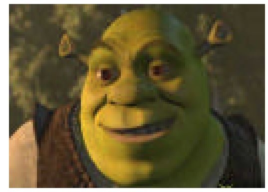
Introduction to Bayesian Learning

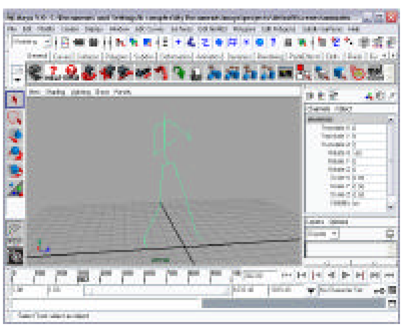
Aaron Hertzmann
University of Toronto
SIGGRAPH 2004 Tutorial

Evaluations: www.siggraph.org/courses_evaluation

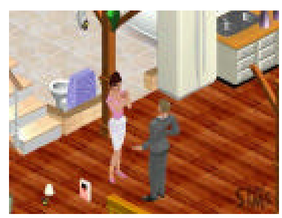
CG is maturing …

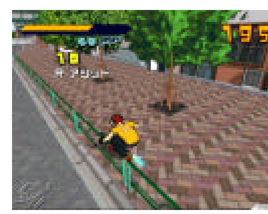


… but it's still hard to create



… it's hard to create in real-time





Data-driven computer graphics

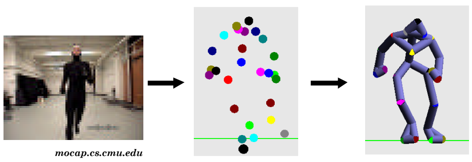
What if we can get models from the real world?

Data-driven computer graphics

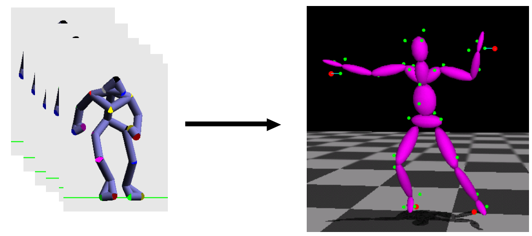
Three key problems:
- Capture data (from video, cameras, mocap, archives, …)
- Build a higher-level model
- Generate new data

Ideally, it should be automatic, flexible

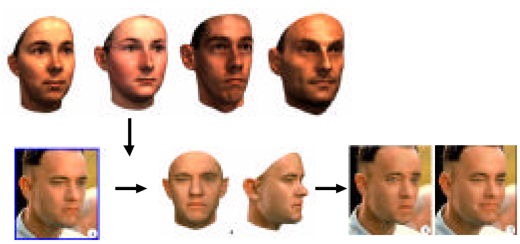
Example: Motion capture



Example: character posing

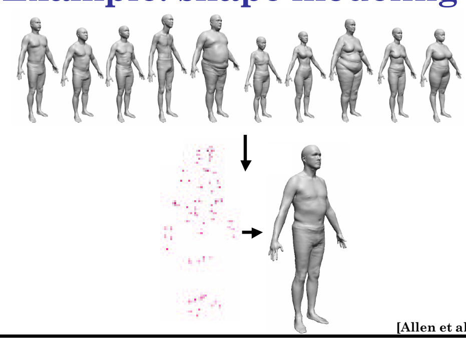


Example: shape modeling



[Blanz and Vetter 1999]

Example: shape modeling



[Allen et al. 2003]

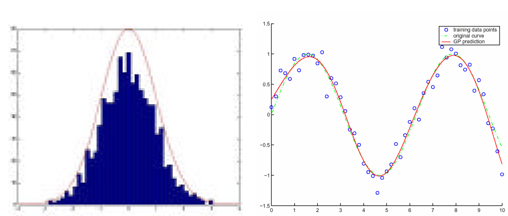
Key problems

- How do you fit a model to data?
  - How do you choose weights and thresholds?
  - How do you incorporate prior knowledge?
  - How do you merge multiple sources of information?
  - How do you model uncertainty?

Bayesian reasoning provides solutions

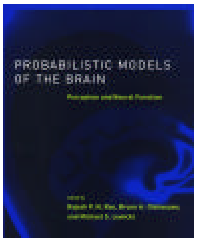
Bayesian reasoning is ...

Probability, statistics, data-fitting



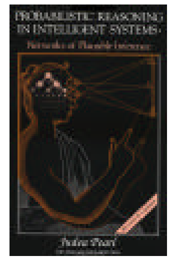
Bayesian reasoning is ...

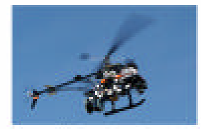
A theory of mind



Bayesian reasoning is ...

A theory of artificial intelligence

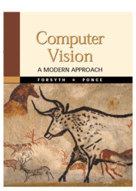


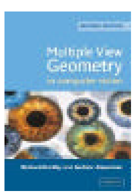


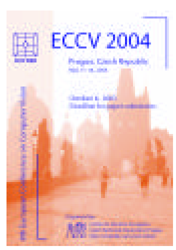
[Thrun et al.]

