An Overview of Probabilistic Modeling and Inference in Medical Image Analysis

Biomedical Image Analysis Summer School: Modalities, Methodologies & Clinical Research, Paris 2012

William (Sandy) Wells

Associate Professor of Radiology
Surgical Planning Laboratory
Harvard Medical School and Brigham and Women’s Hospital

Affiliated Faculty: Harvard – MIT Division of Health Sciences and Technology

Research Scientist – MIT CSAIL
Outline

• Bayesian approach
• graphical Models
• inference
• simple estimators
• overview of mechanisms of Medical Image Analysis
  – summarize with graphical models
Bayesian Approach

• modeling with (lots of) random variables
  – and associated probability distributions
• explicit assumptions
  – especially priors
    • when appropriate: uninformative priors
• marginalization
• emphasis on finding posterior distributions
Graphical Models

- clarify assumptions
- high level summary of mechanism
- convenient for design of methods
- easy to express
  - hierarchy
  - sharing of variables and parameters
  - … which help to control complexity of model
Distribution Notation Examples

\[ p(\alpha; \mu, \sigma^2) = N(\alpha; \mu, \sigma^2) \]

\[ \alpha \sim N(\mu, \sigma^2) \]

\[ p(RV \cdots | RV \cdots; PARAMS \cdots) \]

\[ p(\alpha | \beta) = N(\alpha; \beta, \sigma^2) \]

\[ \alpha | \beta \sim N(\beta, \sigma^2) \]
Random Variables and Parameters

\[ \mu \quad \nu \]  
unknown parameters

\[ \mu \quad \nu \]  
known parameters

\[ \alpha \]  
\( \alpha \sim P(\cdot) \)  
observed random variables

\[ \alpha \]  
latent (hidden) random variables

\[ \alpha \beta \quad \alpha \beta \]  
sometimes package two vars for brevity
Links

\[ p(\alpha, \beta, \delta, \gamma) = p(\gamma|\delta)p(\delta|\alpha, \beta)p(\alpha)p(\beta) \]

(In general: \[ p(\alpha, \beta, \delta, \gamma) = p(\gamma|\alpha, \beta, \delta)p(\delta|\alpha, \beta)p(\alpha|\beta)p(\beta) \])
Joint Probability Fully Characterizes Random Variables

suppose we have \( p(\alpha, \beta, \delta, \gamma) \)

then, e.g.,

\[
p(\delta | \gamma) = \frac{p(\delta, \gamma)}{p(\gamma)} \quad \text{def. conditional probability}
\]

\[
p(\delta, \gamma) = \sum_{\alpha, \beta} p(\alpha, \beta, \delta, \gamma) \quad \text{marginalization}
\]

\[
p(\gamma) = \sum_{\alpha, \beta, \delta} p(\alpha, \beta, \delta, \gamma)
\]
Plates

$\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \ldots \quad \alpha_N$

$\alpha_i$

$N$
Summarize Inference Methods

- Maximum Likelihood (ML) estimation
- Maximum A-Posteriori (MAP) estimation
- EM Algorithm (EM)
- Variational Bayes (VB)
- Markov Chain Monte Carlo (MCMC)
Slide Layout

graphical model

summary of inference

summary of distributions (omit conditioning)
Maximum Likelihood (ML) Parameter Estimation

\[ y_i \sim N(\mu, \sigma^2) \]

\[ p(y; \mu, \sigma^2) = \prod_i N(y_i; \mu, \sigma^2) \]

\( \sigma^2 \) known; observe \( y \), estimate \( \mu \)

\[ \hat{\mu} = \arg \max_\mu p(y; \mu, \sigma^2) \]
Maximum A-Posteriori (MAP) Parameter Estimation

\[ \mu' \quad \sigma'^2 \]

\[ \mu \quad \sigma^2 \]

\[ y_i \sim N(\mu, \sigma^2) \]

\[ \mu \sim N(\mu', \sigma'^2) \]

\[ \mu', \sigma'^2, \sigma^2 \text{ known; observe } y, \text{ estimate } \mu \]

\[ \hat{\mu} = \arg \max_{\mu} p(\mu | y) \propto p(y | \mu) p(\mu) \]
Bayes’ Approach: Characterize Posterior Distribution

\[ \mu \sim N(\mu', \sigma'^2) \]

\[ y_i \sim N(\mu, \sigma^2) \]

\( \mu', \sigma'^2, \sigma^2 \) known; observe \( y \), estimate \( p(\mu|y) \)

\[ p(\mu|y) = z^{-1} p(y|\mu) p(\mu) \]

\[ z = \int_{\mu} p(y|\mu)p(\mu) \]
Student T Distribution by Marginalization

\[ \mu' \]
\[ \sigma'^2 \]
\[ \mu \]
\[ \sigma^2 \]
\[ y_i \]
\[ N \]

latent mean:
\[ \mu \sim N(\mu', \sigma'^2) \]

latent variance:
\[ \sigma^2 \sim \frac{1}{\sigma^2} \]

(improper uninformative prior)

observed data:
\[ y_i \sim N(\mu, \sigma^2) \]

\( \mu', \sigma'^2 \) known; observe \( y \), estimate

\[
p(\mu|y) = \int p(\mu, \sigma^2|y) d\sigma^2 \propto \int p(y|\mu, \sigma^2)p(\sigma^2) d\sigma^2
\]

\[ \mu|y \sim \text{Student-t distribution} \]
Now is the time to speak of heroes
Samuel Adams

• 1722 – 1803
• Brewer
• Patriot
William Gosset

• 1876 – 1937
• Brewer
• Statistician
William Gosset

• 1876 – 1937
• Brewer
  – Employed by Guinness
    • Published under “Student”
    • Industrial Scientist
    • Chief of London Brewery

• Statistician
  – T test
EM Algorithm

EM Algorithm

• ML/MAP parameter estimation
• partition random variables
  – hidden
  – observed
• iterate
  – e step: re-estimate distribution on hidden variables
  – m step: ML/MAP estimate of parameters given estimate of distribution on hidden variables
• converges to marginal ML/MAP est.
Variational Bayes

- make parametric approximations to key distributions
  - mean field approach: use factorized distns
- iterate: cycle through hierarchy
  - update parameters of distributions
Markov Chain Monte Carlo (MCMC)

• random sampling method
• construct “proposal distribution”
  – easy to sample from
  – approximate true posterior
• use Metropolis-Hastings method to accept or reject samples
• resulting samples asymptotically from true posterior
• characterize posterior distribution by samples
• very general (very slow)
Summarize Mechanisms

• methods of Medical Image Analysis
  – STAPLE
  – segmentation
  – registration
  – combined registration / segmentation

• illustrate evolution with graphical models
  – somewhat parochial choice of material
  – take some liberties
    • simplifications
STAPLE

- How to validate segmentation / registration?
- Human raters
  - Do not agree
- STAPLE: combine results of human raters
  - Infer rater performance
  - Estimate missing “true” segmentation

Bernoulli Distribution

binary random variable, e.g., coin flipping

\[ \Gamma \in \{0, 1\} \]
\[ \Gamma \sim B(\theta) \]

\[ \theta \in [0, 1] \]

\[ B(\Gamma; \theta) \doteq \theta^{\Gamma} (1 - \theta)^{(1-\Gamma)} \]
STAPLE

\[ \Gamma_i^T: \text{latent "true" labels} \]

\[ \alpha_j \beta_j: \text{rater params} \]
(sensitivity, specificity)

labels given by raters:
\[ \Gamma_{ij} \sim B((1 - \beta_j)^{T_i} \alpha_j^{(1 - \Gamma_i^T)}) \]

EM:
\[ \hat{\alpha}\hat{\beta} = \arg \max_{\alpha, \beta} p(\alpha \beta | \Gamma) \]
\[ p(\Gamma^T | \Gamma, \hat{\alpha}, \hat{\beta}) \]
Segmentation of Medical Images

- ML, MAP
- Atlas-Based Segmentation
- “EM Segmentation”
  - with tissue parameter estimation
  - with atlas registration
- “latent atlas” segmentation
- Left out: priors models to combat noise: Markov Random Fields
  - solvers, e.g., Mean Field or …
Segmentation of Medical Images

• Early work:
  – Classification based on pixel intensities

• e.g.,
ML Pixel Classification

\[ \hat{\Gamma}_i = \arg \max_{\Gamma_i} p(y_i | \mu_{\Gamma_i}, \sigma^2) \]

tissue label:
\[ \Gamma_i \in \{0, 1\} \]

image intensity:
\[ y_i \sim N(\mu_{\Gamma_i}, \sigma^2) \]
MAP Pixel Classification

\[ \hat{\Gamma}_i = \arg \max_{\Gamma_i} p(\Gamma_i | y_i) \]

\( \theta \): Bernoulli parameter

tissue labels:
\( \Gamma_i \sim B(\theta) \)

image intensities:
\( y_i \sim N(\mu_{\Gamma_i}, \sigma^2) \)
Atlas-Based Segmentation

\[ \theta_i: \text{ Bernoulli parameter (the "atlas") } \]

\[ \Gamma_i \sim B(\theta_i) \]

\[ y_i \sim N(\mu_{\Gamma_i}, \sigma^2) \]

MAP: \[ \hat{\Gamma}_i = \arg \min_{\Gamma_i} p(\Gamma_i | y_i) \]
Atlas-Based Segmentation

\[ \theta_i: \text{ Bernoulli parameter} \]
\[ \text{the "atlas"} \]
\[ \Gamma_i \sim B(\theta_i) \]

\[ y_i \sim N(\mu_{\Gamma_i}, \sigma^2) \]

MAP: \[ \hat{\Gamma}_i = \arg \max_{\Gamma_i} p(\Gamma_i | y_i) \]
Early Atlas-Based Segmentation

• (used knn classifier)
Average Brain Models

• Construct a spatial prior model by averaging tissue distributions over a population [MNI].
Training Subjects

Generic Subjects

Segmentations

Spatial Prior

- register MRIs
- align Segmentation
- produce prior
P(white matter | x y)
Intensity Inhomogeneities in MRI

- MRI signal derived from RF signals...
- Intra Scan Inhomogeneities
  - “Shading” … from coil imperfections
  - interaction with tissue?
- Inter Scan Inhomogeneities
  - Auto Tune
  - Equipment Upgrades
EM Segmentation

\[ \hat{\beta} = \arg \max p(\beta | y) \]
\[ p(\Gamma | y\hat{\beta}) \]

\( \theta_i \): Bernoulli parameter

tissue label:
\( \Gamma_i \sim B(\theta_i) \)

bias field:
\( \beta \sim N(0, \Sigma) \)

image intensities:
\( y_i \sim N(\mu_{\Gamma_i} - \beta_i, \sigma^2) \)
“EM Segmentation”

Dual Echo Longitudinal Study

PDw

T2w
Tissue classification

No Intensity Correction

EM Segmentation
Surface Coil Example
EM-Segmentation

E-Step
Compute tissue posteriors using current intensity correction.

M-Step
Estimate intensity correction using residuals based on current posteriors.

Provided by T Kapur
Segmentation with tissue characterization

EM Seg. and Tissue Characterization

\[ \theta: \text{Bernoulli parameter} \]

\[ \Gamma_i \sim B(\theta) \]

bias field:
\[ \beta \sim N(0, \Sigma) \]

image intensities:
\[ y_i \sim N(\mu_{\Gamma_i} - \beta_i, \sigma^2) \]

EM:
\[ \hat{\beta}\mu\sigma^2 = \arg\max p(\beta | y; \mu\sigma^2) \]
\[ p(\Gamma | y, \hat{\beta}\mu\sigma^2) \]
Joint Segmentation and Atlas Registration

Joint Segmentation / Atlas Registration

T: pixels $\rightarrow$ pixels
$\theta_i$: Bernoulli Params (the "atlas")

Tissue labels:
$\Gamma_i \sim B(\theta_{T(i)})$

Bias field:
$\beta \sim N(0, \Sigma)$

Image intensities:
$y_i \sim N(\mu_{\Gamma_i} - \beta_i, \sigma^2)$

EM:
$\hat{\beta}^T = \underset{\beta_T}{\operatorname{arg~max}} \ p(\beta | y; T)$

$p(\Gamma | y; \hat{\beta}^T)$
segmentation, registration example
Initial registration - segmentation
10 iterations
15 iterations
20 iterations
Latent Atlas segmentation

- Constructing atlases can be expensive, time consuming

Latent Atlas Segmentation

\[ \theta_i: \text{ Bernoulli params (the "atlas") } \]

\[ \Gamma_{ij} \sim B(\theta_i) \]

\[ \beta_j \sim N(0, \Sigma) \]

\[ y_{ij} \sim N(\mu \Gamma_{ij} - \beta_{ij}, \sigma^2) \]

EM:
\[
\hat{\beta} \theta = \arg \max \ p(\beta | y; \theta) \\
p(\Gamma | y, \hat{\beta} \theta)
\]
Medical Image Registration

- ML approach
- generative model for minimum entropy registration
  - incorporate prior info
- groupwise registration: Congealing
- VB approach to estimating posterior on deformations
- MCMC approach to estimating posterior on deformations
A Marginalized MAP Approach and EM Optimization for Pair-Wise Registration

IPMI 2007

Lilla Zollei
Mark Jenkinson
Samson Timoner
William Wells
Probability on Intensity Pairs

\[ p(u_i, v_i | \Theta) = \text{Mult}(\mathcal{B}(u_i, v_i); 1, \Theta) = \theta_{\mathcal{B}(u_i, v_i)} \]

\[ 0 \leq \theta_j \leq 1 \quad \sum_j \theta_j = 1 \]

- \( \mathcal{B}(\cdot, \cdot) \) : Bin index
- Multinomial Distribution
  - One trial
Joint Image Histogram

Data: \((u(x_i), v(x_i))\)

Bin Counts, e.g., \(n_g = 2\)
ML Registration

$T$: pixel coords $\rightarrow$ pixel coords

$\theta_j$: multinomial param

Joint intensities:

$u(x_i), \; v \circ T(x_i) \sim M(1, \theta)$

multinomial in one trial

ML: $\hat{T} = \arg \max_T \sum_j n_j(T) \log(\theta_j)$
ML Registration

Estimate Multinomial Distribution from Data

• Maximum Likelihood Method:
  – get pre-registered images
  – Histogram the data
  – Set the parameters to be normalized histogram counts
Strong vs. Weak Models

• Capture / Bias Tradeoff
  – Strong model, e.g.: ML
    • Robust: large capture
    • Model may be inaccurate for new images
      – Less accurate estimate of $T$
  – Weak Model, e.g.: min Entropy, max MI
    • Less Robust: smaller capture
    • More Accurate estimate of $T$
Dirichlet Prior on $\Theta$

$$p(\Theta|w) = \text{Dir}(\Theta; w)$$

$$= \frac{1}{Z(w)} \prod_{i=1}^{g} \theta_i^{(w_i-1)} = \Gamma(w_0) \prod_j \frac{\theta_j^{(w_j-1)}}{\Gamma(w_j)}$$

$$w_0 = \sum_i w_i$$

- Conjugate prior for Multinomial
- Multi-category generalization of Beta
- Parameterized by pseudo-data counts
- Laplace prior: $W_i = 1$
Examples: order 3 Dirichlet

- Distribution on Multinomial Parameters:

\[ 0 \leq \theta_i \leq 1 \quad \theta_1 + \theta_2 + \theta_3 = 1 \]

Bayesian statistics: a concise introduction. Kevin P. Murphy
(tech note on web, use google)
Dirichlet distribution...

• Conjugate property:

\[ p(\Theta|x) \propto \text{Mult}(x; N, \Theta) \cdot \text{Dir}(\Theta; w) \]
\[ \propto \text{Dir}(\Theta; n(x) + w) \]

• Mean and Variance:

\[ E[\theta_i] = \frac{w_i}{w_0} \]
\[ \text{Var}[\theta_i] = \frac{w_i(w_0 - w_i)}{w_0^2(w_0 + 1)} \]
## Discrete Distributions

<table>
<thead>
<tr>
<th></th>
<th>Single trial</th>
<th>Multiple trials</th>
<th>Conjugate Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary data</td>
<td><em>Bernoulli</em></td>
<td><em>Binomial</em></td>
<td><em>Beta</em></td>
</tr>
<tr>
<td>Number of data values &gt; 2</td>
<td><em>Multinomial (in one trial)</em>&lt;br&gt;PMF&lt;br&gt;Probability Distribution</td>
<td><em>Multinomial</em></td>
<td><em>Dirichlet</em></td>
</tr>
</tbody>
</table>
Minimum Entropy Registration

Laplace Prior: \( w_i = 1 \)

\( T \): pixel coords. \( \rightarrow \) pixel coords.

multinomial parameter:
\( \theta \sim DIR(w) \)

joint intensities:
\( u(x_i), \ v \circ T(x_i) \sim M(1, \theta) \)

multinomial in one trial

pixels \( x_i \)

ML: \( \hat{T} = \arg \min_T \left[ H[M(1, \frac{n(T)+1}{2N})] \right] \) using Stirling’s appx
Min. Entropy Reg. with Prior Info

\[ w \text{ estimated from training data} \]

\[ T: \text{pixel coords.} \rightarrow \text{pixel coords.} \]

\[ \theta \sim \text{DIR}(w) \]

Joint intensities:
\[ u(x_i), \ v \circ T(x_i) \sim M(1, \theta) \]

Multinomial in one trial

ML: \[ \hat{T} = \arg \min_T \left[ H[M(1, \frac{n(T) + w}{N+w_0})] \right] \] using Stirling’s appx
Historical Precursor

- Voxel Similarity Measures for Automated Image Registration. VBC 1994
  - Hill D., Studholme, C., and Hawkes, D.
  - Meeting at Mayo Clinic, Oct 4 – 7 1994
  - 3rd order moments (and other) measures
  - MOVIE by Colin Studholme…
Early Entropy / MI Registration

- Minimum Entropy and Registration:

\[ I(U, V) = H(U) + H(V) - H(UV) \]

- Maximum Mutual Information Registration
MRI-CT registration, ~1996
Registration of Video and 3D Model

Surgical Example: Finite Element Deformation Model
S. Timoner: Compact Representations for Fast Non-rigid Registration of Medical Images (PhD’03)
Groupwise Registration

• Generalizing $p(u,v)$ to a collection of images is difficult

• Alternative: model $p(u_i)$ at each pixel
Congealing…


Congealing

$w$: Laplace prior

multinomial parameter:
$\theta_j \sim \text{DIR}(w)$

joint intensity:
$u_i \circ T_i(x_j) \sim M(1, \theta_j)$

multinomial in one trial

ML: $\hat{T} = \arg \min_T \sum_j H[M(\frac{n_j(T)}{N_j})]$  using Stirling’s appx
127 Adult MRI
Before and After Congealing

Data set: 127 T1w MRI; [256x256x124] with (0.9375, 0.9375, 1.5) mm³ voxels;
Experiment: 3 levels; 12-param. affine; N = 800-1600; iter = 250; time = 6hrs
30 FBIRN Subjects

Affine

B-Spline

Joint Congealing Two Infant Populations

- 17 full term
- 22 pre term

- Group analysis of transform params
  - Significant difference in shape
Uncertainty in Non-Rigid Registration

- characterize the poster distribution
- summarize
Motivation: Uncertainty

• **Uncertainty**: “A state of having limited knowledge where it is impossible to exactly describe existing state or future outcome. More than one possible outcome.”

• If you tell a surgeon where something is, it would be nice to have “error bars” on that
• Want to move away from a deterministic approach
• Formulate registration in a Bayesian framework
• Quantify the most probable deformation
• All modes of the posterior distribution
• The associated deformation uncertainty
Characterize Posterior on Deformations in Nonrigid Registration

$p(u|m) \propto \int \int p(u \tau_r \tau_s | fm) d\tau_r d\tau_s$

$\mu_r, \lambda_r, \mu_s, \lambda_s$ set for uninformative priors

$\ln \tau_r \sim N(\mu_r, \lambda_r)$

$\ln \tau_s \sim N(\mu_s, \lambda_s)$

deformation field:

$p(u) \propto \exp\left(-\frac{E_r}{\tau_r}\right)$

$E_r$: Lin. Elastic Def. Energy of $u$

$f, m$: fixed and moving images

$p(m|fu \tau_s) \propto \exp\left(-\frac{E_s}{\tau_s}\right)$

$E_s = SSE(m, f \circ u)$
fMRI: Marginal Volumes

- For each deformation sample: increment every voxel that is inside the deformed fMRI activated volume
- Determines the probability for a voxel to be in the fMRI activation volume
- Iso-contours define confidence bounds
DTI: Marginal Visitation Volume

- Trace the deformed tractography streamlines through the volume and increment values in each voxel a streamline crosses.
- This is the marginal distribution that a fiber crosses a voxel.
Cumulative Dose Distributions

Submandibular Gland
references


additional recent work in registration

- characterize posterior distribution on deformation in non-rigid registration
- automatically estimate regularization parameter (by marginalization)
- extended existing FNIRT registration
- improved group differences
Simson et al.

\[ y = t(x, w) + e \]

\[ \lambda \sim Ga(s_0, c_0) \]  
(uninformative)

\[ \phi \sim Ga(a_0, b_0) \]  
(uninformative)

\[ w \sim N(0, (\lambda \Lambda)^{-1}) \]  
\( \Lambda \): regularization kernel

\[ x, y: \text{ images} \]

\[ y_i | (x \circ w)_i, \phi \sim N(y_i - (x \circ w)_i, \phi) \]

Variational Bayes:  
\[ p(w|xy) \approx N(\mu, \Psi) \]
emergent themes

• evolution
  – initial models with fixed known parameters
  – online estimation of parameters
  – parameters become latent variables with priors
    • parameters are known…
    • parameters correspond to uninformative priors

• more emphasis on posterior distributions
references

• Pattern Recognition and Machine Learning, Christopher Bishop, Springer 2000.


references…

• HST 582 (6.555)
  – Biomedical Signal and Image Processing.
  – MIT Open Courseware
    • signal processing
    • elementary medical image analysis
      – Segmentation, Registration, etc
    • Lecture notes
    • MATLAB “lab” experiments
The End