Probabilistic Graphical Models

Introduction. Basic Probability and Bayes

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

26/9/2011

Outline

Benefits of Doubt

> Not to be absolutely certain is,
> I think, one of the essential things
> in rationality — B. Russell (1947)

Real-world problems are uncertain

- Measurement errors
- Incomplete, ambiguous data
- Model? Features?

Benefits of Doubt

> Not to be absolutely certain is,
> I think, one of the essential things
> in rationality — B. Russell (1947)

If uncertainty is part of your problem ...

Ignore/remove it

- Costly
- Complicated
- Not always possible

Benefits of Doubt

Not to be absolutely certain is, I think, one of the essential things in rationality B. Russell (1947)

If uncertainty is part of your problem . . .

Ignore/remove it

- Costly
- Complicated
- Not always possible

Live with it

- Quantify it: probabilities
- Compute it: Bayesian inference

Benefits of Doubt

> Not to be absolutely certain is,
> I think, one of the essential things
> in rationality
>
> B. Russell (1947)

If uncertainty is part of your problem ...

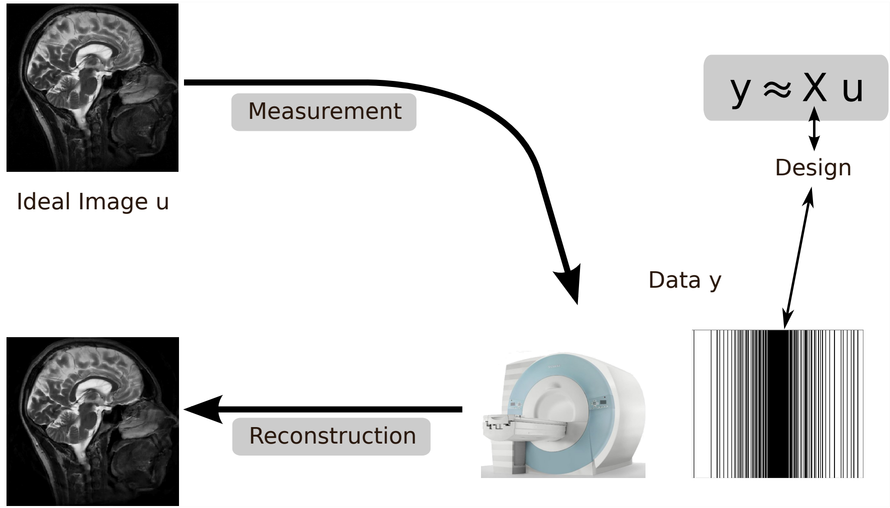
Ignore/remove it

- Costly
- Complicated
- Not always possible

Exploit it

- Experimental design
- Robust decision making
- Multimodal data integration

Image Reconstruction



Reconstruction is Ill Posed

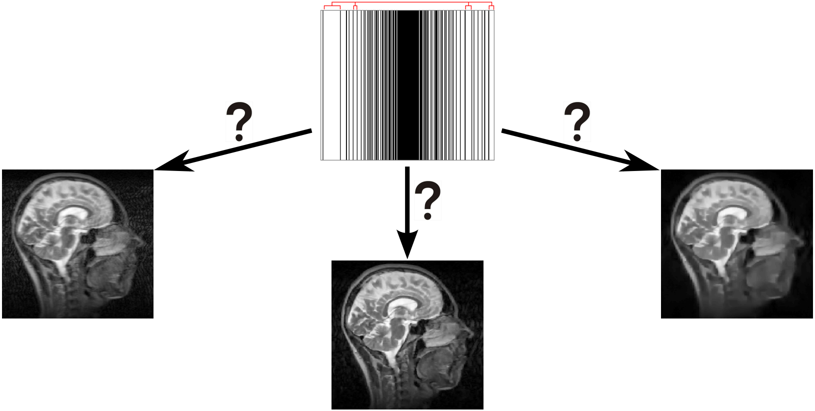
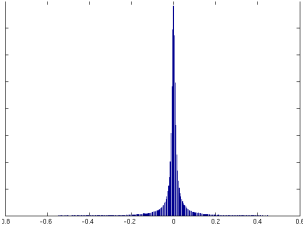


Image Statistics

Whatever images are . . .

they are not Gaussian!

- Image gradient super-Gaussian (“sparse”)



Use sparsity prior distribution $P(\boldsymbol{u})$

Posterior Distribution

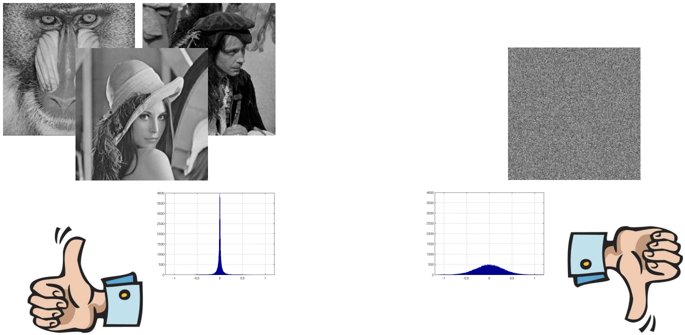
- Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$: Data fit

$$P(\boldsymbol{y}|\boldsymbol{u})$$

Posterior Distribution

- Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$: Data fit
- Prior $P(\boldsymbol{u})$: Signal properties

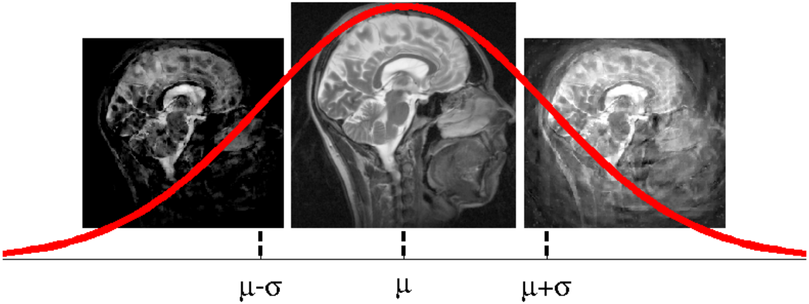
$$P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})$$



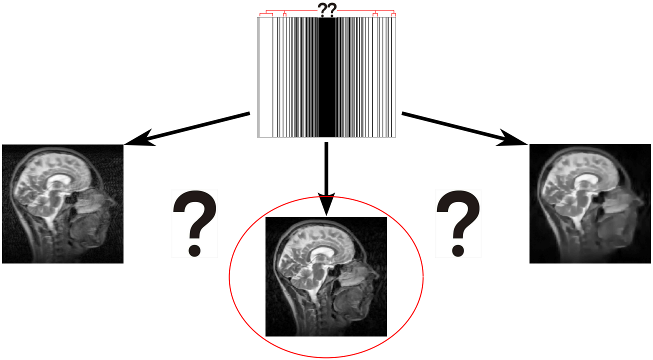
Posterior Distribution

- Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$: Data fit
- Prior $P(\boldsymbol{u})$: Signal properties
- Posterior distribution $P(\boldsymbol{u}|\boldsymbol{y})$: Consistent information summary

$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$



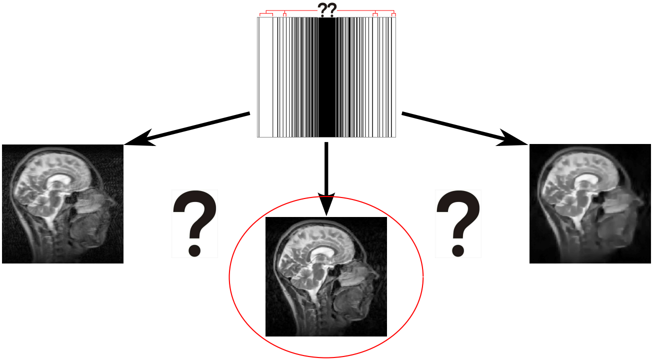
Estimation



Maximum a Posteriori (MAP) Estimation

$$\boldsymbol{u}_* = \operatorname{argmax}_{\boldsymbol{u}} P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})$$

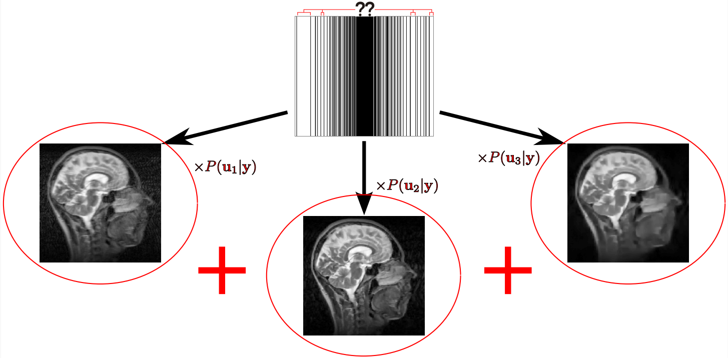
Estimation



Maximum a Posteriori (MAP) Estimation

- There are many solutions. Why settle for any single one?

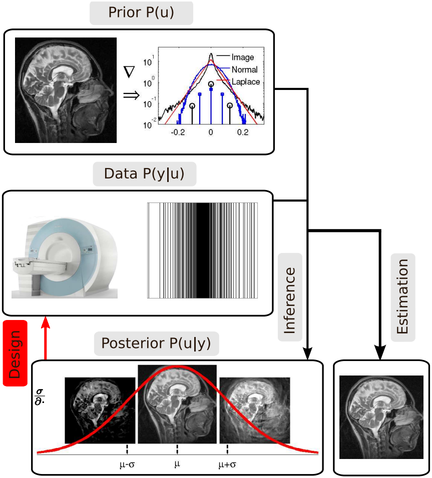
Bayesian Inference



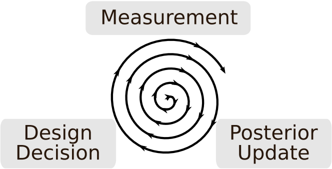
Use All Solutions

- Weight each solution by our uncertainty
- Average over them. Integrate, don't prune

Bayesian Experimental Design



- Posterior: Uncertainty in reconstruction
- Experimental design: Find poorly determined directions
- Sequential search with interjacent partial measurements



Structure of Course

Graphical Models [$\approx$ 6 weeks]

- Probabilistic database. Expert system
- Query for making optimal decisions
- Graph separation $\leftrightarrow$ conditional independence
  $\Rightarrow$ (More) efficient computation (dynamic programming)

Approximate Inference [$\approx$ 6 weeks]

- Bayesian inference is never really tractable
- Variational relaxations (convex duality)
- Propagation algorithms
- Sparse Bayesian models

Course Goals

This course is not:

- Exhaustive (but we will give pointers)
- Playing with data until it works
- Purely theoretical analysis of methods

This course is:

- Computer scientist's view on Bayesian machine learning: Layers above and below formulae
  - Understand concepts (what to do and why)
  - Understand approximations, relaxations, generic algorithmic schemata (how to do, above formulae)
  - Safe implementation on a computer (how to do, below formulae)
- Red line through models, algorithms. Exposing roots in specialized computational mathematics

Why Probability?

- Remember sleeping through Statistics 101 (p-value, t-test, ...)? Forget that impression!
- Probability leads to beautiful, useful insights and algorithms. Not much would work today without probabilistic algorithms, decisions from incomplete knowledge.
  - Numbers, functions, moving bodies $\Rightarrow$ Calculus
  - Predicates, true/false statements $\Rightarrow$ Predicate logic
  - Uncertain knowledge about numbers, predicates, ... $\Rightarrow$ Probability
- Machine learning? Have to speak probability! Crash course here. But dig further, it's worth it:
  - Grimmett, Stirzaker: Probability and Random Processes
  - Pearl: Probabilistic Reasoning in Intelligent Systems

Why Probability?

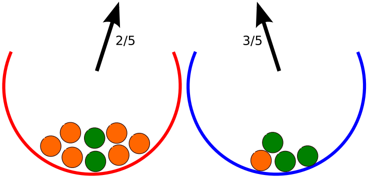
Reasons to use probability (forget “classical” straightjacket)

- We really don’t / cannot know (exactly)
- It would be too complicated/costly to find out
- It would take too long to compute
- Nondeterministic processes (given measurement resolution)
- Subjective beliefs, interpretations

Probability over Finite/Countable Sets

You know databases?
You know probability!

- Probability distribution P:
  Joint table/hypercube ($\cdot \geq 0$; $\sum \cdot = 1$)
- Random variable F: Index of table
- Event $\mathcal{E}$: Part of table
  Probability $P(\mathcal{E})$: Sum over cells in $\mathcal{E}$
- Marginal distribution $P(F)$: Projection of table (sum over others)

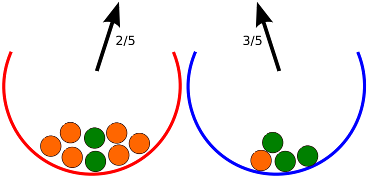


	Box	
Fruit	**red**	**blue**
apple	1/10	9/20
orange	3/10	3/20

Probability over Finite/Countable Sets

You know databases?
You know probability!

- Probability distribution P:
  Joint table/hypercube ($\cdot \geq 0; \sum \cdot = 1$)
- Random variable F: Index of table
- Event $\mathcal{E}$: Part of table
  Probability $P(\mathcal{E})$: Sum over cells in $\mathcal{E}$
- Marginal distribution $P(F)$: Projection of table (sum over others)



	Box	
Fruit	**red**	**blue**
apple	1/10	9/20
orange	3/10	3/20

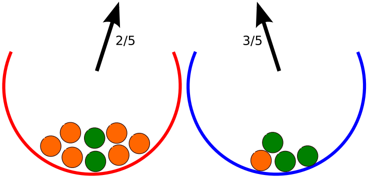
Marginalization (Sum Rule)

Not interested in variable(s) right now?
$\Rightarrow$ Marginalize over them!

$$P(F) = \sum_{B=r,b} P(F,B)$$

Probability over Finite/Countable Sets (II)

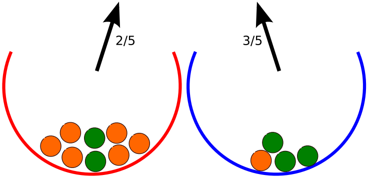
- Conditional probability:
  Factorization of table
  - Chop out part you're sure about (don't marginalize: you know!)
  - Renormalize to 1 (/ marginal)



	Box	
Fruit	**red**	**blue**
apple	1/10	9/20
orange	3/10	3/20

Probability over Finite/Countable Sets (II)

- Conditional probability: Factorization of table
  - Chop out part you're sure about (don't marginalize: you know!)
  - Renormalize to 1 (/ marginal)



Conditioning (Product Rule)

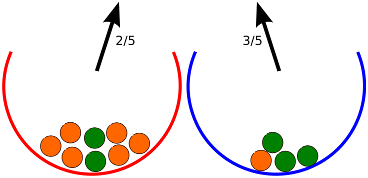
Observed some variable/event?
$\Rightarrow$ Condition on it!
Joint = Conditional $\times$ Marginal

Fruit	Box red	Box blue
apple	1/10	9/20
orange	3/10	3/20

$$P(F,B) = P(F|B)P(B)$$

Probability over Finite/Countable Sets (II)

- Conditional probability: Factorization of table
  - Chop out part you're sure about (don't marginalize: you know!)
  - Renormalize to 1 (/ marginal)



Conditioning (Product Rule)

Observed some variable/event?
$\Rightarrow$ Condition on it!
Joint = Conditional $\times$ Marginal

Fruit	Box red	Box blue
apple	1/10	9/20
orange	3/10	3/20

$$P(F,B) = P(F|B)P(B)$$

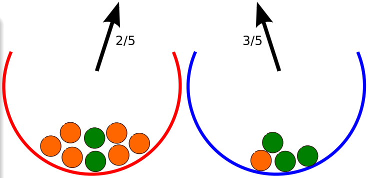
- Information propagation (B$\rightarrow$F)
  - Predict
  - Marginalize

$$P(F) = \sum_B P(F|B)P(B)$$

Probability over Finite/Countable Sets (III)

Bayes Formula

$$P(B|F)P(F) = P(F,B) = P(F|B)P(B)$$



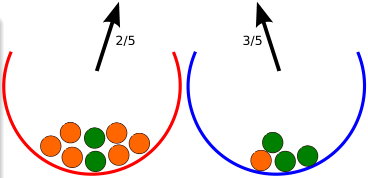
	Box	
Fruit	**red**	**blue**
apple	1/10	9/20
orange	3/10	3/20

Probability over Finite/Countable Sets (III)

Bayes Formula

$$P(B|F) = \frac{P(F|B)P(B)}{P(F)}$$

- Inversion of information flow
- Causal $\rightarrow$ diagnostic (diseases $\rightarrow$ symptoms)
- Inverse problem



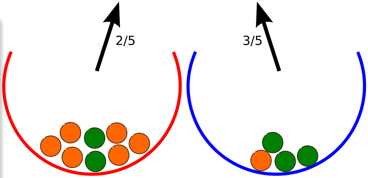
	Box	
Fruit	**red**	**blue**
apple	1/10	9/20
orange	3/10	3/20

Probability over Finite/Countable Sets (III)

Bayes Formula

$$P(B|F) = \frac{P(F|B)P(B)}{P(F)}$$

- Inversion of information flow
- Causal $\rightarrow$ diagnostic (diseases $\rightarrow$ symptoms)
- Inverse problem



	Box	
Fruit	**red**	**blue**
apple	1/10	9/20
orange	3/10	3/20

Chain rule of probability

$$P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \dots P(X_n|X_1, \dots, X_{n-1})$$

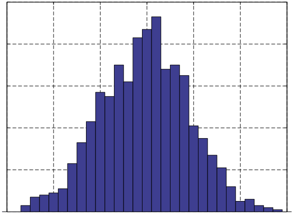
- Holds in any ordering
- Starting point for Bayesian networks [next lecture]

Probability over Continuous Variables

$$\Delta x \longrightarrow dx$$

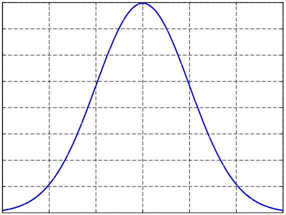
$$P(dx) = p(x)\, dx$$

Distribution



$\longrightarrow$

Density



$$\sum_x \ldots \longrightarrow \int \ldots\, dx$$

Probability over Continuous Variables (II)

- Caveat: Null sets [$P(\{x = 5\}) = 0$; $P(\{x \in \mathbb{N}\}) = 0$]
- Every observed event is a null set!
  $P(y|x) = P(y, x)/P(x)$ cannot work for $P(x) = 0$
- Define conditional density as $P(y|x)$ s.t.

$$P(y|x \in \mathcal{A}) = \int_{\mathcal{A}} P(y|x) P(x)\, dx \quad \text{for all events } \mathcal{A}$$

- Most cases in practice:
  - Look at $y \mapsto P(y, x)$ ("plug in x")
  - Recognize density / normalize

Probability over Continuous Variables (II)

- Caveat: Null sets [$P(\{x = 5\}) = 0$; $P(\{x \in \mathbb{N}\}) = 0$]
- Every observed event is a null set!
  $P(y|x) = P(y,x)/P(x)$ cannot work for $P(x) = 0$
- Define conditional density as $P(y|x)$ s.t.

$$P(y|x \in \mathcal{A}) = \int_{\mathcal{A}} P(y|x) P(x)\, dx \quad \text{for all events } \mathcal{A}$$

- Most cases in practice:
  - Look at $y \mapsto P(y,x)$ ("plug in x")
  - Recognize density / normalize
- Another (technical) caveat: Not all subsets can be events.
  $\Rightarrow$ Events: "Nice" subsets (measurable)

Expectation. Moments of a Distribution

Expectation

$$\mathrm{E}[f(\boldsymbol{x})] = \int f(\boldsymbol{x}) P(\boldsymbol{x})\, d\boldsymbol{x} \quad \text{or} \quad \sum_{\boldsymbol{x}} f(\boldsymbol{x}) P(\boldsymbol{x})$$

Expectation. Moments of a Distribution

Expectation

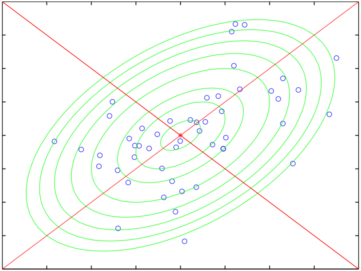
$$\mathrm{E}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x} \quad \text{or} \quad \sum_{\boldsymbol{x}} f(\boldsymbol{x})P(\boldsymbol{x})$$

- $P(\boldsymbol{x})$ complicated. How does $\boldsymbol{x} \sim P(\boldsymbol{x})$ behave?
- Moments: Essential behaviour of distribution
- Mean (1st order)

$$\mathrm{E}[\boldsymbol{x}] = \int \boldsymbol{x} P(\boldsymbol{x})\,d\boldsymbol{x}$$

- Covariance (2nd order)

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathrm{E}[\boldsymbol{x}\boldsymbol{y}^T] - \mathrm{E}[\boldsymbol{x}](\mathrm{E}[\boldsymbol{y}])^T$$
$$= \mathrm{E}[\boldsymbol{v}_x \boldsymbol{v}_y^T],\ \ \boldsymbol{v}_x = \boldsymbol{x} - \mathrm{E}[\boldsymbol{x}]$$



Expectation. Moments of a Distribution

Expectation

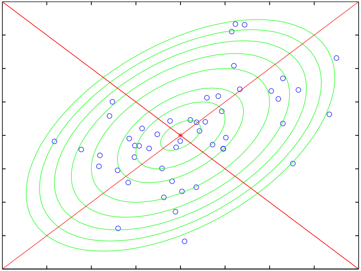
$$\mathrm{E}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})P(\boldsymbol{x})\,d\boldsymbol{x} \quad \text{or} \quad \sum_{\boldsymbol{x}} f(\boldsymbol{x})P(\boldsymbol{x})$$

- $P(\boldsymbol{x})$ complicated. How does $\boldsymbol{x} \sim P(\boldsymbol{x})$ behave?
- Moments: Essential behaviour of distribution
- Mean (1st order)

$$\mathrm{E}[\boldsymbol{x}] = \int \boldsymbol{x} P(\boldsymbol{x})\,d\boldsymbol{x}$$

- Covariance (2nd order)

$$\mathrm{Cov}[\boldsymbol{x}] = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}]$$



Decision Theory in 30 Seconds

Recipe for optimal decisions

1. Choose a loss function L
   (depends on problem, your valuation, susceptibility)
2. Model actions (A), outcomes (O). Compute $P(O|A)$
3. Compute risks (expected losses) $R(A) = \int L(O)P(O|A)\,dO$
4. Go for $A_* = \operatorname{argmin}_A R(A)$

- Special case: Pricing of A (option, bet, car)
  Choose $-R(A) + \text{Margin}$
- Next best measurement? Next best scientific experiment?
- Harder if timing plays a role (optimal control, *etc*)

Maximum Likelihood Estimation

Bayesian inversion hard. In simple cases, with enough data:

Maximum Likelihood Estimation

Observed $\{x_i\}$. Interested in cause θ

- Construct sampling model $P(x|\theta)$
- Likelihood $L(\theta) = P(D|\theta) = \prod_i P(x_i|\theta)$:
  Should be high close to "true" θ_0
- Maximum likelihood estimator:
  $\theta_* = \text{argmax}\, L(\theta) = \text{argmax} \log L(\theta) = \text{argmax} \sum_i \log P(x_i|\theta)$

Maximum Likelihood Estimation

Bayesian inversion hard. In simple cases, with enough data:

Maximum Likelihood Estimation

Observed $\{x_i\}$. Interested in cause θ

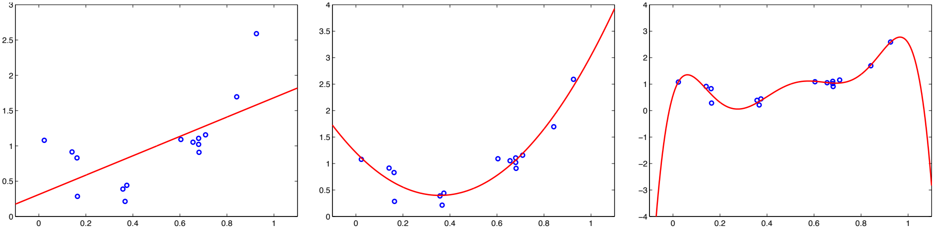
- Construct sampling model $P(x|\theta)$
- Likelihood $L(\theta) = P(D|\theta) = \prod_i P(x_i|\theta)$:
  Should be high close to "true" θ_0
- Maximum likelihood estimator:
  $\theta_* = \text{argmax}\, L(\theta) = \text{argmax} \log L(\theta) = \text{argmax} \sum_i \log P(x_i|\theta)$

- Method of choice for simple θ, lots of data.
  Well understood asymptotically
- Knowledge about θ besides D? Not used
- Breaks down if θ "larger" than D
- And another problem . . .

Overfitting

Overfitting Problem (Estimation)

For finite D, more and more complicated models fit D better and better.
With huge brain, you just learn by heart.
Generalization only comes with a limit on complexity!
Marginalization solves this problem, but even Bayesian estimation ("half-way marginalization") embodies complexity control.

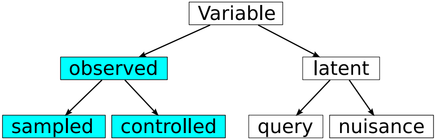


Probabilistic Model

Model

Concise description of joint distribution (generative process) of all variables of interest

- Encoding assumptions:
  What are the entities?
  How do they relate?
- Variables have different roles.
  Roles may change depending on what model is used for



- Model specifies variables and their (in)dependencies
  $\Rightarrow$ Graphical models [next lecture]

The Linear Model (Polynomial Fitting)

Fit data with polynomial (degree k)

- Prior $P(\boldsymbol{w})$
  - Restrictions on $\boldsymbol{w}$
  - Prior knowledge

  $\Rightarrow$ Prefer smaller $\|\boldsymbol{w}\|$
- Likelihood $P(\boldsymbol{y}|\boldsymbol{w})$
- Posterior

$$P(\boldsymbol{w}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{y})}$$

Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \varepsilon$$

$\boldsymbol{y}$	Responses	(observed)
$\boldsymbol{X}$	Design	(controlled)
$\boldsymbol{w}$	Weights	(query)
ε	Noise	(nuisance)

The Linear Model (Polynomial Fitting)

Fit data with polynomial (degree k)

- Prior $P(\boldsymbol{w})$
  - Restrictions on $\boldsymbol{w}$
  - Prior knowledge

  $\Rightarrow$ Prefer smaller $\|\boldsymbol{w}\|$
- Likelihood $P(\boldsymbol{y}|\boldsymbol{w})$
- Posterior

$$P(\boldsymbol{w}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{y})}$$

- Prediction: $y_* = \boldsymbol{x}_*^T \mathrm{E}[\boldsymbol{w}|\boldsymbol{y}]$
- Marginal likelihood

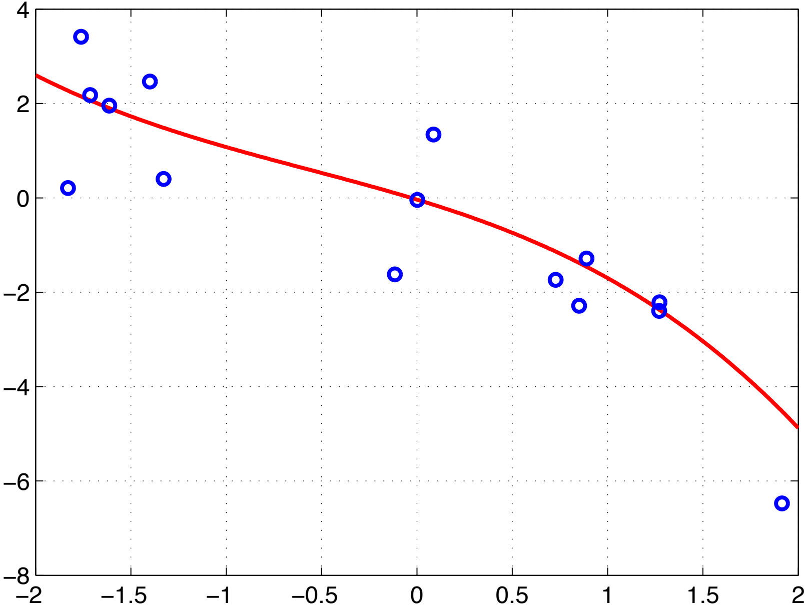
$$P(\boldsymbol{y}) = P(\boldsymbol{y}|k) = \int P(\boldsymbol{y}|\boldsymbol{w})P(\boldsymbol{w})\,d\boldsymbol{w}$$

Linear Model

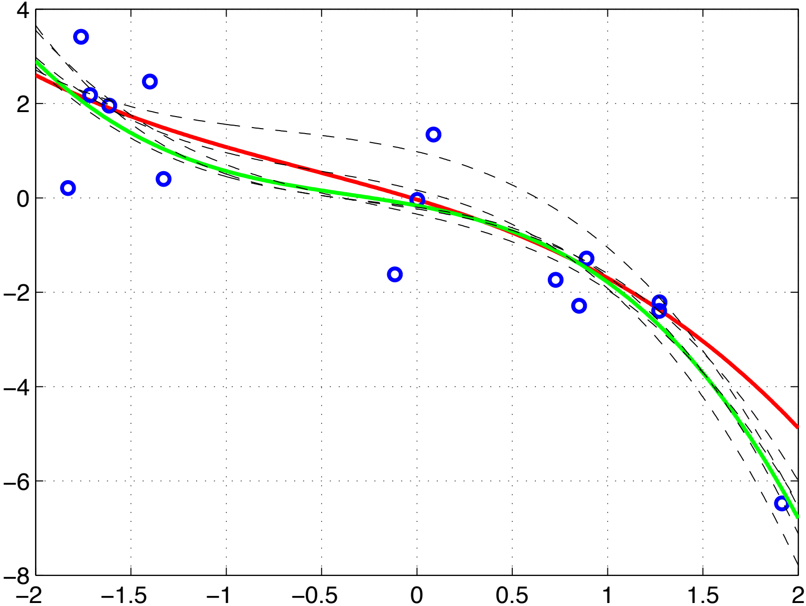
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \varepsilon$$

$\boldsymbol{y}$	Responses	(observed)
$\boldsymbol{X}$	Design	(controlled)
$\boldsymbol{w}$	Weights	(query)
ε	Noise	(nuisance)

The Linear Model (Polynomial Fitting)

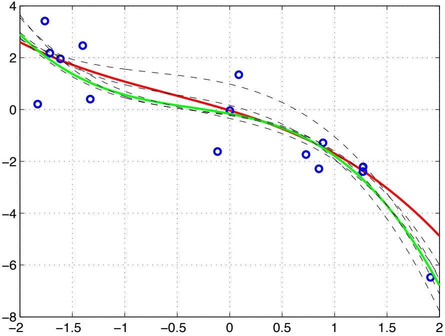


The Linear Model (Polynomial Fitting)

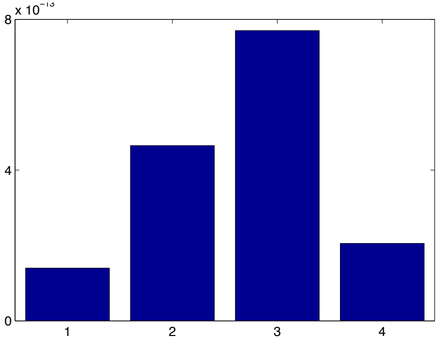


Model Selection

Posterior $P(\boldsymbol{w}|\boldsymbol{y}, k = 3)$



Marginal Likelihood $P(\boldsymbol{y}|k)$



- Simpler hypotheses considered as well
  $\Rightarrow$ Occam's razor

Model Averaging

Don't know polynomial order k $\Rightarrow$ Marginalize out

$$\mathrm{E}[y_*|\boldsymbol{y}] = \sum_{k\geq 1} P(k|\boldsymbol{y})\boldsymbol{x}_*^{(k)\,T}\mathrm{E}[\boldsymbol{w}|\boldsymbol{y},k]$$