Bayesian reasoning is ...

A standard tool of computer vision







and ...

Applications in:
- Data mining
- Robotics
- Signal processing
- Bioinformatics
- Text analysis (inc. spam filters)
- and (increasingly) graphics!

Outline for this course

3:45-4pm: Introduction
4pm-4:45: Fundamentals
- From axioms to probability theory
- Prediction and parameter estimation

4:45-5:15: Statistical shape models
- Gaussian models and PCA
- Applications: facial modeling, mocap

5:15-5:30: Summary and questions

More about the course

- Prerequisites
  - Linear algebra, multivariate calculus, graphics, optimization
- Unique features
  - Start from first principles
  - Emphasis on graphics problems
  - Bayesian prediction
  - Take-home "principles"

Bayesian vs. Frequentist

- Frequentist statistics
  - a.k.a. "orthodox statistics"
  - Probability = frequency of occurrences in infinite # of trials
  - Arose from sciences with populations
  - *p*-values, *t*-tests, ANOVA, etc.
- Bayesian vs. frequentist debates have been long and acrimonious

Bayesian vs. Frequentist

"In academia, the Bayesian revolution is on the verge of becoming the majority viewpoint, which would have been unthinkable 10 years ago."

- Bradley P. Carlin, professor of public health, University of Minnesota
New York Times, Jan 20, 2004

Bayesian vs. Frequentist

If necessary, please leave these assumptions behind (for today):

- "A probability is a frequency"
- "Probability theory only applies to large populations"
- "Probability theory is arcane and boring"

Fundamentals

What is reasoning?

- How do we infer properties of the world?
- How should computers do it?

Aristotelian logic

- If A is true, then B is true
- A is true
- Therefore, B is true

A: My car was stolen
B: My car isn't where I left it

Real-world is uncertain

Problems with pure logic:
- Don't have perfect information
- Don't really know the model
- Model is non-deterministic

So let's build a logic of uncertainty!

Beliefs

Let B(A) = "belief A is true"
B(¬A) = "belief A is false"

e.g., A = "my car was stolen"
B(A) = "belief my car was stolen"

Reasoning with beliefs

Cox Axioms [Cox 1946]

1. Ordering exists
   - e.g., B(A) > B(B) > B(C)
2. Negation function exists
   - B(¬A) = f(B(A))
3. Product function exists
   - B(A Ù Y) = g(B(A|Y),B(Y))

This is all we need!

The Cox Axioms uniquely define a complete system of reasoning: This is probability theory!

Principle #1:

"Probability theory is nothing more than common sense reduced to calculation."

- Pierre-Simon Laplace, 1814

Definitions

P(A) = "probability A is true"
= B(A) = "belief A is true"
P(A) ∈ [0…1]
P(A) = 1 iff "A is true"
P(A) = 0 iff "A is false"

P(A|B) = "prob. of A if we knew B"
P(A, B) = "prob. A and B"

Examples

A: "my car was stolen"
B: "I can't find my car"

$P(A) = .1$
$P(A) = .5$
$P(B \mid A) = .99$
$P(A \mid B) = .3$

Basic rules

<u>Sum rule:</u>

$$P(A) + P(\neg A) = 1$$

<u>Example:</u>
A: "it will rain today"

$p(A) = .9$ — $p(\neg A) = .1$

Basic rules

<u>Sum rule:</u>

$$\text{å}_i \; P(A_i) = 1$$

when exactly one of A_i must be true

Basic rules

<u>Product rule:</u>

$$P(A,B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

Basic rules

Conditioning

<u>Product Rule</u>

$P(A,B) = P(A|B)\,P(B)$

$\rightarrow$ $P(A,B|C) = P(A|B,C)\,P(B|C)$

<u>Sum Rule</u>

$\text{å}_i\, P(A_i) = 1$ $\rightarrow$ $\text{å}_i\, P(A_i|B) = 1$

Summary

Product rule $P(A,B) = P(A|B)\,P(B)$
Sum rule $\text{å}_i\, P(A_i) = 1$

All derivable from Cox axioms;
must obey rules of common sense

Now we can derive new rules

Example

A = you eat a good meal tonight
B = you go to a highly-recommended restaurant
¬B = you go to an unknown restaurant

Model: P(B) = .7, P(A|B) = .8, P(A|¬B) = .5

What is P(A)?

Example, continued

Model: P(B) = .7, P(A|B) = .8, P(A|¬B) = .5

1 = P(B) + P(¬B) — Sum rule
1 = P(B|A) + P(¬B|A) — Conditioning
P(A) = P(B|A)P(A) + P(¬B|A)P(A)
= P(A,B) + P(A,¬B) — Product rule
= P(A|B)P(B) + P(A|¬B)P(¬B) — Product rule
= .8 .7 + .5 (1-.7) = .71

Basic rules

Marginalizing

$$P(A) = \sum_i P(A, B_i)$$

for mutually-exclusive B_i

e.g., p(A) = p(A,B) + p(A, ¬B)

Principle #2:

Given a complete model, we can derive any other probability

Inference

Model: P(B) = .7, P(A|B) = .8, P(A|¬B) = .5

If we know A, what is P(B|A)? ("Inference")

P(A,B) = P(A|B) P(B) = P(B|A) P(A)

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)} = .8\ .7\ /\ .71 \sim .79$$

Bayes' Rule

Inference

Bayes Rule

Likelihood

Prior

$$P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}$$

Posterior

Principle #3:

Describe your model of the world, and then compute the probabilities of the unknowns given the observations

---

Principle #3a:

Use Bayes' Rule to infer unknown model variables from observed data

Likelihood

Prior

$$P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}$$

Posterior

Discrete variables

Probabilities over discrete variables

$C \in \{\text{Heads}, \text{Tails}\}$

$P(C=\text{Heads}) = .5$

$P(C=\text{Heads}) + P(C=\text{Tails}) = 1$

Continuous variables

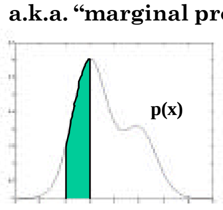
Let $\mathbf{x} \in \mathbb{R}^N$

How do we describe beliefs over **x**?

e.g., **x** is a face, joint angles, …

Continuous variables

Probability Distribution Function (PDF)
a.k.a. "marginal probability"



$P(a \le x \le b) = \int_a^b p(x)\,dx$

Notation: P(x) is prob
p(x) is PDF

Continuous variables

Probability Distribution Function (PDF)

Let $x \in \mathbb{R}$

p(x) can be any function s.t.

$\int_{-\infty}^{\infty} p(x)\,dx = 1$

$p(x) \ge 0$

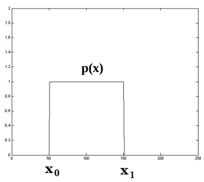
Define $P(a \le x \le b) = \int_a^b p(x)\,dx$

Uniform distribution

$x \sim \mathcal{U}(x_0, x_1)$

$p(x) = 1/(x_0 - x_1)$ if $x_0 \leq x \leq x_1$

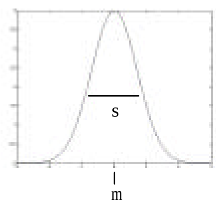
$= 0$ otherwise



Gaussian distributions

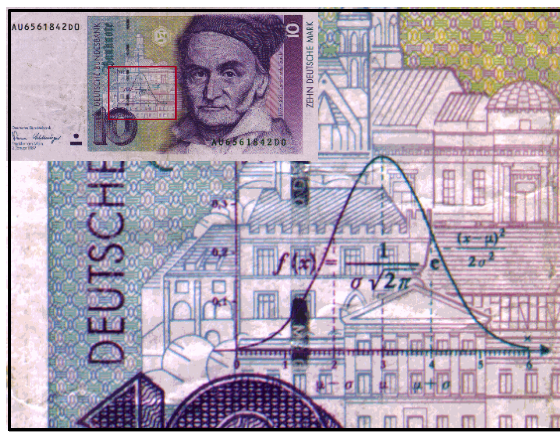
$x \sim \mathcal{N}(m, s^2)$

$p(x \mid m, s^2) = \exp(-(x-m)^2/2s^2) / \sqrt{2ps^2}$



Why use Gaussians?

- Convenient analytic properties
- Central Limit Theorem
- Works well
- Not for everything, but a good building block
- For more reasons, see
  [Bishop 1995, Jaynes 2003]



Rules for continuous PDFs

Same intuitions and rules apply

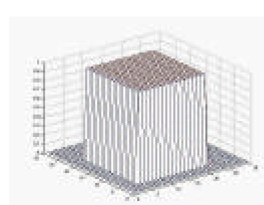
"Sum rule": $\int_{-\infty}^{\infty} p(x)\, dx = 1$

Product rule: $p(x,y) = p(x \mid y)p(x)$

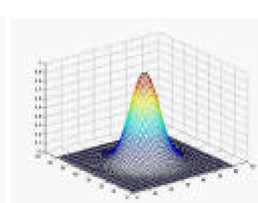
Marginalizing: $p(x) = \int p(x,y) dy$

… Bayes' Rule, conditioning, etc.

Multivariate distributions



Uniform: $x \sim \mathcal{U}(\text{dom})$



Gaussian: $x \sim \mathcal{N}(m, S)$

Inference

How do we reason about the world from observations?

Three important sets of variables:
- observations
- unknowns
- auxiliary ("nuisance") variables

Given the observations, what are the probabilities of the unknowns?

Inference

Example: coin-flipping

P(C = heads | q) = q

p(q) = $\mathcal{U}$(0,1)

Suppose we flip the coin 1000 times and get 750 heads. What is q?

Intuitive answer: 750/1000 = 75%

What is q?

$p(q) = \text{Uniform}(0,1)$

$P(C_i = h | q) = q, \; P(C_i = t | q) = 1\text{-}q$

$P(C_{1:N} \mid q) = \tilde{O}_i \, P(C_i = h \mid q)$

$p(q \mid C_{1:N}) = \dfrac{P(C_{1:N} \mid q) \, p(q)}{P(C_{1:N})}$ Bayes' Rule

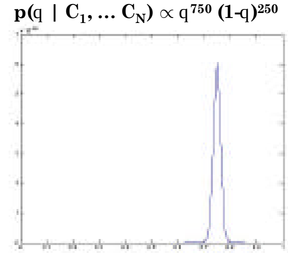
$= \tilde{O}_i \, P(C_i \mid q) \, P(q) / P(C_{1:N})$

$\propto q^H (1\text{-}q)^T$

H = 750, T = 250

What is q?

$p(q \mid C_1, \ldots C_N) \propto q^{750} (1\text{-}q)^{250}$



q

"Posterior distribution:" new beliefs about q

Bayesian prediction

What is the probability of another head?

$P(C{=}h | C_{1:N}) = \int P(C{=}h, q | C_{1:N}) \, dq$

$= \int P(C{=}h | q, C_{1:N}) \, P(q \mid C_{1:N}) \, dq$

$= (H+1)/(N+2)$

$= 751 / 1002 = 74.95 \%$

Note: we never computed q

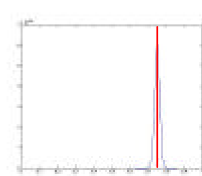
Parameter estimation

- What if we want an estimate of q?
- Maximum A Posteriori (MAP):

$\theta^* = \arg\max_q p(q \mid C_1, \ldots, C_N)$

$= H / N$

$= 750 / 1000 = 75\%$

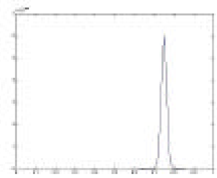


A problem

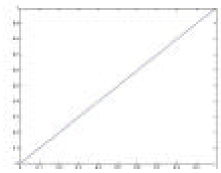
Suppose we flip the coin once
What is $P(C_2 = h \mid C_1 = h)$?

MAP estimate: $q^* = H/N = 1$
This is absurd!
Bayesian prediction:
$P(C_2 = h \mid C_1 = h) = (H+1)/(N+2) = 2/3$

What went wrong?



$p(q \mid C_{1:N})$



$p(q \mid C_1)$

Over-fitting

- A model that fits the data well but does not generalize
- Occurs when an estimate is obtained from a “spread-out” posterior
- Important to ask the right question: estimate C_{N+1}, not q

Principle #4:

Parameter estimation is not Bayesian. It leads to errors, such as over-fitting.

Advantages of estimation

Bayesian prediction is usually difficult and/or expensive

$$p(x \mid D) = \int p(x, q \mid D)\, dq$$

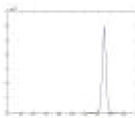
Q: When is estimation safe?

A: When the posterior is “peaked”

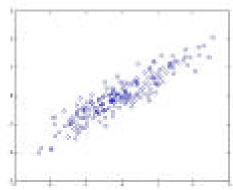
- The posterior “looks like” a spike
- Generally, this means a lot more data than parameters
- But this is not a guarantee (e.g., fit a line to 100 identical data points)
- Practical answer: use error bars (posterior variance)

Principle #4a:

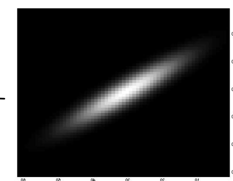
Parameter estimation is easier than prediction. It works well when the posterior is "peaked."



Learning a Gaussian



?



{x²}

m, s²

Learning a Gaussian

p(x|m,s²) = exp(-(x-m)²/2s²) / √2ps²

p($x_{1:K}$|m, s²) = Õ p(x_i | m, s²)

Want: max p($x_{1:K}$|m, s²)

= min –ln p($x_{1:K}$|m, s²)

= å$_i$ (x-m)²/2s² + K/2 ln 2p s²

Closed-form solution:

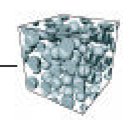
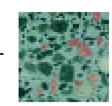
m = å$_i$ x_i / N

s² = å$_i$ (x - m)²/N

Stereology

[Jagnow et al. 2004 (this morning)]

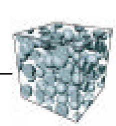
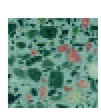
Model:

PDF over solids —  — 

p(q) p(S | q) p(I|S)

Problem: What is the PDF over solids?
Can't estimate individual solid shapes:
arg max p(q, S | I) is underconstrained)

Stereology

PDF over solids —  — 

p(q) p(S | q) p(I|S)

Marginalize out S:
p(q | I) = ∫ p(q, S | I) dS
can be maximized

Principle #4b:

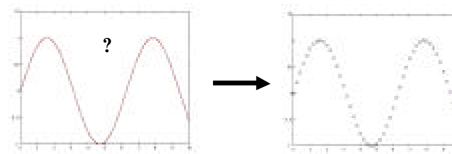
When estimating variables, marginalize out as many unknowns as possible.

Algorithms for this:
- Expectation-Maximization (EM)
- Variational learning

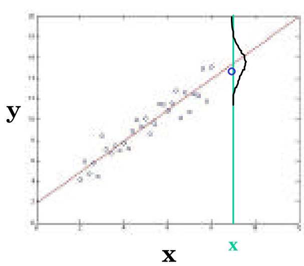
Regression

Regression

Curve fitting



Linear regression



Model:

$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

$y = a\,x + b + \epsilon$

Or:

$p(y|x,a,b,\sigma^2) = \mathcal{N}(ax + b, \sigma^2 I)$

Linear regression

$p(y\,|\,x, a, b,\sigma^2) = \mathcal{N}(ax + b, \sigma^2 I)$

$p(y_{1:K}\;|\;x_{1:K}, a, b, \sigma^2) = \prod_i p(y_i\;|\;x_i, a, b, \sigma^2)$

Maximum likelihood:

$a^*, b^*, \sigma^{2*} = \arg\max \prod_i p(y_i\,|\,x_i, a, b, \sigma^2)$

$= \arg\min -\ln \prod_i p(y_i\,|\,x, a, b, \sigma^2)$

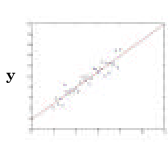
Minimize:

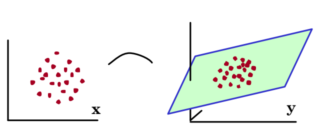
$\underbrace{\sum_i (y_i-(ax_i+b))^2}_{\text{Sum-of-squared differences: "Least-squares"}}/(2\sigma^2) + K/2 \ln 2\pi\sigma^2$

Linear regression

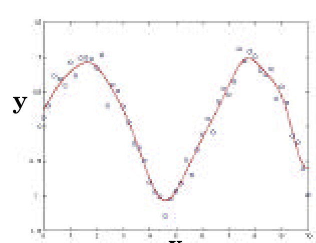
Same idea in higher dimensions

$y = Ax + \mu + \epsilon$





Nonlinear regression



Model:

$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

$y = f(x;w) + \epsilon$

Curve parameters

Or:

$p(y|x,w,\sigma^2) = \mathcal{N}(f(x;w), \sigma^2 I)$

Typical curve models

Line

$f(x;w) = w_0\ x + w_1$

B-spline, Radial Basis Functions

$f(x;w) =$ å$_i\ w_i\ B_i(x)$

Artificial neural network

$f(x;w) = $ å$_i\ w_i\ \tanh($å$_j\ w_j\ x + w_0)+w_1$

Nonlinear regression

p(y | x, w, s^2) = $\mathcal{N}$(f(x;w), s^{2}I)

p($y_{1:K}$ | $x_{1:K}$, w, s^2) = Õ$_i$ p(y_i | x_i, a, b, s^2)

Maximum likelihood:

w*,s^{2*} = arg max Õ$_i$ p(y_i | x_i,a,b,s^2)

= arg min –ln Õ$_i$ p(y_i | x, a, b, s^2)

Minimize:

å$_i$ (y_i-f(x_i;w))^2/(2s^2) + K/2 ln 2 p s^2

Sum-of-squared differences: "Least-squares"

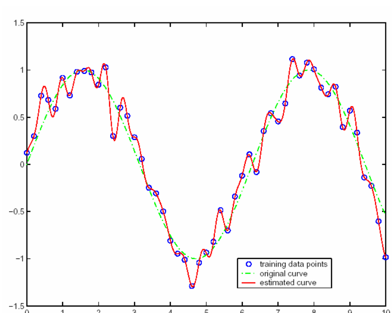
Principle #5:

Least-squares estimation is a special case of maximum likelihood.

Principle #5a:

Because it is maximum likelihood, least-squares suffers from overfitting.

Overfitting



Smoothness priors

Assumption: true curve is smooth

Bending energy:

p(w | l) ~ exp(-∫ ||∇ f||^2 / 2 l^2)

Weight decay:

p(w | l) ~ exp(-||w||^2 / 2 l^2)

Smoothness priors

MAP estimation:

arg max p(w | y) = p(y | w) p(w)/p(y)=

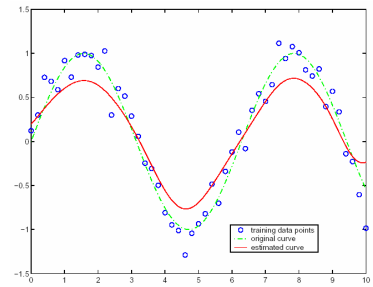
arg min –ln p(y | w) p(w) =

$\Sigma_i (y_i - f(x_i; w))^2/(2s^2) + \|w\|^2/2l^2 + K \ln s$

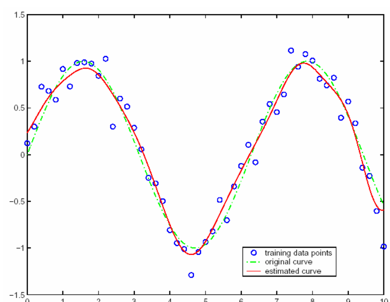
Sum-of-squares differences

Smoothness

Underfitting



Underfitting



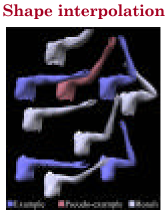
Principle #5b:

MAP estimation with smoothness priors leads to under-fitting.

Applications in graphics

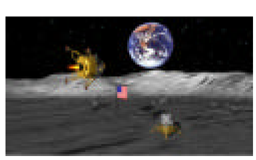
Two examples:

Shape interpolation



[Rose III et al. 2001]

Approximate physics



[Grzeszczuk et al. 1998]

Choices in fitting

- Smoothness, noise parameters
- Choice of basis functions
- Number of basis functions

Bayesian methods can make these choices automatically and effectively

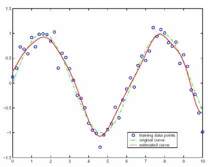
Learning smoothness

Given "good" data, solve

$l^*, \sigma^{2*} = \arg\max\ p(l, s^2 \mid w, x_{1:K}, y_{1:K})$

Closed-form solution

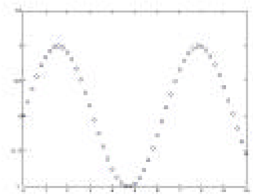
Shape reconstruction in vision [Szeliski 1989]



Learning without shape

Q: Can we learn smoothness/noise without knowing the curve?

A: Yes.



Learning without shape

$l^*, \sigma^{2*} = \arg\max\ p(l, s^2 \mid x_{1:K}, y_{1:K})$

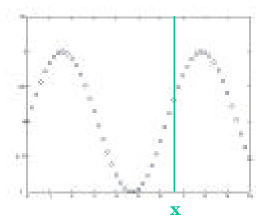
(2 unknowns, K measurements)

$$p(l, s^2 \mid x_{1:K}, y_{1:K}) = \int p(l, s^2, w \mid x_{1:K}, y_{1:K})\, dw$$
$$\propto \int p(x_{1:K}, y_{1:K} \mid w, s^2, l)\, p(w \mid l, s^2)\, dw$$

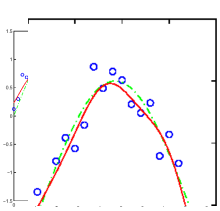
Bayesian regression

don't fit a single curve, but keep the uncertainty in the curve:

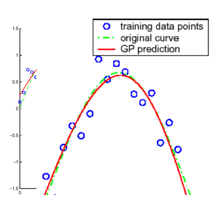
$p(x \mid x_{1:N}, y_{1:N})$



Bayesian regression

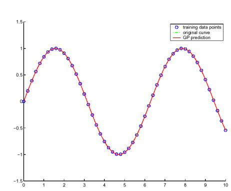


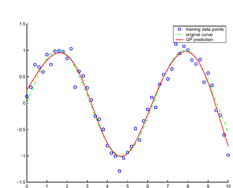
MAP/Least-squares (hand-tuned l, s^2, basis functions)



Gaussian Process regression (learned parameters l, s^2)

Bayesian regression

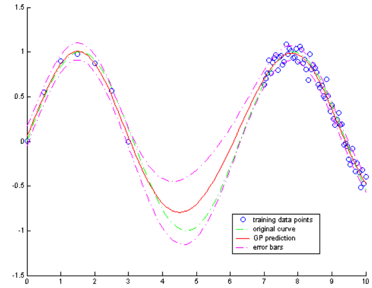




Principle #6:

Bayes' rule provide principle for learning (or marginalizing out) *all* parameters.

Prediction variances



More info: D. MacKay's *Introduction to Gaussian Processes*

NIPS 2003 Feature Selection Challenge

- Competition between classification algorithm, including SVMs, nearest neighbors, GPs, etc.
- Winners: R. Neal and J. Zhang
- Most powerful model they could compute with (1000's of parameters) and Bayesian prediction
- Very expensive computations

Summary of "Principles"

1. Probability theory is common sense reduced to calculation.
2. Given a model, we can derive any probability
3. Describe a model of the world, and then compute the probabilities of the unknowns with Bayes' Rule

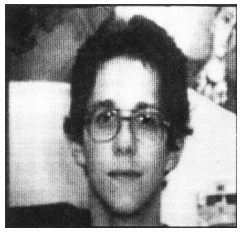
Summary of "Principles"

4. Parameter estimation leads to over-fitting when the posterior isn't "peaked." However, it is easier than Bayesian prediction.
5. Least-squares estimation is a special case of MAP, and can suffer from over- and under-fitting
6. You can learn (or marginalize out) all parameters.

Statistical shape and appearance models with PCA

Key vision problems

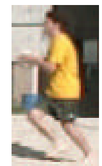
- Is there a face in this image?
- Who is it?
- What is the 3D shape and texture?



Turk and Pentland 1991

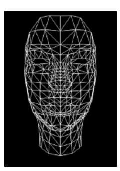
Key vision problems

- Is there a person in this picture?
- Who?
- What is their 3D pose?



Key graphics problems

- How can we easily create new bodies, shapes, and appearances?
- How can we edit images and videos?



The difficulty

- Ill-posed problems
  - Need prior assumptions
  - Lots of work for an artist

Outline

- Face modeling problem
  - Linear shape spaces
  - PCA
  - Probabilistic PCA
- Applications
  - face and body modeling

Background: 2D models

- Eigenfaces
  - Sirovich and Kirby 1987, Turk and Pentland 1991
- Active Appearance Models/Morphable models
  - Beier and Neely 1990
  - Cootes and Taylor 1998

Face representation

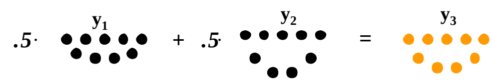
- 70,000 vertices with (x, y, z, r, g, b)
- Correspondence precomputed

[Blanz and Vetter 1999]

Data representation

$\mathbf{y}_i = [x_1, y_1, z_1, \ldots, x_{70,000}, y_{70,000}, z_{70,000}]^T$

Linear blends:



$\mathbf{y}_{new} = (\mathbf{y}_1 + \mathbf{y}_2) / 2$

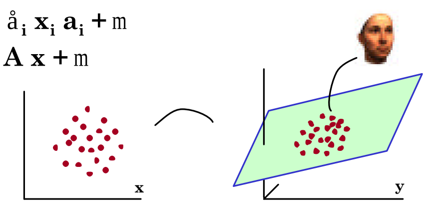
a.k.a. blendshapes, morphing

Linear subspace model

y = å$_i$ w$_i$ y$_i$ (s.t., å$_i$ w$_i$ = 1)

= å$_i$ x$_i$ a$_i$ + m

= A x + m

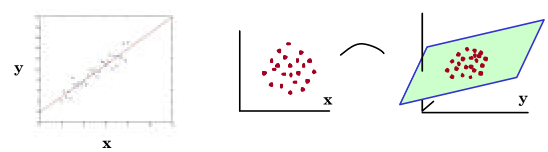


Problem: can we learn this linear space?

Principal Components Analysis (PCA)

Same model as linear regression

Unknown x



Conventional PCA (Bayesian formulation)

x, A, m ~ Uniform, $A^T A = I$

e ~ $\mathcal{N}$(0, s^2 I)

y = A x + m + e

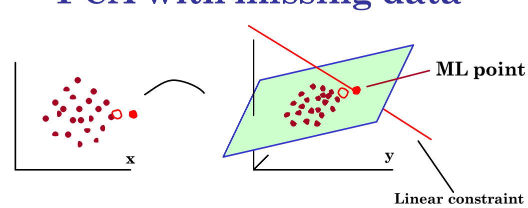
Given training y$_{1:K}$, what are A, x, m, s^2?

Maximum likelihood reduces to:

å$_i$ || y$_i$ – (A x$_i$ + m) ||^2 / 2s^2 + K/2 ln 2 p s^2

Closed-form solution exists

PCA with missing data



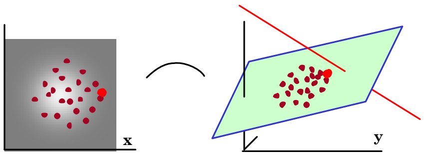
Problems:
- Estimated point far from data if data is noisy
- High-dimensional y is a uniform distribution
- Low-dimensional x is overconstrained

Why? Because x ~ $\mathcal{U}$

Probabilistic PCA

$\mathbf{x} \sim \mathcal{N}(0,I)$

$\mathbf{y} = A\mathbf{x} + \mathbf{b} + \epsilon$



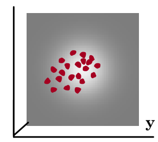
[Roweis 1998, Tipping and Bishop 1998]

Fitting a Gaussian

$\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$

easy to learn, and nice properties

… but Σ is a $70{,}000^2$ matrix



PPCA vs. Gaussians

However…

PPCA: $p(\mathbf{y}) = \int p(\mathbf{x},\mathbf{y})\, d\mathbf{x}$

$= \mathcal{N}(\mathbf{b},\ A\, A^T + \sigma^2 I)$

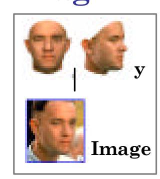
This is a special case of a Gaussian!

PCA is a degenerate case (σ^2=0)

Face estimation in an image

$p(\mathbf{y}) = \mathcal{N}(\mu, \Sigma)$

$p(\text{Image} \mid \mathbf{y}) = \mathcal{N}(I_s(\mathbf{y}), \sigma^2 I)$

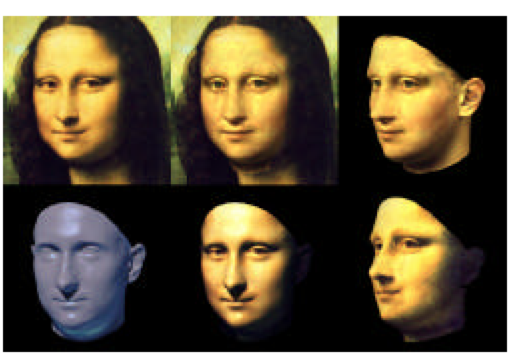


[Blanz and Vetter 1999]

$$-\ln p(S,T \mid \text{Image}) = \underbrace{\|\text{Image} - I_s(\mathbf{y})\|^2/2\sigma^2}_{\text{Image fitting term}} + \underbrace{(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)/2}_{\text{Face likelihood}}$$

Use PCA coordinates for efficiency

Efficient editing in PCA space



Comparison

PCA: unconstrained latent space – not good for missing data

Gaussians: general model, but impractical for large data

PPCA: constrained Gaussian – best of both worlds

Estimating a face from video

[Blanz et al. 2003]

The space of all body shapes

[Allen et al. 2003]

The space of all body shapes

[Allen et al. 2004]

Non-rigid 3D modeling from video

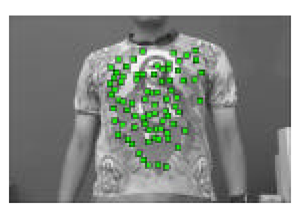
What if we don't have training data?

[Torresani and Hertzmann 2004]

Non-rigid 3D modeling from video

- Approach: learn all parameters
  - shape and motion
  - shape PDF
  - noise and outliers
- Lots of missing data (depths)
  - PPCA is essential
- Same basic framework, more unknowns

Results



Reference frame

Lucas-Kanade tracking

Tracking result

3D reconstruction

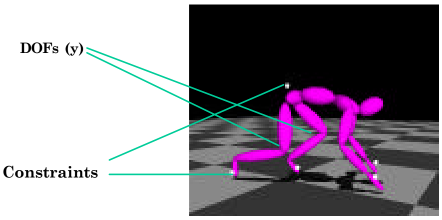
Results

Robust algorithm

3D reconstruction

[Almodovar 2002]

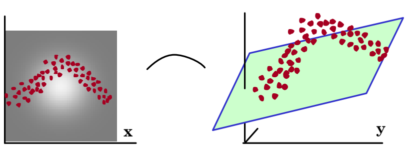
Inverse kinematics



[Grochow et al. 2004 (tomorrow)]

Problems with Gaussians/PCA

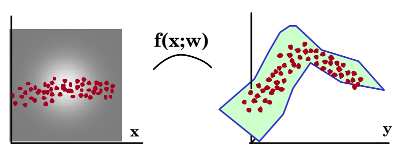
Space of poses may is nonlinear, non-Gaussian



Non-linear dimension reduction

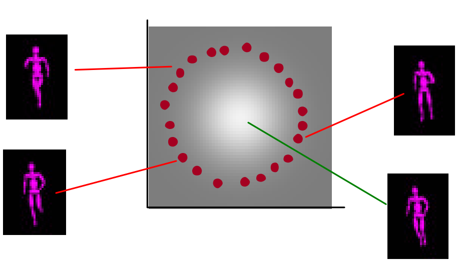
y = f(x;w) + e

Like non-linear regression w/o x



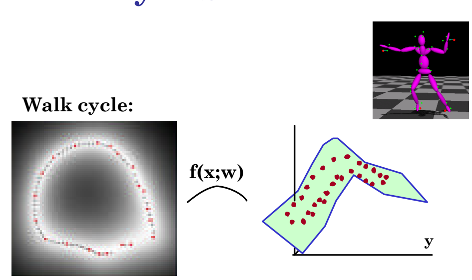
NLDR for BRDFs: [Matusik et al. 2003]

Problem with Gaussians/PPCA



Style-based IK

Walk cycle:



Details: [Grochow 2004 (tomorrow)]

Discussion and frontiers

Designing learning algorithms for graphics

Write a generative model
p(data | model)
Use Bayes' rule to learn the model from data
Generate new data from the model and constraints
(numerical methods may be required)

What model do we use?

- Intuition, experience, experimentation, rules-of-thumb
- Put as much domain knowledge in as possible
  - model 3D shapes rather than pixels
  - joint angles instead of 3D positions
- Gaussians for simple cases; nonlinear models for complex cases (active research area)

Q: Are there any limits to the power of Bayes' Rule?

http://yudkowsky.net/bayes/bayes.html:

A: According to legend, one who fully grasped Bayes' Rule would gain the ability to create and physically enter an alternate universe using only off-the-shelf equipment. One who fully grasps Bayes' Rule, yet remains in our universe to aid others, is known as a Bayesattva.

Problems with Bayesian methods

1. The best solution is usually intractable

- often requires expensive numerical computation
- it's still better to understand the real problem, and the approximations
- need to choose approximations carefully

Problems with Bayesian methods

2. Some complicated math to do

- Models are simple, algorithms complicated
- May still be worth it
- Bayesian toolboxes on the way (e.g., VIBES, Intel OpenPNL)

Problems with Bayesian methods

3. Complex models sometimes impede creativity

- Sometimes it's easier to tune
- Hack first, be principled later
- Probabilistic models give insight that helps with hacking

Benefits of the Bayesian approach

1. Principled modeling of noise and uncertainty
2. Unified model for learning and synthesis
3. Learn all parameters
4. Good results from simple models
5. Lots of good research and algorithms

Course notes, slides, links:
http://www.dgp.toronto.edu/~hertzman/ibl2004

Course evaluation
http://www.siggraph.org/courses_evaluation

Thank you!