MEDICAL IMAGING INFORMATICS: Lecture # 1

Estimation Theory

Norbert Schuff
Professor of Radiology
VA Medical Center and UCSF
Norbert.schuff@ucsf.edu
Objectives Of Today’s Lecture

• Understand basic concepts of data modeling
  – Deterministic
  – Probabilistic
  – Bayesian

• Learn the form of the most common estimators
  – Least squares estimator
  – Maximum likelihood estimator
  – Maximum a-posteriori estimator

• Learn how estimators are used in image processing
What Is Medical Imaging Informatics?

- **Signal Processing**
  - Digital Image Acquisition
  - Image Processing and Enhancement

- **Data Mining**
  - Computational anatomy
  - Statistics
  - Databases
  - Data-mining
  - Workflow and Process Modeling and Simulation

- **Data Management**
  - Picture Archiving and Communication System (PACS)
  - Imaging Informatics for the Enterprise
  - Image-Enabled Electronic Medical Records
  - Radiology Information Systems (RIS) and Hospital Information Systems (HIS)
  - Quality Assurance
  - Archive Integrity and Security

- **Data Visualization**
  - Image Data Compression
  - 3D, Visualization and Multi-media
  - DICOM, HL7 and other Standards

- **Teleradiology**
  - Imaging Vocabularies and Ontologies
  - Transforming the Radiological Interpretation Process (TRIP)[2]
  - Computer-Aided Detection and Diagnosis (CAD).
  - Radiology Informatics Education

- **Etc.**
What Is The Focus Of This Course?

Learn how to maximize information using efficient computational tools

- Collect data
- Image
- Measurements
- Model
- Data drive the model
- Compare with model
- Inference Statistic
- Generative statistic
- knowledge

Data drive the model
Challenge: Maximizing Information Gain

1. Q: How to estimate quantities from a given set of uncertain (noisy) measurements?
   A: *Apply estimation theory* (1\textsuperscript{st} lecture today)

2. Q: How to quantify information?
   A: *Apply information theory* (2\textsuperscript{nd} lecture next week)
Motivation Example I: Tissue Classification

Gray/White Matter Segmentation

Hypothetical Histogram

GM/WM overlap 50:50; Can we do better than flipping a coin?
Motivation Example II: MR Spectroscopy

Magnetic Resonance Spectroscopy Identifies Neural Progenitor Cells in the Live Human Brain
Louis N. Manganas, et al.
Science 318, 980 (2007);
DOI: 10.1126/science.1147851

Colored spectra from singular value decomposition (SVD)

Inlets from Fourier Transformation (FT)
Motivation Example III: Signal Decomposition

Diffusion Tensor Imaging (DTI)
- Sensitive to random motion of water
- Senses the neighborhood on microscopic scale

Goal:
Represent fiber bundles

Quantitative Diffusion Maps
Microscopic tissue sample

Dr. Van Wedeen, MGH
Basic Concepts of Modeling

\( \theta \): unknown world state of interest

\( \varphi \): measurement

\( \hat{\theta} \): Estimator - a good guess of \( \theta \) based on measurements
Deterministic Model

N = number of measurements
M = number of states, M=1 is possible
Usually N > M and \|noise\|^2 > 0

\[
\begin{pmatrix}
\varphi_1 \\
\vdots \\
\varphi_n
\end{pmatrix}
= \begin{pmatrix}
h_{11} & \cdots & h_{1M} \\
\vdots & \ddots & \vdots \\
h_{n1} & \cdots & h_{nm}
\end{pmatrix}
\begin{pmatrix}
\theta_1 \\
\vdots \\
\theta_M
\end{pmatrix}
+ \begin{pmatrix}
\text{noise}_1 \\
\vdots \\
\text{noise}_n
\end{pmatrix}
\]

\[\varphi_N = H_{N \times M} \hat{\theta}_M + \text{noise}_N\]

The model is deterministic, because discrete values of \(\theta\) are solutions.

Note:
1) we make no assumption about the distribution of \(\theta\)
2) Each value is as likely as any another value

What is the best estimator under these circumstances?
Least-Squares Estimator (LSE)

The best what we can do is *minimizing noise*:

\[ \varphi - H\theta = \text{noise} \]

\[ \varphi - H\hat{\theta}_{\text{LSE}} = 0 \]

\[ H^T\varphi - (H^T H)\hat{\theta}_{\text{LSE}} = 0 \]

\[ \hat{\theta}_{\text{LSE}} = (H^T H)^{-1} H^T \varphi \]

• LSE is popular choice for model fitting  
• Useful for obtaining a descriptive measure

But

• LSE makes no assumptions about distributions of data or parameters  
• Has no basis for statistics \( \Rightarrow \) “deterministic model”

Covariance Matrix
Prominent Examples of LSE

Mean Value: \[ \hat{\theta}_{\text{mean}} = \frac{1}{N} \sum_{j=1}^{N} \varphi_j \]

Variance: \[ \hat{\theta}_{\text{variance}} = \frac{1}{N-1} \sum_{j=1}^{N} (\varphi_j - \hat{\theta}_{\text{mean}})^2 \]

Amplitude \( \hat{\theta}_1 \)
Frequency \( \hat{\theta}_2 \)
Phase \( \hat{\theta}_3 \)
Decay \( \hat{\theta}_4 \)
We believe $\theta$ is governed by chance, but we don’t know the governing probability.

We perform measurements for all possible values of $\theta$.

We obtain the likelihood function of $\theta$ given a series of measurements $\varphi$.

Note:
$\theta$ is random and unknown
$\varphi$ has been measured
The likelihood of parameter $\theta$ is the probability of the observed data $\varphi$ as a function of this parameter.
Likelihood Function

Let $\varphi$ be a random variable with a discrete probability distribution $p$ depending on the parameter $\theta$. The likelihood function of $\theta$ (given the outcome $\varphi_j$ of $\varphi$) is:

$$L(\theta | \varphi) = \Pr(\varphi = \varphi_j ; \theta)$$

Example: Coin toss

Let the probability that a coin lands heads up when tossed be $p_H$.

The probability of getting two heads in two tosses (HH) is $p_H \ast p_H$

Thus, if $p_H=0.5$, the probability of seeing two heads is 0.25.

Another way of saying this is that the likelihood that $p_H=0.5$ given observation $HH = 0.25$ is

$$L(p_H = 0.5 | HH) = \Pr(HH ; p_H = 0.5) = 0.25$$

NOTE: this is not the same as saying that the probability that $p_H=0.5$ is 0.25, given the observation HH. The likelihood function is NOT a probability density function which sums to 1.
Likelihood Model (cont’d)

New Goal:
Find an estimator which gives the most likely probability of \( \theta \) underlying \( L(\theta | \phi) \).

Figure 1.2: Perception of 3D angles. The angle \( \theta \) between two line segments in 3D.
Maximum Likelihood Estimator (MLE)

Goal: Find estimator which gives the most likely value of $\theta$ underlying the likelihood function of $\Theta$

Highest probability

$$\hat{\theta}_{MLE} = \max \Pr(\varphi; \theta)$$

The MLE can be found by taking the derivative of the likelihood function
Example I: MLE Of Normal Distribution

Normal distribution

\[
\Pr(\varphi_N | \bar{\theta}, \sigma^2) = \exp\left( \frac{1}{2\sigma^2} \sum_{j=1}^{N}(\varphi(j) - \bar{\theta})^2 \right)
\]

log of the normal distribution (normD)

\[
\ln \Pr(\varphi_N | \bar{\theta}, \sigma^2) = \frac{1}{2\sigma^2} \sum_{j=1}^{N}(\varphi(j) - \bar{\theta})^2
\]

MLE of the mean (1\textsuperscript{st} derivative):

\[
\frac{d}{d \theta} \ln \Pr(\cdot) = \frac{1}{4\hat{\sigma}^2 N} \sum_{j=1}^{N}(\varphi(j) - \theta) = 0
\]

\[
\Theta_{MLE} = \frac{1}{N} \sum_{j=1}^{N} \varphi(j)
\]
Example II: MLE Of Binominal Distribution (Coin Tosses)

\(n\) = # of tosses
\(\varphi\) = # of heads in \(n\) tosses
\(\theta\) = probability of obtaining a head in a toss (unknown)
\(\Pr(\varphi; \theta; n)\) = probability of heads from \(n\) tosses given \(\theta\)

\[\Pr(\varphi; \theta = 0.7; n)\]

\[\Pr(\varphi; \theta = 0.3; n)\]
MLE Of Coin Toss (cont’d)

Goal:

Given $Pr(\phi; \theta)$, estimate the most likely probability distribution $\hat{\theta}_{MLE} = \max Pr(\phi; \theta)$ that produced the data.

Intuitively:

$$\hat{\theta}_{MLE} = \frac{\#\text{heads}}{\#\text{tosses}}$$

For a fair coin $\hat{\theta}_{MLE} \approx 0.5$
MLE Of Coin Toss (cont’d)

Coin tosses follow a binominal distribution

\[ \Pr(Y = y | n, \theta) = \frac{n!}{(y!)(n-y)!} \cdot \theta^y (1 - \theta)^{n-y} \]

Likelihood function of coin tosses

\[ L(\theta | y, n) = \frac{n!}{(y!)(n-y)!} \cdot \theta^y (1 - \theta)^{n-y} \]

What is the likelihood of observing 7 heads given that we tossed a coin 10 times?

For a fair coin \( \theta = 0.5 \):

\[ L(\theta | n = 10, y = 7) = \frac{10!}{(7!)(10-7)!} \cdot 0.5^7 (1 - 0.5)^{10-7} = 0.12 \]

For an unfair coin \( \theta = 0.6 \):

\[ L(\theta | n = 10, y = 7) = \frac{10!}{(7!)(10-7)!} \cdot 0.6^7 (1 - 0.6)^{10-7} = 0.21 \]
MLE Of Coin Toss (cont’d)

Likelihood function of coin tosses

\[ L(\theta \mid y) = \frac{n!}{(y!)(n-y)!} \cdot \theta^y (1-\theta)^{n-y} \]

log likelihood function

\[ \ln L(\theta \mid y) = \ln \frac{n!}{(y!)(n-y)!} + y \ln \theta + (n-y) \ln (1-\theta) \]
MLE Of Coin Toss

Evaluate MLE equation (1st derivative)

$$\ln L(\theta | y) = \ln \frac{n!}{(y!)(n-y)!} + y \ln \theta + (n-y) \ln (1-\theta)$$

$$\frac{d \ln L(\theta)}{d \theta} = \frac{y}{\theta} - \frac{(n-y)}{1-\theta} = 0$$

$$= \frac{y-n\theta}{\theta(1-\theta)} = 0 \Rightarrow \hat{\theta}_{MLE} = \frac{y}{n} = \#heads \div \#tosses$$

According to the MLE principle, the distribution $Pr(y; \varphi_{MLE}; n)$ for a given $n$ is the most likely distribution to have generated the observed data of $y$. 
Relationship between MLE and LSE

Assume: 
- \( \varphi \) is independent of noise
- Probability of measurements and noise have the same distribution

\[
Pr_\theta (\varphi_N \mid \theta) = Pr_{\text{noise}} (\varphi_N - H\theta \mid \theta)
\]

MLE is maximized when LSE is minimized
Now, the daemon comes into play, but we know the daemon’s preference for $\theta$ (prior knowledge).

$$prior(\theta) = Pr(\theta)$$

**New Goal:**
Find the estimator which gives the most likely probability distribution of $\theta$ given everything we know.
Bayesian Model

Prior

Likelihood

Posterior

Figure 1.2: Perception of 3D angles. The angle $\theta$ between two line segments in 3D

$$Pr(\theta \mid \phi) \approx Pr(\phi; \theta) \cdot Pr(\theta)$$
Thomas Bayes (1701-1761)

Gerolamo Cardano (1501-1576)
First to formulate elementary rules of probability

Abraham de Moivre (1667-1754)
First to formally derive the normal distribution curve
Maximum A-Posteriori (MAP) Estimator

Goal:
Find the most likely $\Theta_{\text{MAP}}$ (max. posterior density of ) given $\varphi$.

Maximize joint density

$$\hat{\theta}_{\text{MAP}} = \max(Pr(\varphi; \theta) \cdot Pr(\theta))$$

$\theta_{\text{MAP}}$ can be found by taken the partial derivative of the joint density
With respect to $\theta$
Example III: MAP Of Normal Distribution

The sample mean of MAP is:

\[ \hat{\theta}_{MAP} = \frac{\sigma^2_\mu}{\sigma^2_\varphi + T \sigma^2_\mu} \sum_{j=1}^{N} \varphi_j \]

MAP is a linear combination between the prior mean and sample mean weighted by there respective covariances.

The case \( \sigma^2_\mu \to \infty \) represents an non-informative prior, leading to \( \hat{\theta}_{MAP} = \hat{\theta}_{MLE} \)
Posterior Distribution and Decision Rules

\[(\theta | \rho)\]

\[\theta_{MSE} \quad \theta_{MAP} \quad \theta\]
Some Desirable Properties of Estimators I:

**Unbiased:** Mean value of the error should be zero

\[ E\left(\hat{\Phi} - \Phi\right) = 0 \]

**Consistent:** Error estimator should decrease asymptotically as number of measurements increase. (Mean Square Error (MSE))

\[ MSE = E\left(\|\hat{\Phi} - \Phi\|^2\right) \rightarrow 0 \text{ for large } N \]

What happens to MSE when estimator is biased?

\[ MSE = E\left(\|\hat{\Phi} - \Phi - b\|^2\right) + E\left(\|b\|^2\right) \]

- **variance**
- **bias**
Some Desirable Properties of Estimators II:

Efficient: Co-variance matrix of error should decrease asymptotically to its minimal value for large $N$

$$C_{ik} = E\left\langle \left( \Phi_i - \Phi_i \right) \left( \Phi_k - \Phi_k \right)^T \right\rangle \leq \text{some.very.small.value}$$
Example: Properties Of Estimators \textit{Mean} and \textit{Variance}

Mean:

\[ E\left\langle \hat{\mu} \right\rangle = \frac{1}{N} \sum_{j=1}^{N} E\left\langle \phi(j) \right\rangle = \frac{1}{N} \cdot N \mu = \mu \]

The sample mean is an unbiased estimator of the true mean

Variance:

\[ E\left\langle \left(\hat{\mu} - \mu\right)^2 \right\rangle = \frac{1}{N^2} \sum_{j=1}^{N} E\left\langle \left(\phi(j) - \mu\right)^2 \right\rangle = \frac{1}{N^2} \cdot N \sigma^2 = \frac{\sigma^2}{N} \]

The variance is a consistent estimator because it approaches zero for large number of measurements
Summary

- **LSE** is a descriptive method to accurately fit data to a model.
- **MLE** is a method to seek the probability distribution that makes the observed data most likely.
- **MAP** is a method to seek the most probably parameter value given prior information about the parameters and the observed data.

- If the influence of prior information decreases, MAP approaches MLE.
Some Priors in Imaging

- Smoothness of the brain
- Anatomical boundaries
- Intensity distributions
- Anatomical shapes
- Physical models
  - Point spread function
  - Bandwidth limits
- Etc.
Estimation Theory: Motivation Example I

Gray/White Matter Segmentation

Hypothetical Histogram

What works better than flipping a coin?

Design likelihood functions based on
- anatomy
- co-occurrence of signal intensities
- others

Determine prior distribution
- population based atlas of regional intensities
- model based distributions of intensities
- others
Data-Driven Brain MRI Segmentation On Edge Confidence And Prior Tissue Information

J.R.Jiménez-Alaniz, et al.

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 25, NO. 1, JANUARY 2006

Fig. 6. Two examples of synthetic brain segmentations: column (a) original images, column (b) segmented images, and column (c) reference images.
Estimation Theory: Motivation Example III

Diffusion Spectrum Imaging – Human Cingulum Bundle

Goal:
Capture directions of fiber bundles

Improvements to identify tracts:
Design likelihood functions based on similarity measures of adjacent voxels, e.g. correlations fiber anatomy maximum bend

Determine prior distributions from anatomy fiber skeletons from a population others

Dr. Van Wedeen, MGH
Bayesian Framework For Global Tractography

J.R.S. Jbabdi, et al.

MAP Estimation In Image Reconstruction

Human brain MRI. (a) The original LR data. (b) Zero-padding interpolation. (c) SR with box-PSF. (d) SR with Gaussian-PSF

From: A. Greenspan in
Improved ASL Perfusion Results

By Dr. John Kornak, UCSF

zDFT = zero-filled DFT
Bayesian Automated Image Segmentation

Bruce Fischl, MGH
Population Atlases As Priors

Dr. Sarang Joshi, U Utah, Salt Lake City
## Imaging Software Using MLE And MAP

<table>
<thead>
<tr>
<th>Packages</th>
<th>Applications</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoxBo</td>
<td>fMRI</td>
<td>C/C++/IDL</td>
</tr>
<tr>
<td>MEDx</td>
<td>sMRI, fMRI</td>
<td>C/C++/Tcl/Tk</td>
</tr>
<tr>
<td>SPM</td>
<td>fMRI, sMRI</td>
<td>matlab/C</td>
</tr>
<tr>
<td>iBrain</td>
<td></td>
<td>IDL</td>
</tr>
<tr>
<td>FSL</td>
<td>fMRI, sMRI, DTI</td>
<td>C/C++</td>
</tr>
<tr>
<td>fmrstat</td>
<td>fMRI</td>
<td>matlab</td>
</tr>
<tr>
<td>BrainVoyager</td>
<td>sMRI</td>
<td>C/C++</td>
</tr>
<tr>
<td>BrainTools</td>
<td></td>
<td>C/C++</td>
</tr>
<tr>
<td>AFNI</td>
<td>fMRI, DTI</td>
<td>C/C++</td>
</tr>
<tr>
<td>Freesurfer</td>
<td>sMRI</td>
<td>C/C++</td>
</tr>
<tr>
<td>NiPy</td>
<td></td>
<td>Python</td>
</tr>
</tbody>
</table>