ASEN 5044, Fall 2018

Statistical Estimation for Dynamical Systems

Note for Fall 2024: You may want to wait to watch or perhaps rewatch this lecture after Lecture 21, since we are a little bit behind as of Lecture 19.

Lecture 20 [Special Topic #4]: Introduction to Maximum Likelihood and Bayesian Point Estimation Theory

Prof. Nisar Ahmed (Nisar.Ahmed@Colorado.edu)

Friday 10/19/2018

University of Colorado Boulder

Aerospace Engineering

Overview

Introduce alternative criteria for non-deterministic estimation, which are based on probabilistic modeling and are more general than least-squares

- Maximum likelihood point estimation
- Bayesian point estimation
- Focus on static state/parameter estimation for now

General Problem Setup for Static State/Parameter Estimation

- Consider unknown static state (or model parameter) x with measurements $y_k = h_k(x,v_k)$ where $v_k \sim p(v_k)$ is some unobserved measurement error

- Find best guess/optimal estimate for x from i.i.d. data set $y_1,\ldots,y_T$

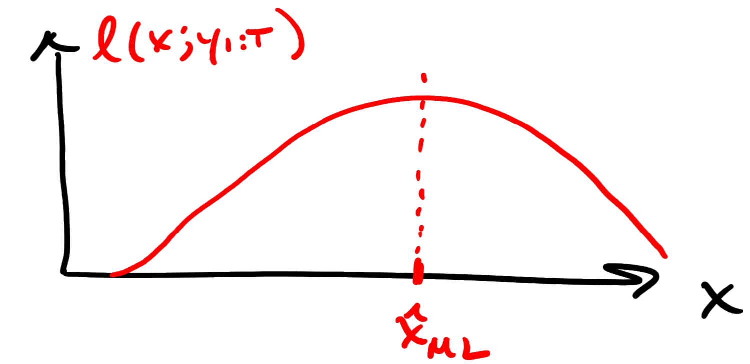
- Note: not restricted to linear $h_k(x)$ or AWGN for v_k – can have arbitrary dependencies between x and y_k, arbitrary pdfs for uncertainties/noise

- x may or may not be random, but y_k is always assumed random

Maximum Likelihood Point Estimators

- Popularized by Sir Ronald Fisher in the 1920's
- Assume that x must be some **non-random**, but unknown constant
- <u>Principle of Maximum Likelihood:</u> optimal estimate of x is value that makes observed $y_{1:T}$ most probable, i.e. the value of x which maximizes the so-called **likelihood score**

$$\ell(x; y_{1:T}) \triangleq p(y_{1:T} \mid x) \underset{\text{for iid } y_k}{=} \prod_{k=1}^{T} p(y_k \mid x)$$



$$\rightarrow \hat{x}_{ML} = \underset{x \in \mathbb{R}^n}{\arg\max}\ \ell(x; y_{1:T})$$

- Often work with maximizing **log-likelihood score** instead, to make math easier:

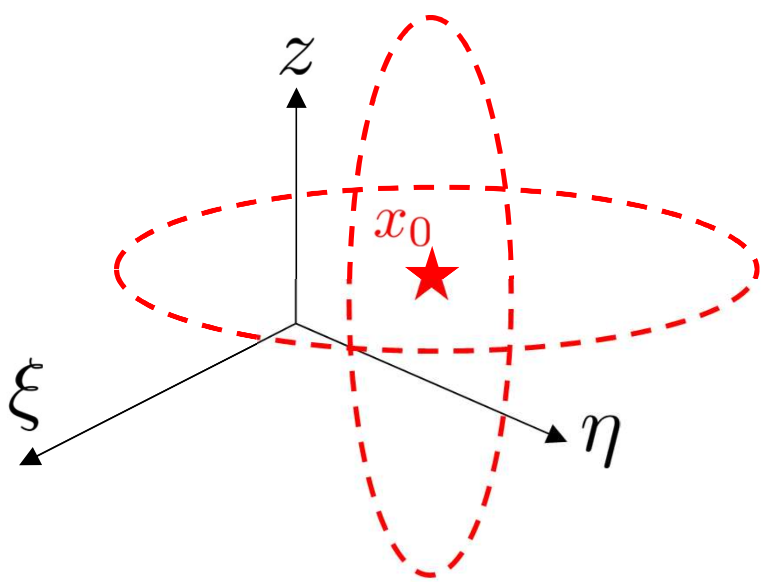
$$\mathcal{L}(x; y_{1:T}) = \log \ell(x; y_{1:T}) = \log p(y_{1:T} \mid x) = \sum_{k=1}^{T} \log p(y_k \mid x)$$

b/c log is a convex fxn (preserves local maxima)

$$\rightarrow \hat{x}_{ML} = \underset{x \in \mathbb{R}^n}{\arg\max}\ \mathcal{L}(x; y_{1:T})$$

