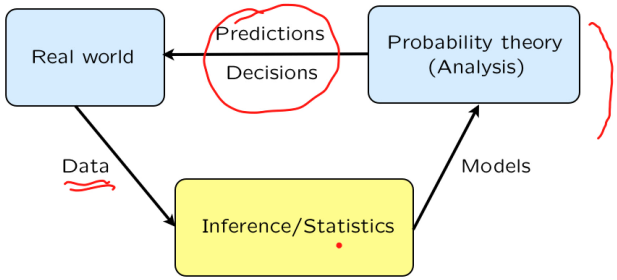
LECTURE 14: Introduction to Bayesian inference

- The big picture
  - motivation, applications
  - problem types (hypothesis testing, estimation, etc.)
- The general framework
  - Bayes' rule $\longrightarrow$ posterior (4 versions)
  - point estimates (MAP, LMS)
  - performance measures) (prob. of error; mean squared error)
  - examples

Inference: the big picture



Inference then and now

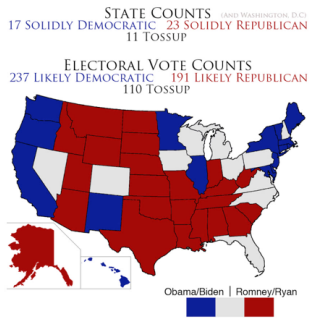
- Then:
  10 patients were treated: 3 died
  10 patients were not treated: 5 died
  Therefore ...

Now:

- Big data
- Big models
- Big computers

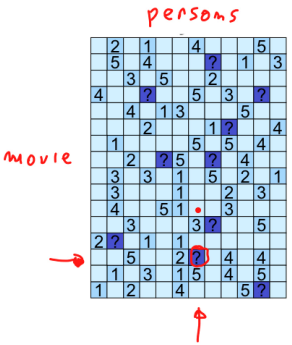
A sample of application domains

- Design and interpretation of experiments
  - polling



A sample of application domains

- marketing, advertising
- recommendation systems
  - Netflix competition

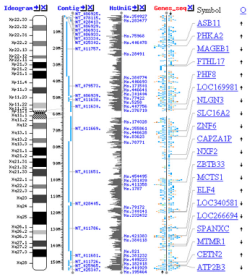


A sample of application domains

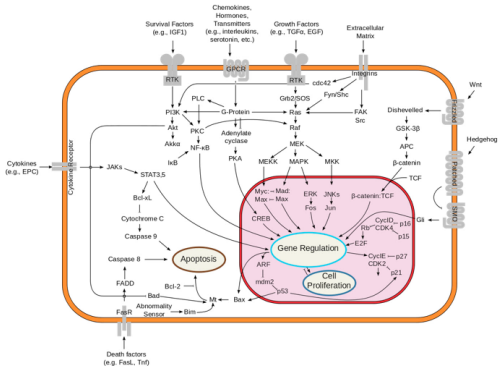
- Finance

A sample of application domains

- Life sciences
  - genomics



  - systems biology



- neuroscience, etc., etc.

A sample of application domains

- Modeling and monitoring the oceans
- Modeling and monitoring global climate
- Modeling and monitoring pollution
- Interpreting data from physics experiments
- Interpreting astronomy data

A sample of application domains

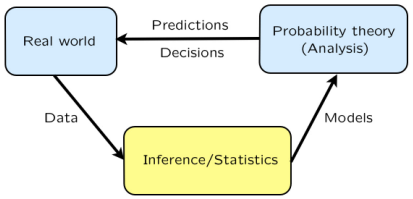
- Signal processing
  - communication systems (noisy ...)
  - speech processing and understanding
  - image processing and understanding
  - tracking of objects
  - positioning systems (e.g., GPS)
  - detection of abnormal events

Model building versus inferring unobserved variables

$X = aS + W$

- Model building:
  - know "signal" S, observe X
  - infer a

- Variable estimation:
  - know a, observe X
  - infer S



Hypothesis testing versus estimation

- Hypothesis testing:
  - unknown takes one of few possible values
  - aim at small probability of incorrect decision

  Is it an airplane or a bird?

