Curve Fitting & Multisensory Integration

Using Probability Theory

Hannah Chen, Rebecca Roseman and Adam Rule | COGS 202 | 4.14.2014

Overheard at Porters

“You’re optimizing for the wrong thing!”

What would you say?

What makes for a good model?

- Best performance?
- Fewest assumptions?
- Most elegant?
- Coolest name?

Agenda

Motivation	Finding Patterns
Tools	Probability Theory
Applications	Curve Fitting Multimodal Sensory Integration

The Motivation

FINDING PATTERNS

You have data... now find patterns

You have data... now find patterns

Unsupervised

x = data (training)

y(**x**) = model

Clustering

Density estimation

You have data... now find patterns

Unsupervised

$\mathbf{x}$ = data (training)

$y(\mathbf{x})$ = model

Clustering

Density estimation

Supervised

$\mathbf{x}$ = data (training)

$\mathbf{t}$ = (target vector)

$y(\mathbf{x})$ = model

Classification

Regression

You have data... now find patterns

Unsupervised

$\mathbf{x}$ = data (training)

$y(\mathbf{x})$ = model

Clustering

Density estimation

Supervised

$\mathbf{x}$ = data (training)

$\mathbf{t}$ = (target vector)

$y(\mathbf{x})$ = model

Classification

Regression

Important Questions

What kind of model is appropriate?

What makes a model accurate?

Can a model be too accurate?

What are our prior beliefs about the model?

The Tools

PROBABILITY THEORY

Properties of a distribution

x = event

p(x) = prob. of event

1. $$p(\mathbf{x}) \geqslant 0$$
2. $$\int p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$$

Rules

Sum

$$p(X) = \sum_Y p(X, Y)$$

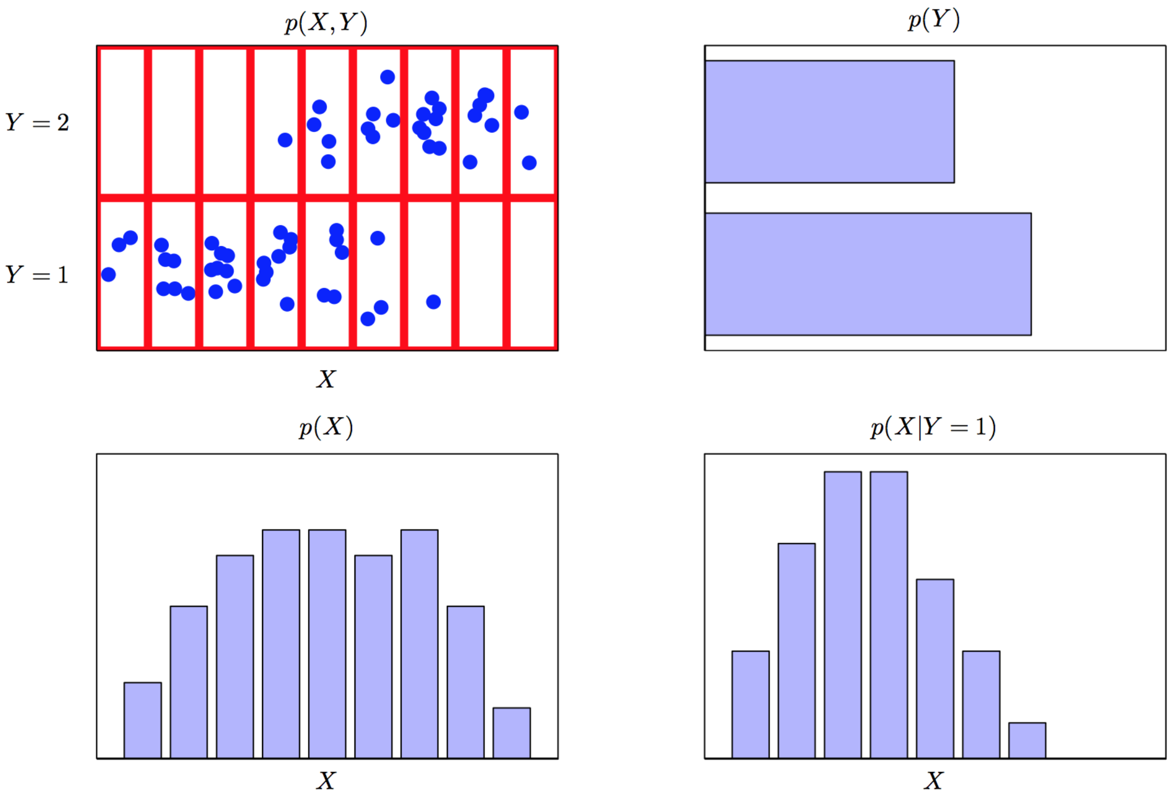
$$p(x) = \int p(x, y) \, \mathrm{d}y$$

Product

$$p(X, Y) = p(Y|X)p(X)$$

$$p(x, y) = p(y|x)p(x)$$

Example (Discrete)



Bayes Rule (review)

posterior likelihood prior

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

evidence

Probability Density vs. Mass Function

$$\int_a^b f(x)dx$$

PDF	PMF
continuous	discrete
Intuition: How much probability f(x) concentrated near x per length dx, how dense is probability near x	Intuition: Probability mass is same interpretation but from discrete point of view: f(x) is probability for each point, whereas in PDF f(x) is probability for an interval dx
Notation: p(x)	Notation: P(x)
p(x) can be greater than 1, as long as integral over entire interval is 1	

Expectation & Covariance

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x.$$

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2 \right]$$

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\} \right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Application

CURVE FITTING

Curve Fitting

Observed Data: Given n points (x,y)

Curve Fitting

Observed Data: Given n points (x,t)

Textbook Example): generate **x** uniformly from range [0,1] and calculate target data **t** with sin(2πx) function + noise with Gaussian distribution

Why?

Curve Fitting

Observed Data: Given n points (x,t)

Textbook Example): generate **x** uniformly from range [0,1] and calculate target data **t** with sin(2πx) function + noise with Gaussian distribution

Why?

Real data sets typically have underlying regularity that we are trying to learn.

Curve Fitting

Observed Data: Given n points (x,y)

Goal: use observed data to predict new target values t’ for new values of x’

Curve Fitting

Observed Data: Given n points (x,y)

Can we fit a polynomial function with this data?

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

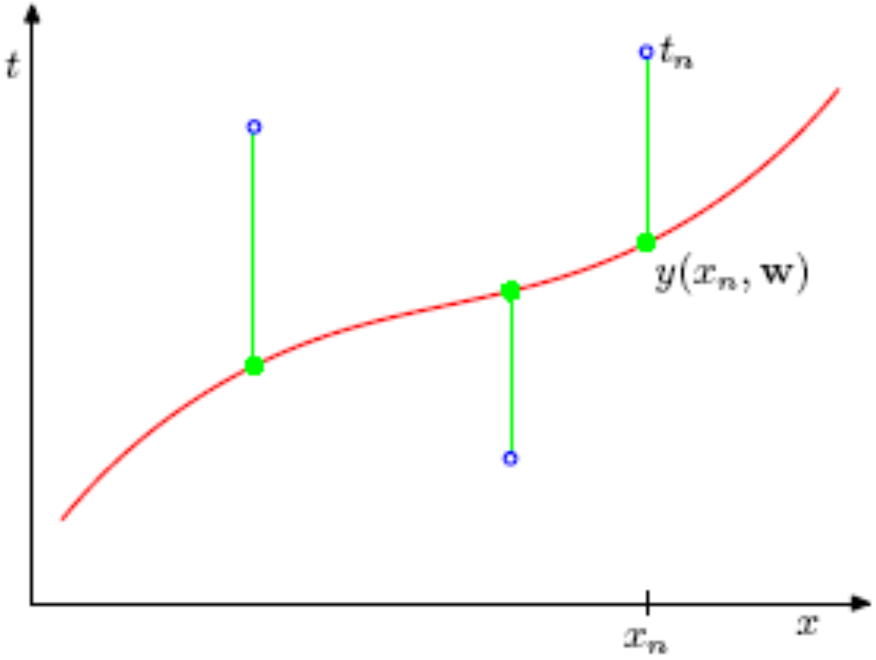
What values for **w** and M fits this data well?

How to measure goodness of fit?

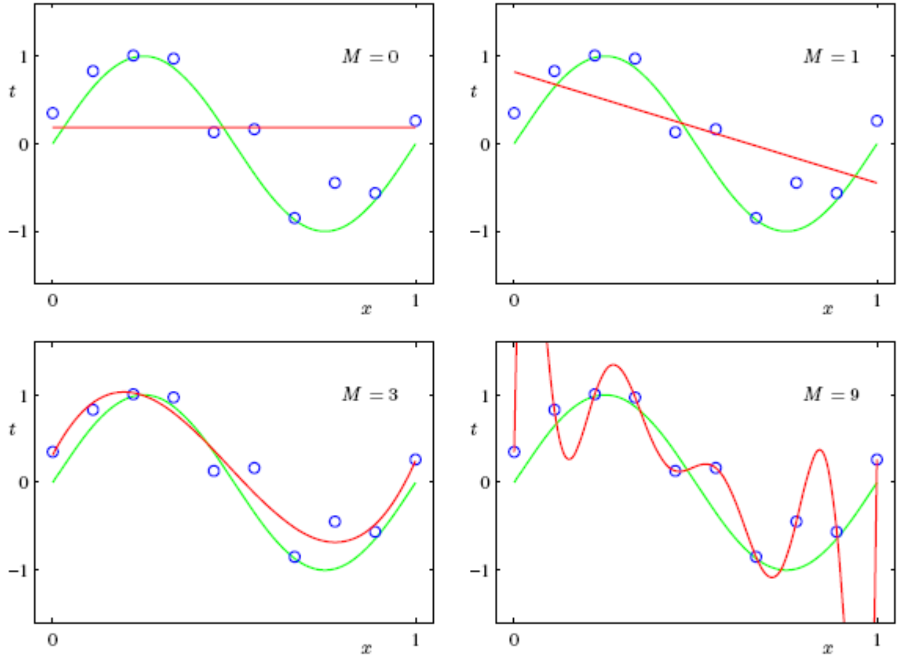
Minimize an error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Sum of squares
of error



Overfitting



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Combatting Overfitting

1. Increase data points

Combatting Overfitting

1. Increase data points

    Data observations may have just been noisy

    With more data, can see if data variation is due to noise or if is part of underlying relationship between observations

Combatting Overfitting

1. Increase data points

    Data observations may have just been noisy

    With more data, can see if data variation is due to noise or if is part of underlying relationship between observations

2. Regularization

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

    Introduce penalty term

    Trade off between good fit and penalty

    Hyperparameter, λ, is input to model. Hyperparam will reduce overfitting, in turn reducing variance and increasing bias (difference between estimated and true target)

How to check for overfitting?

Training and validation subset

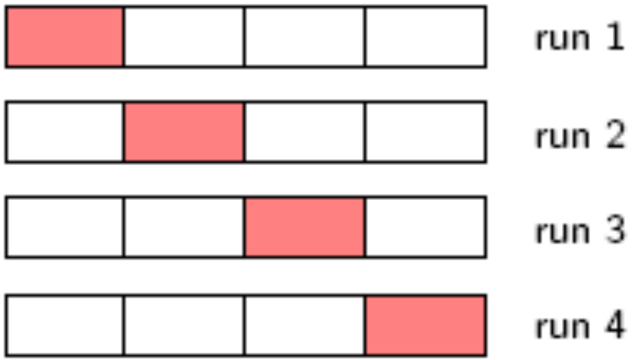
heuristic: data points > (multiple)(# parameters)

Training vs Testing

don’t touch test set until you are actually evaluating experiment!!

Cross-validation

1. Use portion, (S-1)/S, for training (white)
2. Assess performance (red)
3. Repeat for each run
4. Average performance scores



4-fold cross-validation (S=4)

Cross-validation

When to use?

Cross-validation

When to use?

Validation set is small. If very little data, use S=number of observed data points

Cross-validation

When to use?

Validation set is small. If very little data, use S=number of observed data points

Limitations:

- Computationally expensive - # of training runs increases by factor of S
- Might have multiple complexity parameters - may lead to training runs that is exponential in # of parameters

Alternative Approach:

Want an approach that depends only on size of training data rather than # of training runs

e.g.) Akaike information criterion (AIC), Bayesian information criterion (BIC, Sec 4.4)

Akaike Information Criterion (AIC)

Choose model with largest value for

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M$$

best-fit log likelihood

# of adjustable model parameters

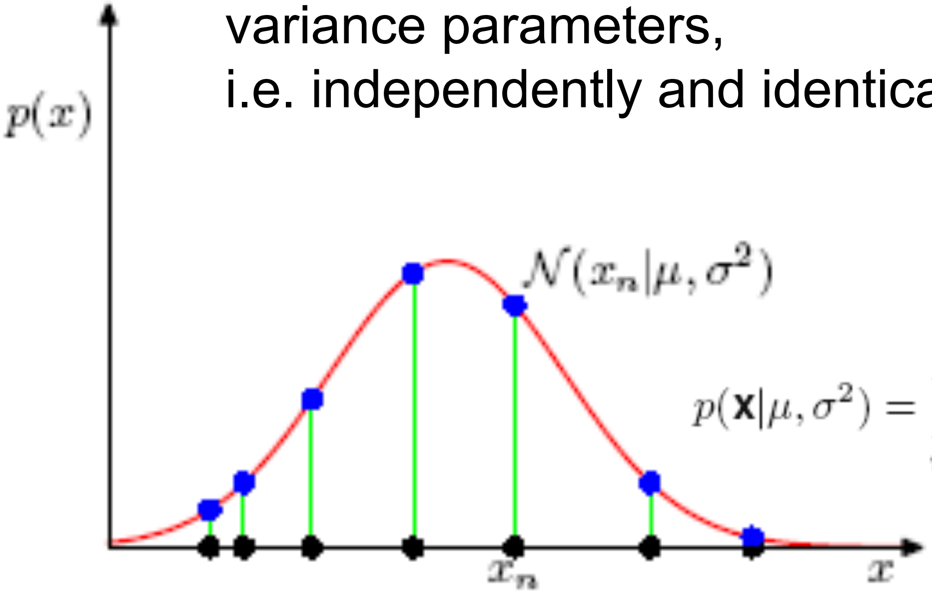
Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

This satisfies the two properties for a probability density! (what are they?)

Likelihood Function for Gaussian

Assumption: data points x drawn *independently* from same Gaussian distribution defined by unknown mean and variance parameters,
i.e. independently and identically distributed (i.i.d)



$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Curve Fitting (ver. 1)

Assumption:

1. Given value of x, corresponding target value t has a Gaussian distribution with mean y(x,**w**)

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

2. Data {**x**,**t**} drawn independently from distribution:

likelihood = $$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

Maximum Likelihood Estimation (MLE)

Log likelihood: $\ln p(\mathbf{t}|\mathbf{x}, \mathrm{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathrm{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$

What does maximizing the log likelihood look similar to?

Maximum Likelihood Estimation (MLE)

Log likelihood: $\ln p(\mathbf{t}|\mathbf{x}, \mathrm{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$

What does maximizing the log likelihood look similar to?

wrt **w**: $$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Maximum Posterior (MAP)

Simpler example: use Gaussian of form

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

With Bayes' can calculate posterior:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Maximum Posterior (MAP)

Determine **w** by maximizing posterior distribution

Equivalent to taking negative log of:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

and combining the Gaussian & log likelihood function from earlier...

Maximum Posterior (MAP)

Minimum of $$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

What does this look like?

Maximum Posterior (MAP)

Minimum of $\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$

What does this look like?

$$\widetilde{E}(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\lambda}{2}\|\mathbf{w}\|^2$$

What is regularization parameter?

Maximum Posterior (MAP)

Minimum of $$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

What does this look like?

$$\widetilde{E}(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\lambda}{2}\|\mathbf{w}\|^2$$

What is regularization parameter? $\lambda=\alpha/\beta$

Bayesian Curve Fitting

However…

MLE and MAP are not fully Bayesian because they involve using point estimates for $\mathbf{w}$

Curve Fitting (ver. 2)

Given training data {**x**, **t**} and new point x, predict the target value t

Assume parameters α and β are fixed

Evaluate predictive distribution: $p(t|x, \mathbf{x}, \mathbf{t})$

Bayesian Curve Fitting

Fully Bayesian approach requires integrating over all values of **w**, by applying sum and product rules of probability

$$p(t|x, \mathsf{x}, \mathsf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathsf{x}, \mathsf{t})\, \mathrm{d}\mathbf{w}$$

Bayesian Curve Fitting

marginalization

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w}$$

$$\mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$

posterior

This posterior is Gaussian and can be evaluated analytically (Sec 3.3)

Bayesian Curve Fitting

Predictive is Gaussian of form

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

with mean and variance and matrix

$$\begin{aligned} m(x) &= \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n \\ s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x). \end{aligned}$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x)^{\mathrm{T}}$$

Bayesian Curve Fitting

Need to define $\phi_i(x) = x^i \text{ for } i = 0, \ldots, M$

Mean and variance depend on x as a result of marginalization

$$\begin{aligned} m(x) &= \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \\ s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x). \end{aligned}$$

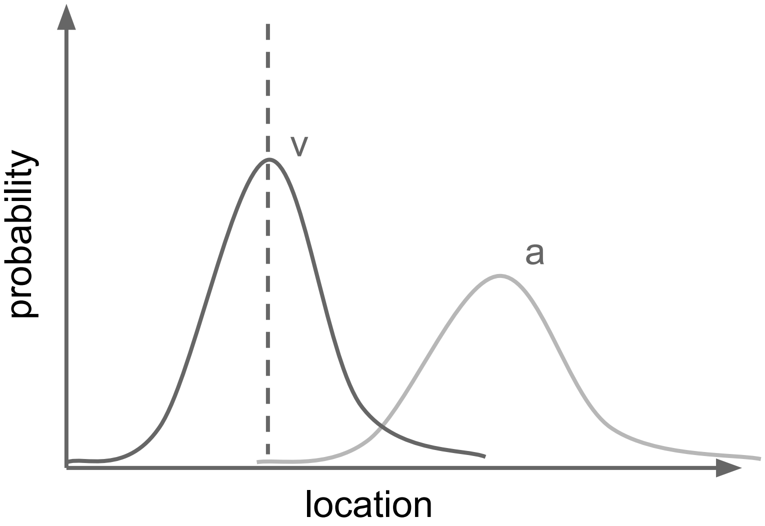
(not the case in MLE/MAP $\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$)

Application

MULTIMODAL SENSORY INTEGRATION

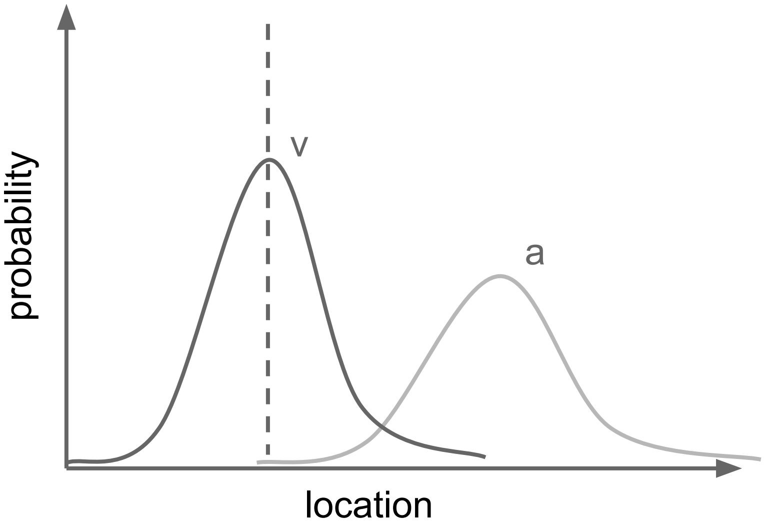
Two Models

Visual Capture

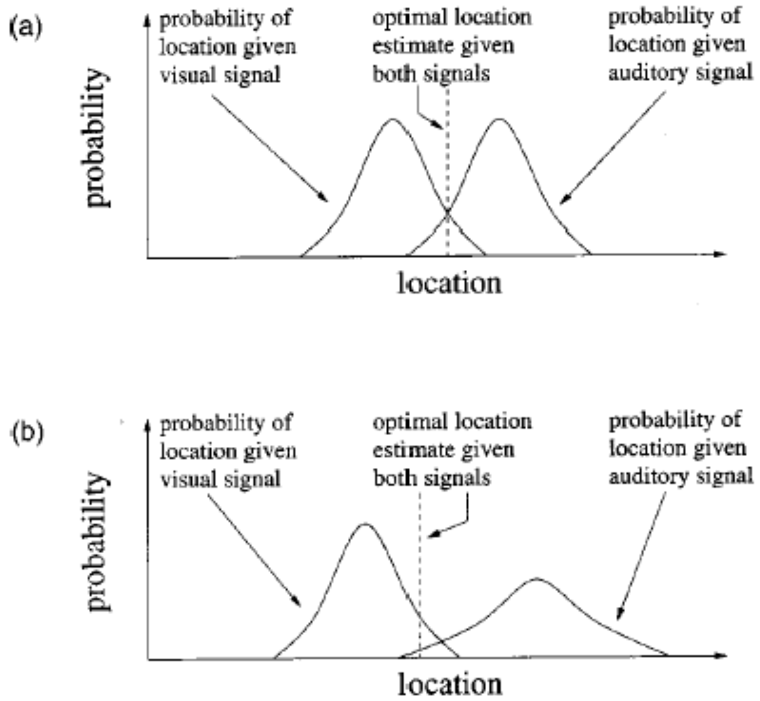


Two Models

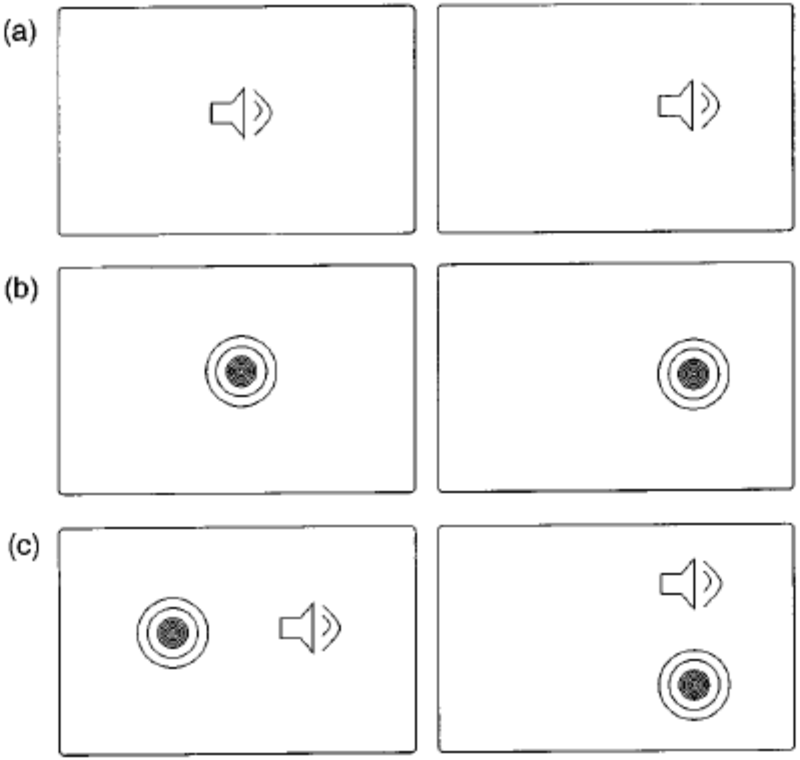
Visual Capture

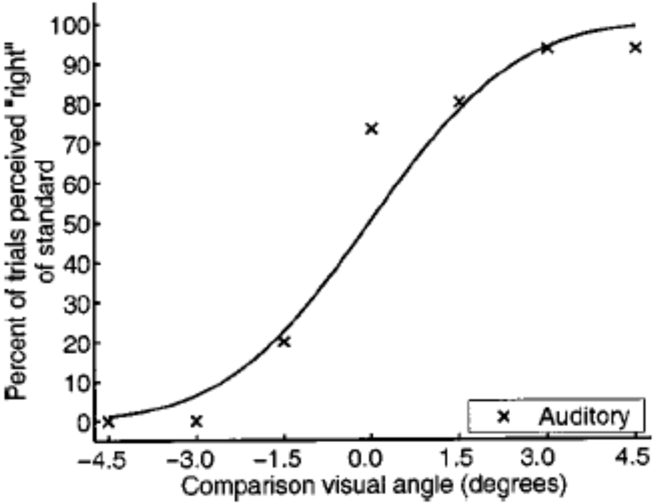


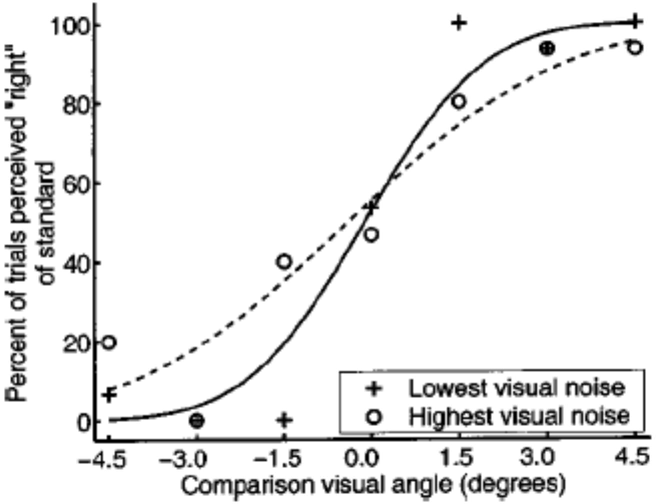
MLE



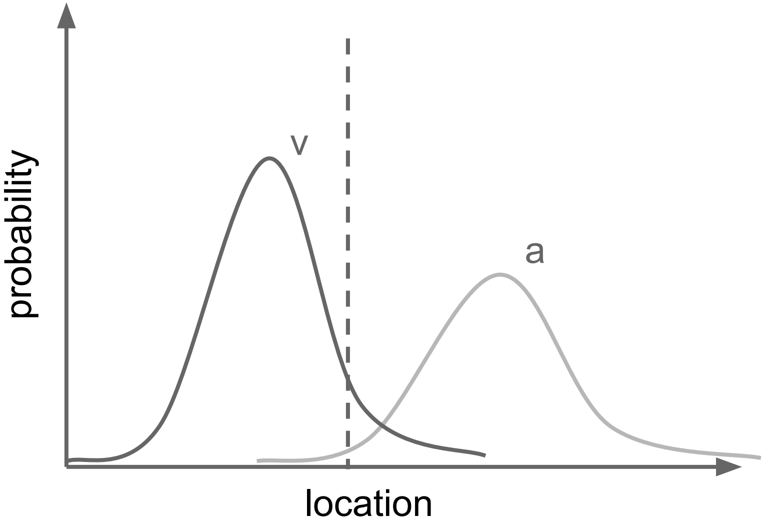
Procedure

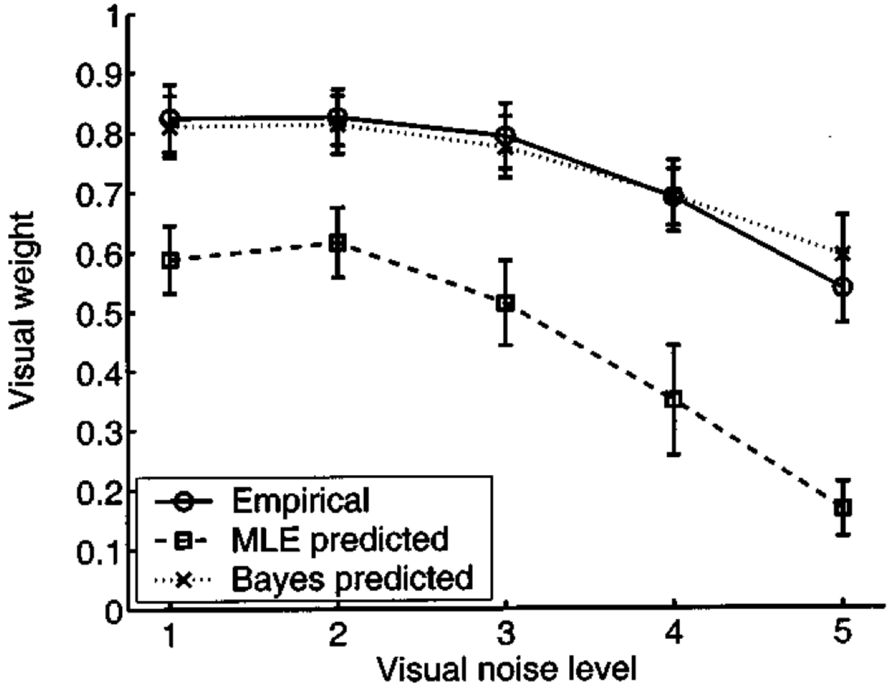

(a)
(b)
(c)


Percent of trials perceived "right" of standard
Comparison visual angle (degrees)
Auditory
100
90
80
70
60
50
40
30
20
10
0
−4.5
−3.0
−1.5
0.0
1.5
3.0
4.5


Percent of trials perceived "right" of standard
Comparison visual angle (degrees)
Lowest visual noise
Highest visual noise
100
80
60
40
20
0
−4.5
−3.0
−1.5
0.0
1.5
3.0
4.5

Final Result





$$w_v = \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_a^2} \text{ and } w_a = \frac{1/\sigma_a^2}{1/\sigma_v^2 + 1/\sigma_a^2}$$

The Math (MLE Model)

Likelihood

$$p(\mathcal{R}|\mu,\ \sigma^2) = \prod_{t=1}^{T} p_t^{r_t}(1 - p_t)^{1-r_t}$$

$$p_t = p(r_t|\mu, \sigma^2)$$

The Math (MLE Model)

Likelihood

$$p(\mathcal{R}|\mu,\, \sigma^2) = \prod_{t=1}^{T} p_t^{r_t}(1 - p_t)^{1-r_t}$$

$$p_t = p(r_t|\mu, \sigma^2)$$

$$w_v = \frac{1/\sigma_v^2}{1/\sigma_v^2 + 1/\sigma_a^2} \text{ and } w_a = \frac{1/\sigma_a^2}{1/\sigma_v^2 + 1/\sigma_a^2}$$

The Math ("Bayesian" Model)

Likelihood * Prior $\quad p(\mathcal{R}|\mu,\ \sigma^2)\, p(\mu, \sigma^2)$

Uniform

Inverse Gamma

The Math (Empirical)

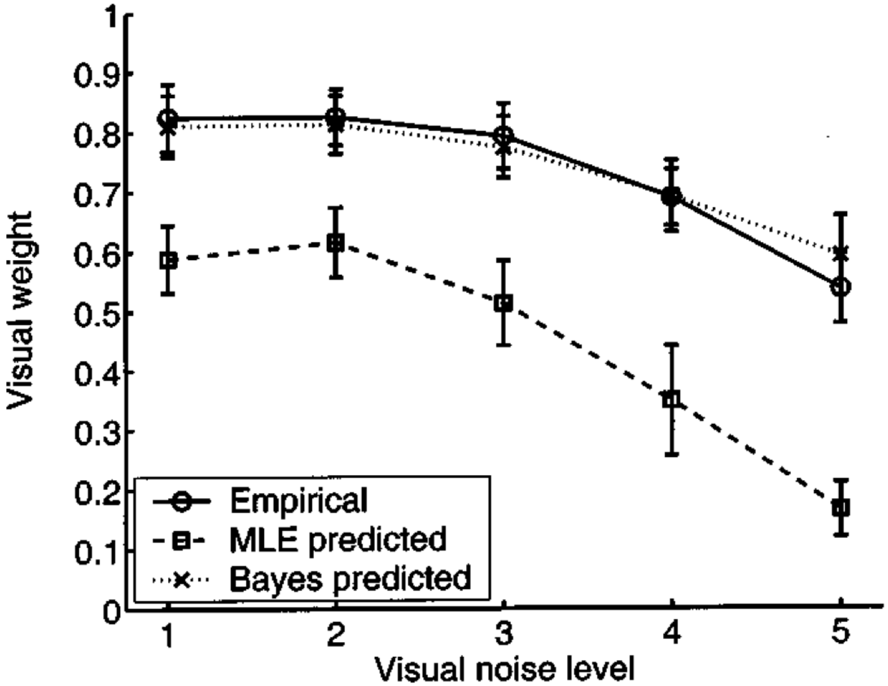
likelihood $$p(\mathcal{R}|\mu,\ \sigma^2) = \prod_{t=1}^{T} p_t^{r_t}(1 - p_t)^{1-r_t}$$

logistic function $$p_t = p(r_t = 1|w_v,\ w_a) = \frac{1}{1 + \exp[-(L_c - L_s)/\tau]}$$

location estimates $$L^c = w_v L_v^c + w_a L_a^c$$
$$L^s = w_v L_v^s + w_a L_a^s$$

weight constraint $$w_v + w_a = 1$$

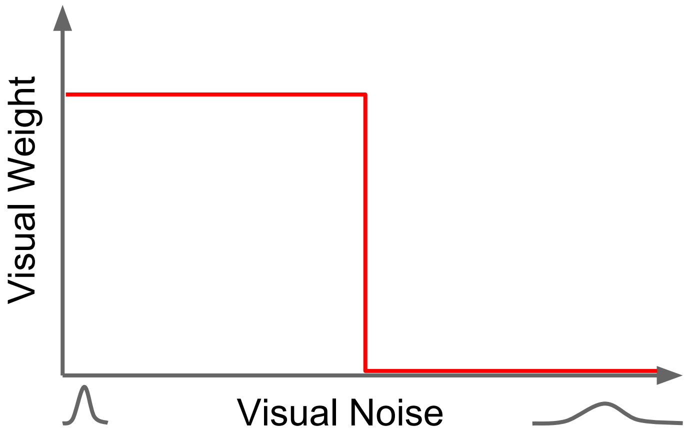
Final Result



QUESTIONS?

Two Models (Prediction)

Visual Capture



MLE