Example 1: Maximum Likelihood Positioning

- Suppose we set up for 3D positioning problem



$$x(k) = \begin{bmatrix} \xi(k) \\ \eta(k) \\ z(k) \end{bmatrix} \quad \begin{pmatrix} \text{Easting} \\ \text{Northing} \\ \text{height} \end{pmatrix}$$

$$x(k+1) = x(k) = \text{const.} = \begin{bmatrix} \xi(0) \\ \eta(0) \\ z(0) \end{bmatrix} = x_0$$

$$y(k+1) = \underbrace{x(k+1) + v(k+1)} \longleftrightarrow H = I_{3\times 3}$$

$$v(k+1) \sim \mathcal{N}(0, R)$$

$$p(y_k | x) = N(x_0, R) = \frac{1}{(2\pi)^{\frac{3}{2}} |R|^{1/2}} \cdot \exp\left\{ -\frac{1}{2} [y_k - x_0]^T R^{-1} [y_k - x_0] \right\}$$

Example 1: Maximum Likelihood Positioning

- To find maximum likelihood estimate $\hat{x}_{\mathrm{ML}} = \arg\max_{x \in \mathbb{R}^n} \mathcal{L}(x_0; y_{1:T})$

→ log-likelihood score: $\mathcal{L}(x; y_{1:T}) = \log p(y_{1:T} | x_0) = \sum_{k=1}^{T} \log p(y_k | x_0)$, where $p(y_k|x_0) = \mathcal{N}(x_0, R)$

→ $\log p(y_k|x_0) = \log \left[\frac{1}{(2\pi)^{\frac{n}{2}} |R|^{\frac{1}{2}}} \cdot \exp\left\{ -\frac{1}{2} [y_k - x_0]^T R^{-1} [y_k - x_0] \right\} \right]$

$= \underset{(\perp \text{ of } x_0)}{\text{const.}} - \frac{1}{2} [y_k - x_0]^T R^{-1} [y_k - x_0] = \log p(y_k|x_0)$, $k = 1, \dots, T$

→ So: $\mathcal{L}(x; y_{1:T}) = \underset{(\perp\, x_0)}{\text{const.}} - \frac{1}{2} \sum_{k=1}^{T} [y_k - x_0]^T R^{-1} [y_k - x_0]$ → looks like LS estimator cost fxn!

→ $\boxed{\hat{x}_{ML} = \underset{(\text{show lec 21})}{\hat{x}_{LS}} = (H^T R^{-1} H)^{-1} \cdot H^T R^{-1} \vec{y}}$ $\left(\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}, H = \begin{bmatrix} I \\ \vdots \\ I \end{bmatrix}, \dots \right)$ See lec 19

→ if $R_1 = R_2 = \dots = R_T = R$ → $\boxed{\hat{x}_{ML} = \frac{1}{T} \sum_{k=1}^{T} y_k}$ (mean of meas. vectors)

→ LS is a special case of maximum likelihood

→ Max likelihood can handle more complex noise & measurement models!

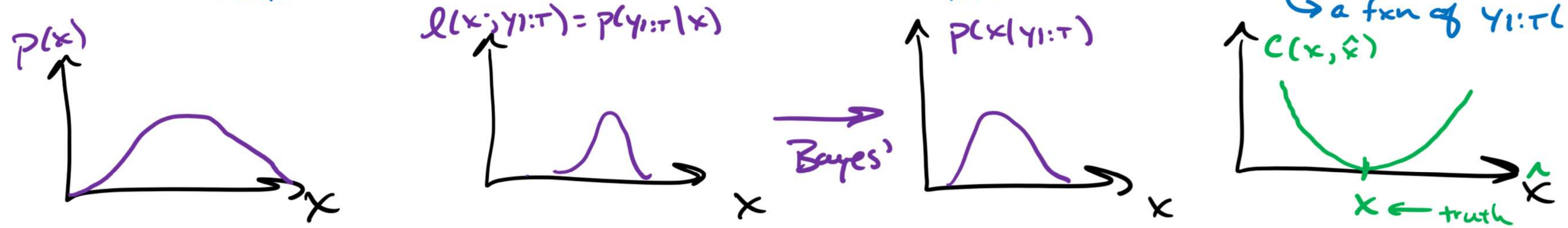
$p(v_k) = \mathcal{U}[a, b]$ or mixture model || y_k = nonlinear fxn of x (range &/or bearing from origin)

Bayesian Point Estimation

- Now suppose x is a **random variable**, with some prior p(x)
- What if estimate $\hat{x}$ ought to instead mitigate cost of making a mistake?
- Suppose we assume a **cost function** $C(x,\hat{x})$ for guessing $\hat{x}$ when true value is in fact x
- Since x is never available in practice, we should **minimize the expected value of** $C(x,\hat{x})$ **in light of whatever data** $y_{1:T}$ **is available**
- That is: pick $\hat{x}$ to **minimize the conditional expectation** of $C(x,\hat{x})$ w.r.t. $p(x|y_{1:T})$:

(*) $$\hat{x}_* = \underset{x\in\mathbb{R}^n}{\arg\min}\; E[C(x,\hat{x})\,|\,y_{1:T}] = \underset{x\in\mathbb{R}^n}{\arg\min} \int_{-\infty}^{\infty} C(x,\underline{\hat{x}})\,p(x|y_{1:T})\,dx$$

↳ a fxn of $y_{1:T}$ (observed)



- "Bayesian": find/take expectation w.r.t. posterior pdf $p(x|y_{1:T}) \propto p(x)\; p(y_{1:T}|x)$

Bayesian Minimum Mean Squared Error (MMSE) Estimation

- Many possible choices for $C(x, \hat{x})$
- One very popular choice is the **square error:** $C(x,\hat{x}) = (x-\hat{x})^T(x-\hat{x}) = ||x-\hat{x}||^2$
- This leads to the so-called minimum mean squared error (MMSE) estimate $\hat{x}_{\mathrm{MMSE}}$

$$\hat{x}_{\mathrm{MMSE}} = \arg\min_{x\in\mathbb{R}^n} E[(x-\hat{x})^T(x-\hat{x}) \mid y_{1:T}]$$

- Given some $p(x|y_{1:T})$, then what does $\hat{x}_{\mathrm{MMSE}}$ correspond to?

$$E[(x-\hat{x})^T(x-\hat{x}) \mid y_{1:T}] = E[x^Tx - 2\hat{x}^Tx + \hat{x}^T\hat{x} \mid y_{1:T}]$$

$$= E[x^Tx \mid y_{1:T}] - 2\hat{x}^T \cdot E[x \mid y_{1:T}] + \hat{x}^T\hat{x}$$

to find opt. $\hat{x}$, take deriv. w.r.t. $\hat{x}$ & set = 0

$$\left.\frac{\partial(\cdot)}{\partial\hat{x}}\right|_{\hat{x}_*} = 0 = 0 - 2\cdot E[x \mid y_{1:T}] + 2\hat{x}_* \rightarrow 0 = 2\hat{x}_* - 2\cdot E[x \mid y_{1:T}]$$

$$\rightarrow \boxed{\hat{x}_* = E[x \mid y_{1:T}] = \hat{x}_{\mathrm{MMSE}}} \rightarrow (*) \text{ true for any PDF } p(x \mid y_{1:T})!$$

Conditional mean of x given $y_{1:T}$ (mean of the posterior)

Example 2: Bayesian MMSE Position Estimation

- 3D positioning problem: this time let's assume a prior on unknown initial state

$$x(0) = \begin{bmatrix} \xi(0) \\ \eta(0) \\ z(0) \end{bmatrix} \rightarrow p(x(0)) = \mathcal{N}(\mu_0, P_0) \qquad \mu_0 = \begin{bmatrix} \bar{\xi}(0) \\ \bar{\eta}(0) \\ \bar{z}(0) \end{bmatrix} \qquad P_0 = \begin{bmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_\eta^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}$$

→ so given data $y_{1:T}$ w/ $p(y_{1:T} | x_0) = \prod_{k=1}^{T} p(y_k | x_0) = \prod_{k=1}^{T} \mathcal{N}(x_0, R)$

→ we have $p(x_0 | y_{1:T}) \propto p(x_0) \cdot p(y_{1:T} | x_0) = \mathcal{N}_{x_0}(\mu_0, P_0) \cdot \prod_{k=1}^{T} \mathcal{N}_{y_k}(x_0, R) \big|_{y_k}$

→ From Lecture 17: we know that the posterior is conditional Gaussian pdf

$p(x_0 | y_{1:T}) = \mathcal{N}(\underline{\mu_+}, \underline{P_+})$, where $\mu_+ = \mu_0 + P_0 H^T [H P_0 H^T + R]^{-1} (\vec{y} - H\mu_0)$

$P_+ = P_0 - P_0 H^T [H P_0 H^T + R]^{-1} H P_0$

where in this case: $H = \begin{bmatrix} I_{3\times3} \\ \vdots \\ I_{3\times3} \end{bmatrix} \in \mathbb{R}^{3T\times3}$

$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \in \mathbb{R}^{3T\times1}$, $R = \begin{bmatrix} R & & \\ & \ddots & \\ & & R \end{bmatrix} \in \mathbb{R}^{3T\times3T}$

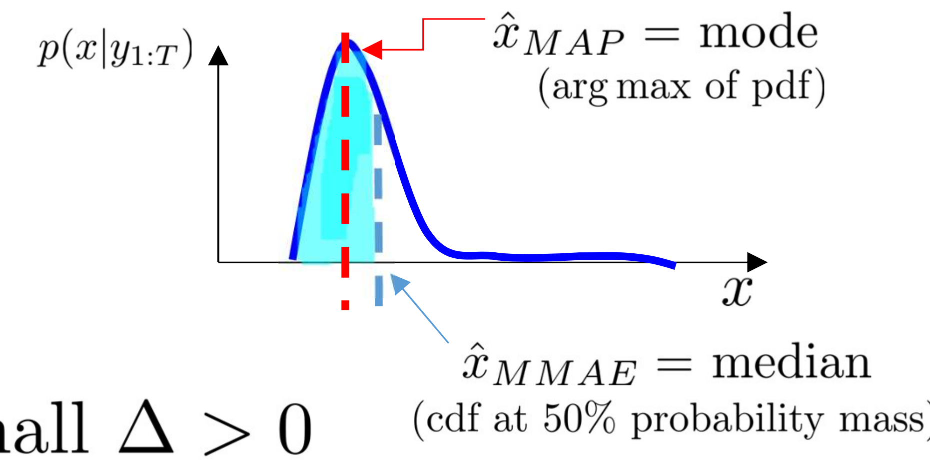
→ $\boxed{\hat{x}_{MMSE} = \mu_+}$

Other Cost Functions for Bayesian Point Estimation

- **L_1 norm:** $C(x,\hat{x}) = ||x-\hat{x}||_1 = \sum_{i=1}^{n} |x(i) - \hat{x}(i)|$

$\rightarrow \hat{x}_{\text{MMAE}} = \arg\min_{x \in \mathbb{R}^n} E[\ ||x-\hat{x}||_1 \mid y_{1:T}]$ (MMAE: minimum mean absolute error)

$\Rightarrow$ can show (for any $p(x|y_{1:T})$): $\hat{x}_{\text{MMAE}} =$ **median** of $p(x|y_{1:T})$



- **"uniform cost":** $C(x,\hat{x}) = \begin{cases} 0, & \text{if } ||x-\hat{x}||_1 \leq \Delta \\ 1, & \text{if } ||x-\hat{x}||_1 > \Delta \end{cases}$ for any small $\Delta > 0$

$\rightarrow \hat{x}_{\text{MAP}} = \arg\min_{x \in \mathbb{R}^n} 1 - P(||x-\hat{x}||_1 \leq \Delta |\ y_{1:T})$ (MAP: maximum a posteriori)

$\Rightarrow$ can show (for any $p(x|y_{1:T})$): $\hat{x}_{\text{MAP}} =$ **mode** (maximum) of $p(x|y_{1:T})$

- **For Gaussian posterior pdfs**: the MMSE, MMAE, and MAP estimators all coincide (posterior mean = posterior median = posterior mode) and are obtained in closed-form (conditional Gaussian mean)!
  - **But generally not true for arbitrary pdfs** (e.g. AQ1 from HW 5)