- Estimation:
  - numerical unknown(s)
  - aim at an estimate that is "close" to the true but unknown value

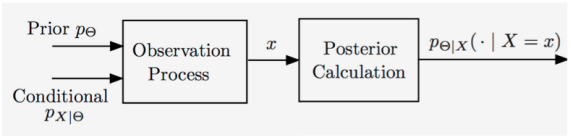
The Bayesian inference framework

- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ

- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$

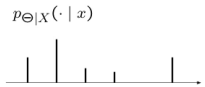
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \,|\, X = x)$ or $f_{\Theta|X}(\cdot \,|\, X = x)$

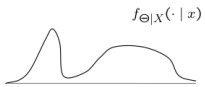
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

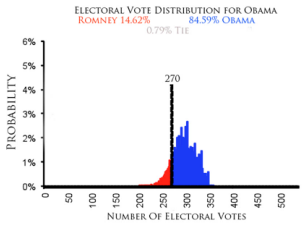


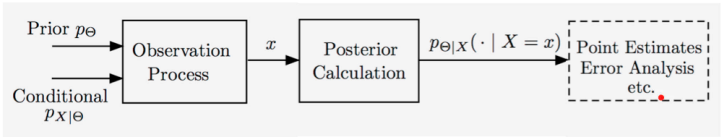
The output of Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$



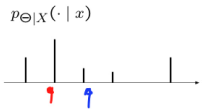


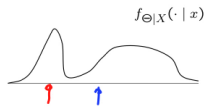




Point estimates in Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$





estimate: $\hat{\theta} = g(x)$
(number)

estimator: $\widehat{\Theta} = g(X)$
(random variable)

- Maximum a posteriori probability (MAP):

$$p_{\Theta|X}(\theta^* \mid x) = \max_\theta p_{\Theta|X}(\theta \mid x)$$

$$f_{\Theta|X}(\theta^* \mid x) = \max_\theta f_{\Theta|X}(\theta \mid x)$$

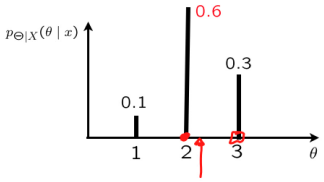
- Conditional expectation: $\mathbf{E}[\Theta \mid X = x]$ (LMS: Least Mean Squares)

Discrete Θ, discrete X

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_\Theta(\theta')\, p_{X|\Theta}(x \mid \theta')$$

- values of Θ: alternative hypotheses



- MAP rule: $\hat{\theta} = 2$

LMS: $\hat{\theta} = E[\Theta \mid X=x] = 2.2$

- conditional prob of error:

$\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x) = 0.4$

smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \sum_x \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)\, p_X(x)$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \Theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

Discrete Θ, continuous X

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

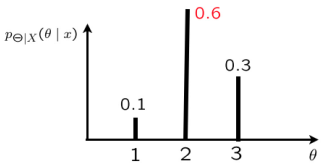
$$f_X(x) = \sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \mid \theta')$$

- Standard example:
  - send signal $\Theta \in \{1, 2, 3\}$

$X = \Theta + W$

$W \sim N(0, \sigma^2)$, indep. of Θ

$f_{X|\Theta}(x \mid \theta) = f_W(x - \theta)$



- MAP rule: $\hat{\theta} = 2$

- conditional prob of error:

$\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x)$

smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \int \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x) f_X(x)\, dx$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

Continuous Θ, continuous X

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$
$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- linear normal models

  estimation of a noisy signal

  $X = \Theta + W$

  Θ and W: independent normals

  multi-dimensional versions (many normal parameters, many observations)

- estimating the parameter of a uniform

  X: uniform$[0, \Theta]$

  Θ: uniform $[0, 1]$

- $\widehat{\Theta} = g(X)$ MAP LMS
- interested in:

$$\begin{cases} \mathbf{E}\big[(\widehat{\Theta} - \Theta)^2 \mid X = x\big] \\ \mathbf{E}\big[(\widehat{\Theta} - \Theta)^2\big] \end{cases}$$

Inferring the unknown bias of a coin and the Beta distribution

$$f_{\Theta|K}(\theta \mid k) = \frac{f_\Theta(\theta)\, p_{K|\Theta}(k \mid \theta)}{p_K(k)}$$
$$p_K(k) = \int f_\Theta(\theta') p_{K|\Theta}(k \mid \theta')\, d\theta'$$

- Standard example:
  - coin with bias Θ; prior $f_\Theta(\cdot)$
  - fix n; K =number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$f_{\Theta|K}(\theta \mid k) = \frac{1 \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)} \qquad \theta \in [0,1]$$
$$= \frac{1}{d(n,k)} \theta^k (1-\theta)^{n-k}$$ "Beta distribution, with parameters $(k+1, n-k+1)$"

- If prior is Beta: $f_\Theta(\theta) = \frac{1}{c}\theta^\alpha (1-\theta)^\beta$ $\alpha, \beta \geq 0$

$$f_{\Theta|K}(\theta \mid k) = \frac{\frac{1}{c}\theta^\alpha(1-\theta)^\beta \binom{n}{k}\theta^k(1-\theta)^{n-k}}{p_K(k)} = d\,\theta^{\alpha+k}(1-\theta)^{\beta+n-k}$$

Inferring the unknown bias of a coin: point estimates

$$\int_0^1 \theta^\alpha (1-\theta)^\beta\, d\theta = \frac{\alpha!\,\beta!}{(\alpha+\beta+1)!} \qquad \alpha \geq 0,\ \beta \geq 0$$

- Standard example:
  - coin with bias Θ; prior $f_\Theta(\cdot)$
  - fix n; K =number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$f_{\Theta|K}(\theta \mid k) = \frac{1}{d(n,k)} \theta^k (1-\theta)^{n-k}$$

- MAP estimate:

$$\hat{\theta}_{\text{MAP}} = k/n$$

$$\max_\theta \big[k \log\theta + (n-k)\log(1-\theta)\big]$$
$$k/\theta - (n-k)/(1-\theta) = 0$$

$$\widehat{\Theta}_{\text{MAP}} = K/n$$

$$\mathbf{E}[\Theta \mid K = k] = \int_0^1 \theta\, f_{\Theta|K}(\theta|k)\, d\theta$$
$$= \frac{1}{d(n,k)} \int_0^1 \theta^{k+1}(1-\theta)^{n-k}\, d\theta$$
$$= \frac{1}{k!\,(n-k)!} \cdot \frac{(k+1)!\,(n-k)!}{(n+2)!}$$
$$= \frac{k+1}{n+2} \approx \frac{k}{n} \quad (n \text{ large})$$

Summary

- Problem data: $p_\Theta(\cdot)$, $p_{X|\Theta}(\cdot \mid \cdot)$
- Given the value x of X: find, e.g., $p_{\Theta|X}(\cdot \mid x)$
  - using appropriate version of the Bayes rule (4 choices)
- Estimator $\widehat{\Theta} = g(X)$ Estimate $\hat{\theta} = g(x)$
  - MAP: $\hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta \mid x)$
  - LMS: $\hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \mathbf{E}[\Theta \mid X = x]$
- Performance evaluation of an estimator $\widehat{\Theta}$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta)$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2 \mid X = x\big]$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2\big]$

total prob/exp thm.

LECTURE 15: Linear models with normal noise

$X_i = \sum_{j=1}^{m} a_{ij}\Theta_j + W_i$ $\quad$ W_i, Θ_j: independent, normal

- Very common and convenient model
- Bayes' rule: normal posteriors
- MAP and LMS estimates coincide
  - simple formulas
    (linear in the observations)
- Many nice properties
- Trajectory estimation example

Recognizing normal PDFs

$X \sim N(\mu, \sigma^2)$ $\quad$ $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ $\quad$ $2\alpha x + \beta = 0$

$c \cdot e^{-8(x-3)^2}$ $\quad$ $\mu = 3$ $\quad$ $\frac{1}{2\sigma^2} = 8 \Rightarrow \sigma^2 = \frac{1}{16}$ $\quad$ $c = \frac{1}{\frac{1}{4}\sqrt{2\pi}}$

$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$ $\quad$ $\alpha > 0$ $\quad$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

$\alpha x^2 + \beta x + \gamma = \alpha\left(x^2 + \frac{\beta}{\alpha}x + \frac{\gamma}{\alpha}\right) = \alpha\left(\left(x + \frac{\beta}{2\alpha}\right)^2 - \frac{\beta^2}{4\alpha^2} + \frac{\gamma}{\alpha}\right)$

$f_x(x) = c\, e^{-\alpha\left(x+\frac{\beta}{2\alpha}\right)^2} e^{-\alpha\left(-\frac{\beta^2}{4\alpha^2} + \frac{\gamma}{\alpha}\right)}$ $\quad$ $\mu = -\frac{\beta}{2\alpha}$

$\frac{1}{2\sigma^2} = \alpha \Rightarrow \sigma^2 = 1/2\alpha$

Estimating a normal random variable in the presence of additive normal noise

$X = \Theta + W$ $\quad$ $\Theta, W : N(0,1)$, $\quad$ independent

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$f_{X|\Theta}(x \mid \theta)$: $X = \theta + W$ $\quad$ $N(\theta, 1)$

$f_{\Theta|X}(\theta \mid x) = \frac{1}{f_x(x)}\, c\, e^{-\frac{1}{2}\theta^2}\, c\, e^{-\frac{1}{2}(x-\theta)^2} = c(x)\, e^{-\text{quadratic}(\theta)}$ $\quad$ Normal!

Fix x $\quad$ $\min_\theta \left[\frac{1}{2}\theta^2 + \frac{1}{2}(x-\theta)^2\right]$ $\quad$ $\theta + (\theta - x) = 0$

$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = x/2$ $\quad$ $\widehat{\Theta}_{\text{MAP}} = \mathbf{E}[\Theta \mid X] = X/2$

Estimating a normal parameter in the presence of additive normal noise

$X = \Theta + W$ $\quad$ $\Theta, W : N(0,1)$, $\quad$ independent

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$\widehat{\Theta}_{\text{MAP}} = \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X] = \frac{X}{2}$

- Even with general means and variances:
  - posterior is normal
  - LMS and MAP estimators coincide
  - these estimators are "linear," of the form $\widehat{\Theta} = aX + b$

The case of multiple observations

$$X_1 = \Theta + W_1 \qquad \Theta \sim N(x_0, \sigma_0^2) \qquad W_i \sim N(0, \sigma_i^2)$$
$$\vdots$$
$$X_n = \Theta + W_n \qquad \Theta, W_1, \ldots, W_n \text{ independent}$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$
$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$f_{X_i|\Theta}(x_i \mid \theta) =$ $c_i\, e^{-(x_i-\theta)^2/2\sigma_i^2}$

given $\Theta=\theta$: $X_i = \theta + W_i \sim N(\theta, \sigma_i^2)$

$f_{X|\Theta}(x \mid \theta) = f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^n f_{X_i|\Theta}(x_i \mid \theta)$

given $\Theta=\theta$: W_i independent $\Rightarrow$ X_i independent

$f_{\Theta|X}(\theta \mid x) = \frac{1}{f_X(x)} \cdot c_0\, e^{-(\theta-x_0)^2/2\sigma_0^2} \prod_{i=1}^n c_i\, e^{-(x_i-\theta)^2/2\sigma_i^2}$ Normal!

The case of multiple observations

$$f_{\Theta|X}(\theta \mid x) = c \cdot \exp\{-\text{quad}(\theta)\} \qquad \text{quad}(\theta) = \frac{(\theta-x_0)^2}{2\sigma_0^2} + \frac{(\theta-x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta-x_n)^2}{2\sigma_n^2}$$

find peak

$\frac{d}{d\theta}\text{quad}(\theta) = 0$: $\sum_{i=0}^n \frac{(\theta-x_i)}{\sigma_i^2} = 0 \Rightarrow \theta \sum_{i=0}^n \frac{1}{\sigma_i^2} = \sum_{i=0}^n \frac{x_i}{\sigma_i^2}$

$$\hat\theta_{\text{MAP}} = \hat\theta_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

The case of multiple observations

- Key conclusions:
  - posterior is normal
  - LMS and MAP estimates coincide
  - these estimates are "linear," of the form $\hat\theta = a_0 + a_1 x_1 + \cdots + a_n x_n$
- Interpretations:
  - estimate $\hat\theta$: weighted average of x_0 (prior mean) and x_i (observations)
  - weights determined by variances

$$\hat\theta_{\text{MAP}} = \hat\theta_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

σ_i^2 large

X_i very noisy

$\Rightarrow$ small weight

The mean squared error

$X_i = \Theta + W_i$

$$f_{\Theta|X}(\theta \mid x) = c \cdot \exp\{-\text{quad}(\theta)\}$$

$$\text{quad}(\theta) = \frac{(\theta-x_0)^2}{2\sigma_0^2} + \frac{(\theta-x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta-x_n)^2}{2\sigma_n^2}$$

$$\hat\theta = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}$$

- Performance measures:

$\mathbf{E}\big[(\Theta-\widehat\Theta)^2 \mid X=x\big] = \mathbf{E}\big[(\Theta-\hat\theta)^2 \mid X=x\big] = \text{var}(\Theta \mid X=x) = \Big/ \sum_{i=0}^n \frac{1}{\sigma_i^2}$

$\mathbf{E}\big[(\Theta-\widehat\Theta)^2\big] = \int E\big[(\Theta-\hat\Theta)^2 \mid X=x\big] f_X(x)\, dx$ $=$

$f_X(x) = c \cdot e^{-(\alpha x^2+\beta x+\gamma)}$ $\quad \alpha > 0$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

$\alpha = \frac{1}{2\sigma_0^2} + \cdots + \frac{1}{2\sigma_n^2}$

some σ_i^2 small $\rightarrow$ MSE small

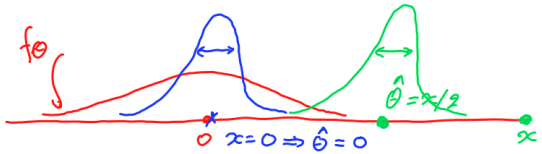
all σ_i^2 large $\rightarrow$ MSE large

The mean squared error

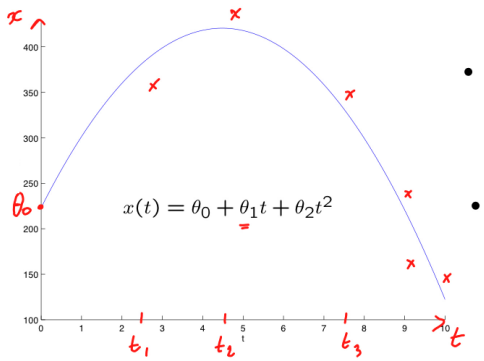
$$\mathbf{E}\big[(\Theta-\widehat{\Theta})^2 \mid X=x\big] = \mathbf{E}\big[(\Theta-\widehat{\Theta})^2\big] = 1\Big/\sum_{i=0}^{n}\frac{1}{\sigma_i^2}$$

$$\hat{\theta} = \frac{\sum_{i=0}^{n}\frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n}\frac{1}{\sigma_i^2}}$$

- Example: $\sigma_0^2=\sigma_1^2=\cdots=\sigma_n^2=\sigma^2$ $\quad \frac{1}{(n+1)\frac{1}{\sigma^2}} = \frac{\sigma^2}{n+1}$
- conditional mean squared error same for all x
- Example: $X=\Theta+W$ $\quad \Theta\sim N(0,1)$, $\quad W\sim N(0,1)$
  independent Θ, W $\quad \widehat{\Theta}=X/2$ $\quad \mathbf{E}\big[(\Theta-\widehat{\Theta})^2 \mid X=x\big] = 1/2$



The case of multiple parameters: trajectory estimation



- Random variables $\Theta_0, \Theta_1, \Theta_2$
  independent; priors f_{Θ_j}
- Measurements at times $t_1,\ldots,t_n$
  $X_i=\Theta_0+\Theta_1 t_i+\Theta_2 t_i^2+W_i$
  noise model: f_{W_i}
  independent W_i; independent from Θ_j

A model with normality assumptions

$$X_i=\Theta_0+\Theta_1 t_i+\Theta_2 t_i^2+W_i \qquad i=1,\ldots,n$$

$$f_{\Theta|X}(\theta\mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x\mid\theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x\mid\theta)\, d\theta$$

- assume $\Theta_j\sim N(0,\sigma_j^2)$, $W_i\sim N(0,\sigma^2)$; independent
- Given $\Theta=\theta=(\theta_0,\theta_1,\theta_2)$, X_i is: $N(\theta_0+\theta_1 t_i+\theta_2 t_i^2, \sigma^2)$

  $f_{X_i|\Theta}(x_i\mid\theta) = c\cdot\exp\Big\{-(x_i-\theta_0-\theta_1 t_i-\theta_2 t_i^2)^2/2\sigma^2\Big\}$
- posterior: $f_{\Theta|X}(\theta\mid x) = \frac{1}{f_X(x)}\prod_{j=0}^{2} f_{\Theta_j}(\theta_j)\prod_{i=1}^{n} f_{X_i|\Theta}(x_i\mid\theta)$

  $$c(x)\exp\Big\{-\frac{1}{2}\Big(\frac{\theta_0^2}{\sigma_0^2}+\frac{\theta_1^2}{\sigma_1^2}+\frac{\theta_2^2}{\sigma_2^2}\Big)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0-\theta_1 t_i-\theta_2 t_i^2)^2\Big\}$$

A model with normality assumptions

$$f_{\Theta|X}(\theta\mid x) = c(x)\exp\Big\{-\frac{1}{2}\Big(\frac{\theta_0^2}{\sigma_0^2}+\frac{\theta_1^2}{\sigma_1^2}+\frac{\theta_2^2}{\sigma_2^2}\Big)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0-\theta_1 t_i-\theta_2 t_i^2)^2\Big\}$$

- MAP estimate: maximize over $(\theta_0,\theta_1,\theta_2)$;
  (minimize quadratic function)

  $\frac{\partial}{\partial\theta_j}(\text{quad}(\theta)) = 0$ $\quad$ 3 equations, 3 unknowns
  linear

Linear normal models

- Θ_j and X_i and are linear functions of independent normal random variables
- $f_{\Theta|X}(\theta\,|\,x) = c(x)\exp\big\{-\text{quadratic}(\theta_1,\ldots,\theta_m)\big\}$
- MAP estimate: maximize over $(\theta_1,\ldots\theta_m)$; (minimize quadratic function)

  $\widehat{\Theta}_{\text{MAP},j}$: linear function of $X = (X_1,\ldots,X_n)$
- Facts:
  - $\widehat{\Theta}_{\text{MAP},j} = \mathbf{E}[\Theta_j\,|\,X]$
  - marginal posterior PDF of Θ_j: $f_{\Theta_j|X}(\theta_j\,|\,x)$, is normal
  - MAP estimate based on the joint posterior PDF: same as MAP estimate based on the marginal posterior PDF
  - $\mathbf{E}\big[(\widehat{\Theta}_{i,\text{MAP}} - \Theta_i)^2\,|\,X = x\big]$: same for all x

An illustration

Estimating the trajectory of a free-falling object

$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$

$X(t) = \Theta_0 + \Theta_1 t + \Theta_2 t^2$

$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$

$$\underset{\theta_0,\theta_1,\theta_2}{\text{minimize}} \quad \frac{1}{2}\Big(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\Big) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2$$

An illustration

Estimating the trajectory of a free-falling object

$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$

$X(t) = \Theta_0 + \Theta_1 t + \Theta_2 t^2$

$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$

$$\underset{\theta_0,\,\theta_1}{\text{minimize}} \quad (\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration

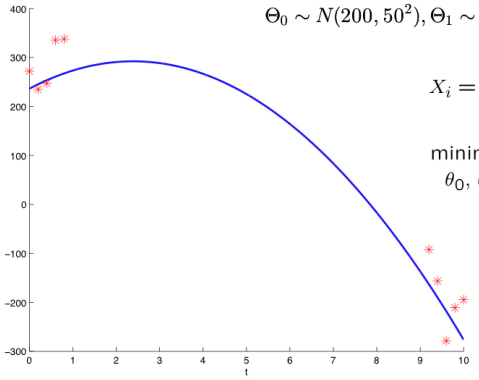
Estimating the trajectory of a free-falling object

$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$

$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$

$$\underset{\theta_0,\,\theta_1}{\text{minimize}} \quad (\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration Estimating the trajectory of a free-falling object

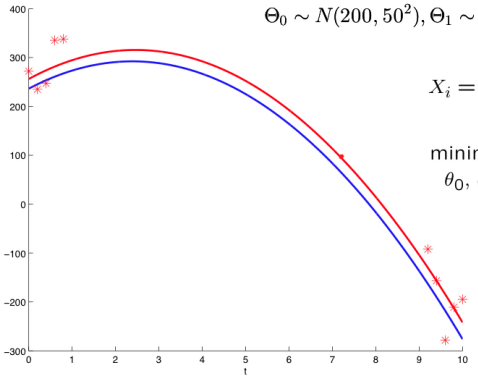


$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\underset{\theta_0,\, \theta_1}{\text{minimize}} \quad (\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration Estimating the trajectory of a free-falling object

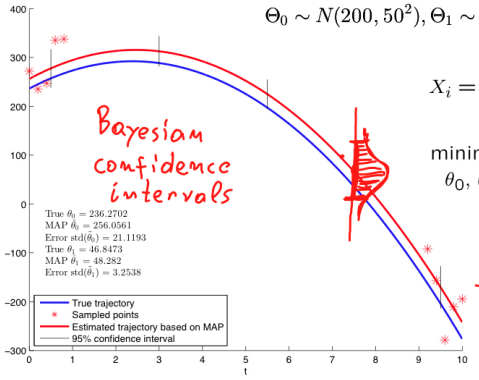


$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \overbrace{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2}^{x(t)} + W_i$$

$$\underset{\theta_0,\, \theta_1}{\text{minimize}} \quad (\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

An illustration Estimating the trajectory of a free-falling object



$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

$$X_i = \overbrace{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2}^{x(t)} + W_i$$

$$\underset{\theta_0,\, \theta_1}{\text{minimize}} \quad (\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

$$P(x(t) \in \text{interval} \mid \text{data}) = 0.95$$

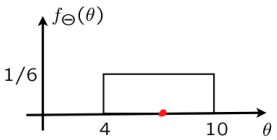
LECTURE 16: Least mean squares (LMS) estimation

- minimize (conditional) mean squared error $\mathbf{E}\left[(\Theta-\hat{\theta})^2 \,|\, X=x\right]$
  - solution: $\hat{\theta}=\mathbf{E}[\Theta \,|\, X=x]$
  - general estimation method
- Mathematical properties
- Example

LMS estimation in the absence of observations

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
  - no observations available
  - MAP rule: any $\hat{\theta} \in [4, 10]$
  - (Conditional) expectation: $\hat{\theta}=7$
- Criterion: Mean Squared Error (MSE): $\mathbf{E}\left[(\Theta-\hat{\theta})^2\right]$

minimize mean squared error



LMS estimation in the absence of observations

- Least mean squares formulation:

minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta-\hat{\theta})^2\right]$: $\hat{\theta}=\mathbf{E}[\Theta]$

$E[\Theta^2] - 2E[\Theta]\hat{\theta} + \hat{\theta}^2 \quad \frac{d}{d\hat{\theta}} = 0: \; -2E[\Theta]+2\hat{\theta}=0$

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

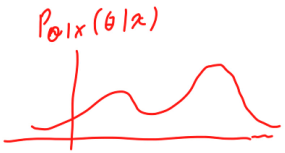
$Var(\Theta-\hat{\theta}) + (E[\Theta-\hat{\theta}])^2$ — minimized when $\hat{\theta}=E[\Theta]$

$Var(\Theta)$

- Optimal mean squared error: $\mathbf{E}\left[(\Theta-\mathbf{E}[\Theta])^2\right]=\text{var}(\Theta)$

LMS estimation of Θ based on X

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
- observation X; model $p_{X|\Theta}(x \,|\, \theta)$
  - observe that $X=x$



minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta-\hat{\theta})^2\right]$: $\hat{\theta}=\mathbf{E}[\Theta]$

minimize conditional mean squared error, $\mathbf{E}\left[(\Theta-\hat{\theta})^2 \,|\, X=x\right]$: $\hat{\theta}=\mathbf{E}[\Theta \,|\, X=x]$

- LMS estimate: $\hat{\theta}=\mathbf{E}[\Theta \,|\, X=x]$

estimator: $\hat{\Theta}=\mathbf{E}[\Theta \,|\, X]$

LMS estimation of Θ based on X

- $\mathbf{E}[\Theta]$ minimizes $\mathbf{E}\big[(\Theta - \hat{\theta})^2\big]$

$E[(\Theta - E[\Theta])^2] \le E[(\Theta - c)^2]$, for all c

- $\mathbf{E}[\Theta \mid X = x]$ minimizes $\mathbf{E}\big[(\Theta - \hat{\theta})^2 \mid X = x\big]$

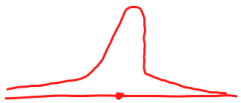
$E[(\Theta - E[\Theta|X=x])^2 \mid X=x] \le E[(\Theta - g(x))^2 \mid X=x]$ for all x

$E[(\Theta - E[\Theta|X])^2 \mid X] \le E[(\Theta - g(X))^2 \mid X]$

$E[(\Theta - E[\Theta|X])^2] \le E[(\Theta - g(X))^2]$

$\widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X]$ minimizes $\mathbf{E}\big[(\Theta - g(X))^2\big]$, over all estimators $\widehat{\Theta} = g(X)$

LMS estimation of Θ based on X



- LMS relevant to estimation (not hypothesis testing)
- Same as MAP if the posterior is unimodal and symmetric around the mean
  - e.g., when posterior is normal (the case in "linear–normal" models)

LMS performance evaluation

- LMS estimate: $\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$

  estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \mid X]$

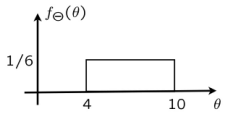
- Expected performance, once we have a measurement:

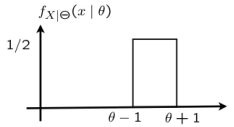
  $\text{MSE} = \mathbf{E}\big[\big(\Theta - \mathbf{E}[\Theta \mid X = x]\big)^2 \mid X = x\big] = \text{var}(\Theta \mid X = x)$

- Expected performance of the design:

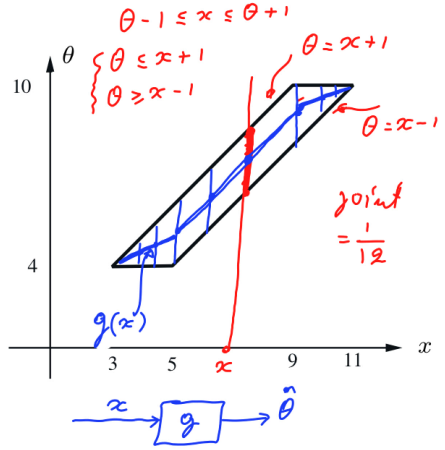
  $\text{MSE} = \mathbf{E}\big[\big(\Theta - \mathbf{E}[\Theta \mid X]\big)^2\big] = \mathbf{E}\big[\text{var}(\Theta \mid X)\big]$

Example

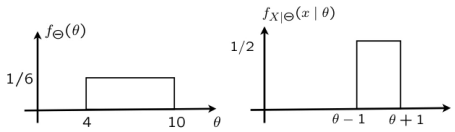


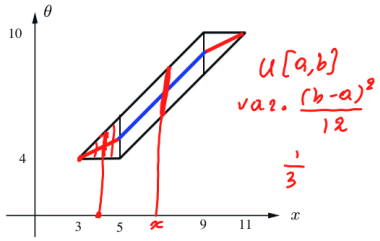


$X = \Theta + U$ $\quad U \sim \text{unif}(-1, 1)$



Conditional mean squared error





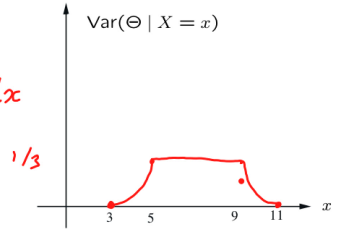
$U[a,b]$

var. $\frac{(b-a)^2}{12}$

$\frac{1}{3}$

- $\mathbf{E}[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x]$
  - same as $\mathrm{Var}(\Theta \mid X = x)$: variance of conditional distribution of Θ

$E[\mathrm{Var}(\Theta|X)] = \int f_X(x)\,\mathrm{Var}(\Theta|X=x)\,dx$



LMS estimation with multiple observations or unknowns

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
- observations $X = (X_1, X_2, \ldots, X_n)$; model $p_{X|\Theta}(x\,|\,\theta)$
  - observe that $X = x$
  - new universe: condition on $X = x$
- LMS estimate: $\mathbf{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$
- If Θ is a vector, apply to each component separately

$\Theta = (\Theta_1, \ldots, \Theta_m)$ $\quad \hat{\Theta}_j = E[\Theta_j \mid X_1 = x_1, \ldots, X_m = x_m]$

Some challenges in LMS estimation

$$f_{\Theta|X}(\theta\,|\,x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x\,|\,\theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x\,|\,\theta')\,d\theta'$$

- Full correct model, $f_{X|\Theta}(x\,|\,\theta)$, may not be available
- Can be hard to compute/implement/analyze

$E[\Theta_j \mid X = x] = \int\cdots\int \theta_j\, f_{\Theta|X}(\theta|x)\,d\theta_1 \cdots d\theta_m$

Properties of the estimation error in LMS estimation

- Estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \mid X]$
- Error: $\widetilde{\Theta} = \widehat{\Theta} - \Theta$

$E[\hat{\Theta}] = E[\Theta]$

$E[\tilde{\Theta}] = 0$

$\mathbf{E}[\widetilde{\Theta} \mid X = x] = 0$ $\quad E[\hat{\Theta} - \Theta \mid X = x] = \hat{\Theta} - E[\Theta|X=x] = 0$

$\mathrm{cov}(\widetilde{\Theta}, \widehat{\Theta}) = 0$ $\quad E[\tilde{\Theta}\hat{\Theta}] - E[\tilde{\Theta}]E[\hat{\Theta}]$ $\quad 0$

$E[\tilde{\Theta}\hat{\Theta} \mid X = x] = \hat{\Theta}E[\tilde{\Theta}|X=x] = 0$

$\mathrm{var}(\Theta) = \mathrm{var}(\widehat{\Theta}) + \mathrm{var}(\widetilde{\Theta})$ •

$\Theta = \hat{\Theta} - \tilde{\Theta}$