$$x[k] = A + Bk + w[k]$$

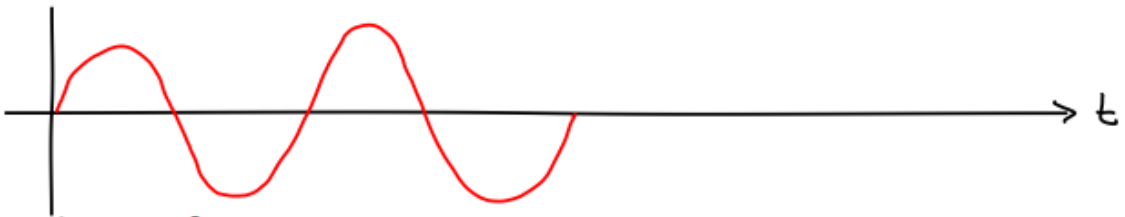
A, B unknown parameters, $w[k]$ white Gaussian noise to model fluctuations.

$$p(x|A, B) =$$

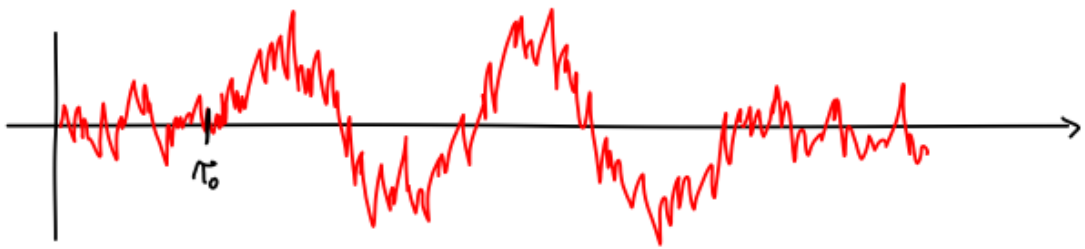
Example: A Classic Example – Radar

Radar example:

Transmit pulse:



Received waveform:



$\tau_0 = 2d/c$,
c = speed of propagation

estimate $\tau_0 \to \hat{\tau}_0$

$\hat{d} = \frac{c\hat{\tau}_0}{2}$

Example: A Classic Example – Radar (cont.)

The received waveform is time-dilated and shifted version of original waveform $g(t)$ plus noise

$$x(t) = g(\alpha t - \tau) + w(t).$$

The parameter of interest is $\theta = \begin{bmatrix} \alpha \\ \tau \end{bmatrix}$, where α is related to velocity / Doppler shift, τ is related to distance

$$\tau = \frac{2d}{c}, \qquad c = \text{speed of light}$$

$$x \to \hat{\tau} \to \hat{d} = \frac{c\hat{\tau}}{2}$$

Example: A modern Example - Imaging

Image processing can involve complicated estimation problems. For example, suppose we observe a moving object with noise. This image is blurry and noisy and our goal is to restore this image by debarring and denoising.

Example: A modern Example - Imaging

We can model the moving part of the observed image as

$$x = \underbrace{h * \theta}_{\text{motion blur (convolution)}} + \underbrace{w}_{\text{noise}}$$

where the parameter of interest θ is the ideal image.

Observation model:

$$\begin{aligned} X =& H\theta + w, \qquad w \sim N(0, \sigma^2 I) \\ X \sim& \mathcal{N}(H\theta, \sigma^2 I) \end{aligned}$$

Estimator:

$$\hat{\theta} = f(x)$$

a function of x

Basic Ingredients of Estimation Theory

- Observations (data):
$$x = [x_1, ..., x_n]^T$$
- A observation probability model parameterized by θ:
$$p(x|\theta), \theta \in \Theta, x \in \mathcal{X}$$
- A class or collection of possible parameter values Θ: e.g., $\mathbb{R}$, {linear estimators}, {unbiased estimators}
- An estimator $\hat{\theta} : \mathcal{X} \to \Theta$, a mapping from $\mathcal{X}$ to Θ
- A loss/error function $\ell : \Theta \times \Theta \to \mathbb{R}^+$. $\ell(\theta, \hat{\theta})$ measures proximity of $\hat{\theta}$ to θ
- Risk (average/expected loss)

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}[\ell(\theta, \hat{\theta}(x))] \\ &= \int_{\mathcal{X}} \ell(\theta, \hat{\theta}(x))\, p(x|\theta) dx \end{aligned}$$

Optimal Estimator

$$\hat{\theta}_{\text{opt}} = \arg \min_{\hat{\theta}:\mathcal{X}\in\Theta} R(\theta, \hat{\theta})$$

$\hat{\theta}_{\text{opt}}$ is optimal with respect to the chosen loss function. **Example losses include:**

- Squared Error (ℓ_2 loss)

$$\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2$$

- Absolute Error (ℓ_1 loss) (penalizes large errors less than ℓ_2)

$$\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1 = \sum_{i=1}^{n} |\theta_i - \hat{\theta}_i|$$

- 0/1 loss

$$\ell(\theta, \hat{\theta}) = \begin{cases} 1, & \hat{\theta} \neq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$R(\theta, \hat{\theta}) = \mathbb{E}[1_{\{\hat{\theta}(x)\neq\theta\}}] = P(\hat{\theta} \neq \theta)$$

Special Case - Hypothesis Testing:

Hypothesis testing can be viewed as a special case in which the parameter θ takes one of two possible values, 0 or 1, and the loss is $0/1$ loss or some weighted version of it.

Terminology in Estimation Theory

Define

$$\epsilon(\widehat{\theta}) := \widehat{\theta} - \theta$$

Recall that $\widehat{\theta} = \widehat{\theta}(x)$ is a function of data $\Longrightarrow \epsilon(\widehat{\theta})$ is a statistic!
Mean Squared Error:

$$\mathbb{E}[\epsilon^T \epsilon] := \mathsf{MSE}(\widehat{\theta})$$

Bias:

$$\mathsf{Bias}(\widehat{\theta}) := \mathbb{E}[\widehat{\theta}] - \theta$$

Variance / Covariance:

$$\mathsf{Var}(\widehat{\theta}) := \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^T]$$

Bias-variance decomposition

Constraints

Often we are interested in estimators with special properties and/or estimators that satisfy certain constraints.

Example: Minimum Variance, Unbiased Estimator

Let θ be an unknown parameter and let $\widehat{\theta}$ be an estimator of θ given the data x.

Choose the estimate $\widehat{\theta}$ such that the Bias $= 0$ and the variance is as small as possible.

Example: Best Linear Unbiased Estimator

Often we would like to keep our estimator simple. One common constraint is to find the best estimator that is linear in the data, unbiased and minimizes variance among all linear functions of the data.

Asymptotics

Estimators are often studied as a function of the number of observations:

$$\widehat{\theta}(x) = \widehat{\theta}_n \text{ where } n = \dim x$$

$\widehat{\theta}$ is asymptotically unbiased if

$$\lim_{n\to\infty} \text{Bias}(\widehat{\theta}_n) =$$

An estimator is consistent if

$$\lim_{n\to\infty} \text{MSE}(\widehat{\theta}_n) =$$

A consistent estimator is *at least* asymptotically unbiased. Some estimators are unbiased, but inconsistent.

The latter basically means that our estimation does not improve as the number of data increase. Inconsistent estimators can provide reasonable estimates when we have a small number of data. However, consistent estimators are usually favored in practice.

Prior information

In the problem formulation above, we do not make any assumptions on what θ might be; we simply minimize the cost function over all allowable possibilities. But suppose that we believe that some values of θ are more likely than others.

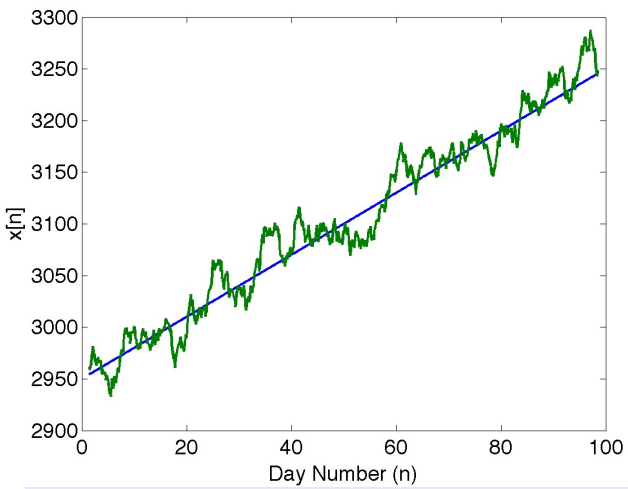
We can reflect this prior belief by viewing the parameter θ as a random variable itself with a prior density $p(\theta)$

This leads to the joint density function:

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

We can specify cost functions based on this joint density. The prior $p(\theta)$ effectively steers or directs our search in what we believe are the right directions.

Example: Stock Market



We might believe that $2800 < A < 3200$. We can incorporate this belief with

Example: Getting our feet wet

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$$

$$\hat{\mu}_n = \hat{\mu}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- ℓ_2 loss: $\ell(\mu, \hat{\mu}) = \|\mu - \hat{\mu}_n\|_2^2$
- MSE risk:

$$\begin{aligned} R(\mu, \hat{\mu}) &= \mathbb{E}[(\mu - \hat{\mu}_n)^2] \\ &= \underbrace{\mathbb{E}[(\mu - \mathbb{E}[\hat{\mu}_n])^2]}_{\text{bias}^2} + \underbrace{\mathbb{E}[(\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n])^2]}_{\text{variance}} \end{aligned}$$

- bias: $\mathbb{E}[\hat{\mu}_n] =$

Example: Getting our feet wet (cont.)

- variance:

- consistency: