Basics of Medical Imaging Informatics: Estimation Theory

Norbert Schuff
Professor of Radiology
VA Medical Center and UCSF
Norbert.schuff@ucsf.edu
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.

UCSF VA Department of Radiology & Biomedical Imaging
What Is The Focus Of This Course?

Learn using computational tools to maximize information for knowledge gain:

Pro-active

- Improve Image Measurements
- Data Collection

Re-active

- Extract Information
- Compare with model

Refine Model

UCSF VA Medical Imaging Informatics 2009, Nschuff Course # 170.03
Slide 3/31
Challenge: Maximize Information Gain

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

2. Q: How can we measure (quantify) information?
   A: *Apply information theory* (2\textsuperscript{nd} lecture next week)
Estimation Theory: Motivation Example I

Gray/White Matter Segmentation

Hypothetical Histogram

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

Goal: Capture dynamic signal on a static background

D. Feinberg Advanced MRI Technologies, Sebastopol, CA
Estimation Theory: Motivation Example III

Diffusion Imaging

• Sensitive to random motion of water
• Probes structures on a microscopic scale

Goal:
Capture directions of fiber bundles

Quantitative Diffusion Maps

Microscopic tissue sample

Dr. Van Wedeen, MGH
Basic Concepts of Modeling

\( \Theta \): target of interest and unknown

\( \rho \): 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

\[ \varphi_N = H\theta_M + noise_N \]

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

Note:
1) we make no assumption about \( \Theta \)
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:

\[ \varphi_N - H\theta_M = noise_M \]

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

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

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

- 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”
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 \[ \hat{\theta}_4 \]
Likelihood Model

Pretend we know something about $\Theta$

We perform measurements for all possible values of $\Theta$

We obtain the likelihood function of $\Theta$ given our measurements $\rho$

Note:
$\Theta$ is random
$\varphi$ is a fixed parameter
Likelihood is a function of both the unknown $\Theta$ and known $\varphi$
Likelihood Model (cont’d)

\[ L_\varphi (\Theta) = p(\varphi_N | \Theta) \]

New Goal:
Find an estimator which gives the most likely probability distribution underlying \( L_\varphi (\Theta) \)

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 probability distribution underlying $x_N$.

$$\widehat{\Theta}_{MLE} = \max p(\varphi_N | \Theta)$$

$\Theta_{MLE}$ can be found by taking the derivative of Likelihood $F$

$$\left. \frac{d}{d\Theta} \ln p(\varphi_N | \Theta) \right|_{\Theta=\Theta_{MLE}} = 0$$
Example I: MLE Of Normal Distribution

Normal distribution

\[ p\left(\varphi_N \mid \Theta, \sigma^2\right) = \exp\left[\frac{1}{2\sigma^2} \sum_{j=1}^{N}(\varphi(j) - \Theta)^2\right] \]

log of the normal distribution (normD)

\[ \ln p\left(\varphi_N \mid \Theta, \sigma^2\right) = \frac{1}{2\sigma^2} \sum_{j=1}^{N}(\varphi(j) - \Theta)^2 \]

MLE of the mean (1st derivative):

\[ \frac{d}{d\Theta_{MLE}} \ln p(\Theta) = \frac{1}{4\hat{\sigma}^2} \sum_{j=1}^{N}(\varphi(j) - \Theta_{MLE}) = 0 \]

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

Distribution function $f(y|n,w)$:
- $n =$ number of tosses
- $w =$ probability of success

$f(y|n=10,w=7)$
$f(y|n=10,w=3)$
MLE Of Coin Toss (cont’d)

Goal:
Given the observed data $f(y|w=0.7, n=10)$, find the parameter $\Theta_{MLE}$ that most likely produced the data.

$$L(\Theta_{MLE} \mid y = 7, n = 10)$$

For a fair coin $\Theta_{MLE} = 0.5$
MLE Of Coin Toss (cont’d)

Likelihood function of coin tosses

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

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

\[ L(\Theta \mid n = 10, w = 7) = \frac{10!}{(7!)(10-7)!} \cdot 0.5^7 (1-0.5)^{10-7} = 0.12 \]

unfair coin \( \Theta = 0.6 \)

\[ L(\Theta \mid n = 10, w = 7) = \frac{10!}{(7!)(10-7)!} \cdot 0.6^7 (1-0.6)^{10-7} = 0.21 \]

log likelihood function

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

Evaluate MLE equation (1st derivative)

\[
\frac{d \ln L}{d \Phi_{MLE}} = \frac{y}{\Phi_{MLE}} - \frac{(n-y)}{1-\Phi_{MLE}} = 0
\]

\[
= \frac{y - n\Phi_{MLE}}{\Phi_{MLE}(1 - \Phi_{MLE})} = 0 \Rightarrow \Phi_{MLE} = \frac{y}{n}
\]

According to the MLE principle, the distribution \( f(y/n) \) for a given \( n \) is the most likely distribution to have generated the observed data of \( y \).
Relationship between MLE and LSE

Assume: \( \Theta \) is independent of \( noise_N \)
MLE and \( noise_N \) have the same distribution

\[
p_\theta (\varphi_N \mid \Theta) = p_{noise} (\varphi_N - H\Theta \mid \Theta)
\]

\( noise_N \) is zero mean and gaussian

\( p(\rho \mid \Theta) \) is maximized when LSE is minimized
Bayesian Model

Now, the daemon comes into play, but we know the daemon’s preferences for $\Theta$ (prior knowledge).

$$prior(\Theta) = p(\Theta)$$

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

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

\[ \text{posterior}_\varphi (\Theta) = C \cdot L_\varphi (\varphi_N | \Theta) \cdot p(\Theta) \]
Maximum A-Posteriori (MAP) Estimator

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

$$\Theta_{\text{MAP}} = \max L(\phi_N | \Theta) p(\phi_N)$$

$\Theta_{\text{MAP}}$ can be found by taking the partial derivative

$$\frac{d}{d\Theta} \ln L(\phi_N | \Theta) p(\Theta) = \frac{\partial}{\partial \Theta} \ln L(\phi_N | \Theta) + \frac{\partial}{\partial \Theta} \ln p(\Theta) = 0$$
Example III: MAP Of Normal Distribution

The sample mean of MAP is:

\[
\hat{\Phi}_{\text{MAP}} = \frac{\sigma^2_{\mu}}{\sigma^2_{\varphi} + T\sigma^2_{\mu}} \sum_{j=1}^{N} \varphi(j)
\]

If we do not have prior information on \(\mu, \sigma_{\mu} \to \infty \) or \(T \to \infty\)

\[
\hat{\mu}_{\text{MAP}} \Rightarrow \hat{\mu}_{\text{ML}}, \hat{\mu}_{\text{LSE}}
\]
Posterior Distribution and Decision Rules

$p(\Theta | \rho)$

$\Theta_{MSE}$ $\Theta_{MAP}$
Decision Rules

Measurements → Likelihood function → Posterior Distribution

Prior Distribution → Gain Function → Result
Some Desirable Properties of Estimators I:

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

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

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

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

What happens to MSE when estimator is biased?

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

- **Variance**
- **Bias**
Some Desirable Properties of Estimators II:

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

\[ C_\theta = E\left( (\tilde{\Phi}_i - \Phi_i)(\tilde{\Phi}_k - \Phi_k)^T \right) \leq \text{some.very.small.value} \]
Example:
Properties Of Estimators **Mean** and **Variance**

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

The sample mean is an unbiased estimator of the true mean.

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

The variance is a consistent estimator because it approaches zero for large number of measurements.
Properties Of MLE

- **is consistent**: the MLE recovers asymptotically the true parameter values that generated the data for \( N \to \infty \);

- **Is efficient**: The MLE achieves asymptotically the minimum error (= max. information)
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, i.e. many measurements, 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
Estimation Theory: Motivation Example II

Goal: Capture dynamic signal on a static background

Improvements to identify the dynamic signal:

Design likelihood functions based on auto-correlations anatomical information

Determine prior distributions from serial measurements multiple subjects anatomy

D. Feinberg Advanced MRI Technologies, Sebastopol, CA
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 fiber anatomy

Determine prior distributions from anatomy fiber skeletons from a population others

Dr. Van Wedeen, MGH
MAP Estimation in Image Reconstructions with Edge-Preserving Priors
MAP in Image Reconstructions with Edge-Preserving Priors

For DTI, use the fact that coregistered DTI images have common edge features:

Dr. Justin Haldar Urbana-Champaign
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
Segmentation Using MLE

A: Raw MRI
B: SPM2
C: EMS
D: HBSA

from
Habib Zaidi, et al,
NeuroImage 32
(2006) 1591 – 1607
Population Atlases As Priors

Dr. Sarang Joshi, U Utah, Salt Lake City
Population Shape Regressions Based Age-Selective Priors

Age = 29  33  37  41  45  49

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>fmristat</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>
Literature

2 Probability Distributions 67
2.1 Binary Variables ......................................... 68
2.1.1 The beta distribution ................................... 71
2.2 Multinomial Variables ..................................... 74
2.2.1 The Dirichlet distribution .............................. 76
2.3 The Gaussian Distribution ............................... 78
2.3.1 Conditional Gaussian distributions ................. 85
2.3.2 Marginal Gaussian distributions ...................... 88
2.3.3 Bayes' theorem for Gaussian variables .......... 90
2.3.4 Maximum likelihood for the Gaussian ............ 93
2.3.5 Sequential estimation ................................ 94
2.3.6 Bayesian inference for the Gaussian .............. 97
2.3.7 Student's t-distribution ............................... 102
2.3.8 Periodic variables ................................... 105
2.3.9 Mixtures of Gaussians ............................... 110
2.4 The Exponential Family ................................. 113
2.4.1 Maximum likelihood and sufficient statistics .... 116
2.4.2 Conjugate priors .................................... 117
2.4.3 Noninformative priors ............................... 117
2.5 Nonparametric Methods ................................. 120
2.5.1 Kernel density estimators .............................. 122
2.5.2 Nearest-neighbour methods ......................... 124
Exercises .................................................. 127
3 Linear Models for Regression 137
3.1 Linear Basis Function Models ............................ 138
3.1.1 Maximum likelihood and least squares .......... 140
3.1.2 Geometry of least squares ........................... 143
3.1.3 Sequential learning ................................. 143
3.1.4 Regularized least squares ............................ 144
3.1.5 Multiple outputs ................................. 146
3.2 The Bias-Variance Decomposition .................... 147
3.3 Bayesian Linear Regression ............................. 152
3.3.1 Parameter distribution ............................... 153
3.3.2 Predictive distribution ............................... 156
3.3.3 Equivalent kernel ................................. 157
3.4 Bayesian Model Comparison .......................... 161
3.5 The Evidence Approximation .......................... 161
4.1 Basic concepts 78
4.2 Properties of estimators 80
4.3 Method of moments 84
4.4 Least-squares estimation 86
4.4.1 Linear least-squares method 86
4.4.2 Nonlinear and generalized least squares * 88
4.5 Maximum likelihood method 90
4.6 Bayesian estimation * 94
4.6.1 Minimum mean-square error estimator 94
4.6.2 Wiener filtering 96
4.6.3 Maximum a posteriori (MAP) estimator 97
4.7 Concluding remarks and references 99
Problems 101
5 Information Theory 105
5.1 Entropy 105
5.1.1 Definition of entropy 105
5.1.2 Entropy and coding length 107
5.1.3 Differential entropy 108
5.1.4 Entropy of a transformation 109
5.2 Mutual information 110
5.2.1 Definition using entropy 110
5.2.2 Definition using Kullback-Leibler divergence 110
Part I: Theories of Perception and Learning
Chapter 1: Bayesian Modelling of Visual Perception, by P. Mamassian, M. Landy and L. Maloney
Chapter 2: Vision, Psychophysics, and Bayes, by P. Schrater and D. Kersten
Chapter 3: Visual Cue Integration for Depth Perception, by R. Jacobs
Chapter 4: Velocity Likelihoods in Biological and Machine Vision, by Y. Weiss and D. Fleet
Chapter 5: Learning Motion Analysis, by W. Freeman, J. Haddon and E. Pasztor
Chapter 6: Information Theoretic Approach to Neural Coding and Parameter Estimation: A Perspective, by J.-P. Nadal
Chapter 7: From Generic to Specific: An Information Theoretic Perspective on the Value of High-Level Information, by A. Yuille and J. Coughlan
Chapter 8: Sparse Correlation Kernel Reconstruction and Superresolution, by C. Papageorgiou, F. Girosi and T. Poggio
Literature
Literature

Mathematical

Signal Processing

Statistics: