Gestalt psychology, Bayesian networks, and Bayesian model comparison

Lecture 5

Done so far

- Population encoding and decoding
- Role of correlations in populations
- Perception as Bayesian inference; explaining visual illusions
- Cue combination: a simple Bayesian computation

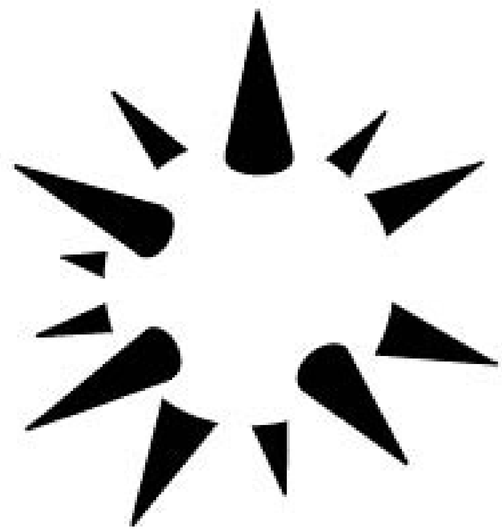
This lecture

- Gestalt psychology: cornerstone of higher-level vision in psychology: beyond sensory uncertainty
- Bayesian models in practice: how to compute probabilities when it gets hard; how to generate behavioral predictions
- Bayesian model comparison: how to show that model A is better than model B; Occam's razor

Gestalt psychology

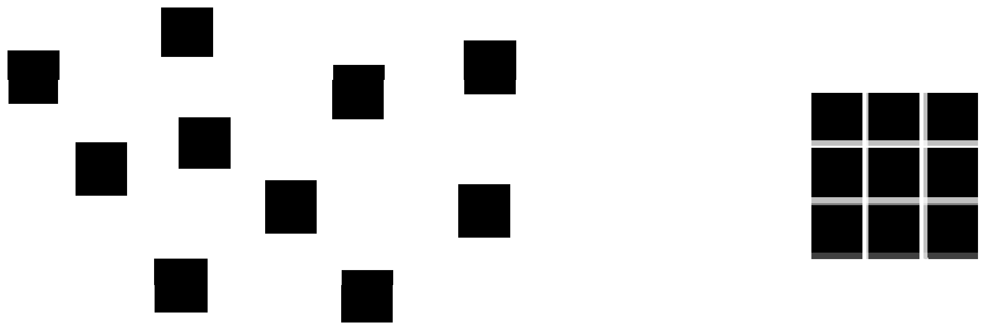
- Observers tend to order their experience in a manner that is regular, orderly, symmetric, and simple.
- “The whole is different than the some of its parts.”
- Gestalt psychologists attempt to discover refinements of this idea → Gestalt “laws of grouping”

Law of closure



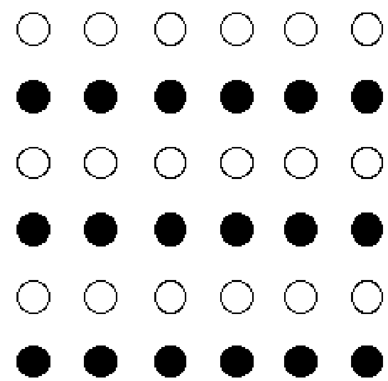
The mind tends to complete incomplete figures (that is, to increase regularity). We may experience elements that are not physically present.

Law of proximity



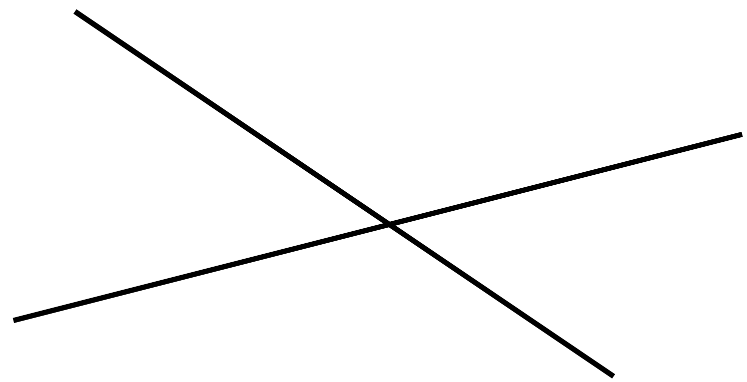
Spatial or temporal proximity of elements may induce the mind to perceive a collective entity.

Law of similarity



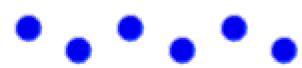
The mind groups similar elements into collective entities. This similarity might depend on relationships of form, color, size, or brightness.

Law of continuity



The mind continues visual, auditory, and kinetic patterns. When something is introduced as a series, the mind tends to perpetuate the series.

Law of common fate

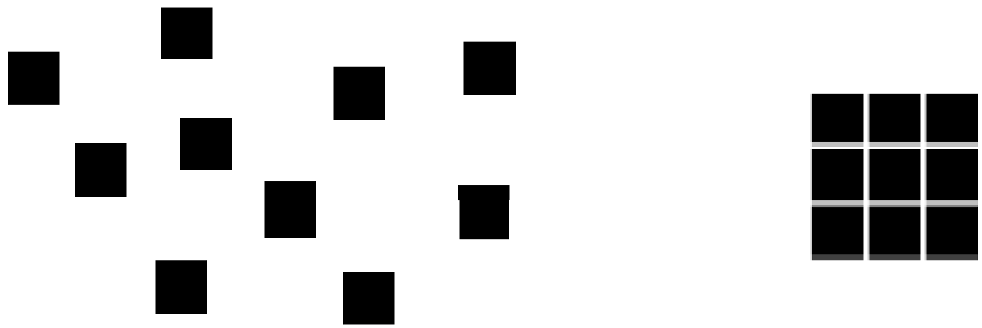


When element move in the same direction, we tend to see them as a collective entity.

Criticisms

- “Vague and inadequate” – V. Bruce et al., 1996
- “Redundant and uninformative” – Wikipedia
- “Haphazard” – Trevor Holland, March 29, 2009
- Descriptive rather than explanatory

Gestalt as Bayesian inference

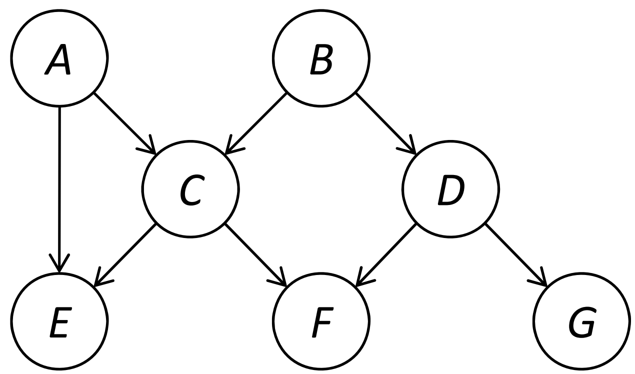


$$p\left(\text{single object} \mid x_1, x_2, \ldots, x_9\right)$$
$$p\left(\text{independent objects} \mid x_1, x_2, \ldots, x_9\right)$$

No sensory uncertainty, but uncertainty about higher-level structure

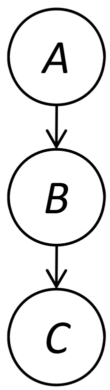
How to compute Bayesian probabilities when it gets hard

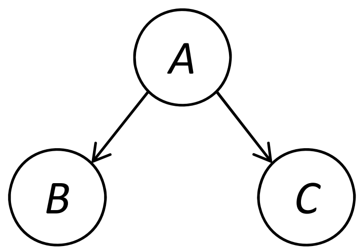
Bayesian networks

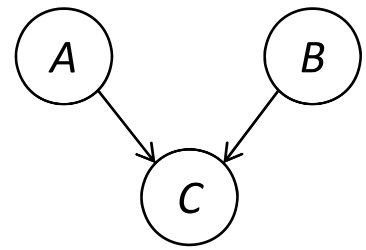


Exercise: Compute $p(A|E,F)$ based on the conditional probabilities indicated in this Bayesian network.

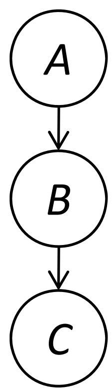
How to compute probabilities in practice







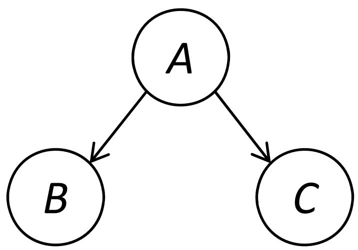
Markov chain



$$p(A,B,C) = p(A)\,p(B \mid A)\,p(C \mid B)$$

$$p(A \mid C) = \frac{p(A,C)}{p(C)} = \frac{\sum_{B} p(A,B,C)}{\sum_{A,B} p(A,B,C)} = \frac{p(A)\sum_{B} p(B \mid A)\,p(C \mid B)}{\sum_{A,B} p(A)\,p(B \mid A)\,p(C \mid B)}$$

Conditional independence

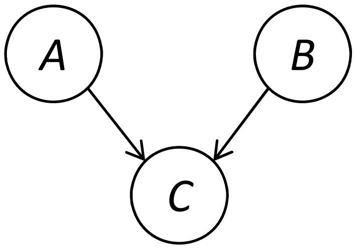


$$p(A,B,C) = p(A)\,p(B\mid A)\,p(C\mid A)$$

$$p(A\mid B,C) = \frac{p(A,B,C)}{p(B,C)} = \frac{p(A)\,p(B\mid A)\,p(C\mid A)}{\sum_A p(A)\,p(B\mid A)\,p(C\mid A)}$$

$$p(A\mid B) = \frac{\sum_C p(A,B,C)}{\sum_{A,C} p(A,B,C)} = \frac{p(A)\,p(B\mid A)}{\sum_A p(A)\,p(B\mid A)}$$

Independent sources



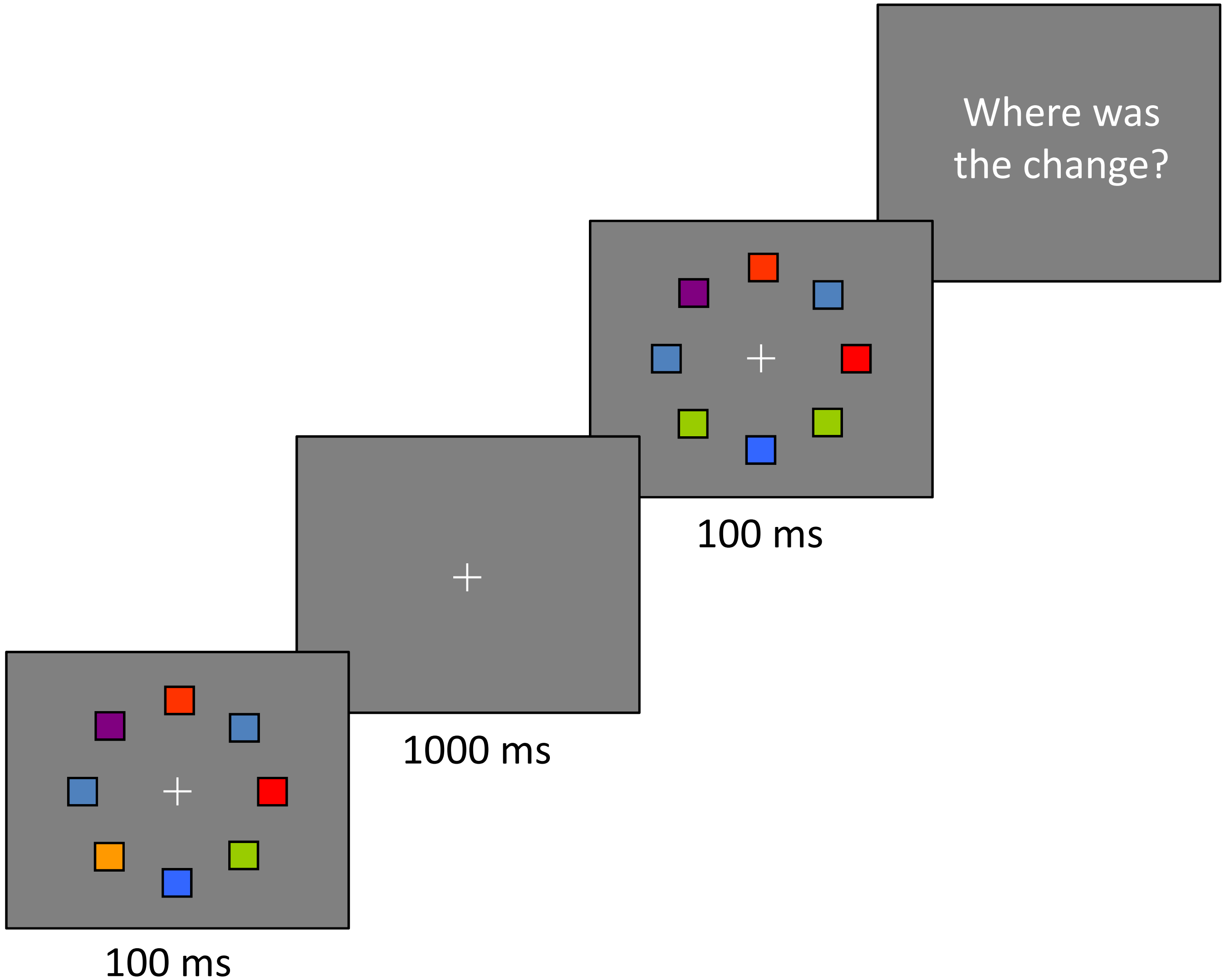
$$p(A,B,C) = p(A)\,p(B)\,p(C \mid A,B)$$

$$p(A \mid B,C) = \frac{p(A)\,p(B)\,p(C \mid A,B)}{\sum_{A} p(A)\,p(B)\,p(C \mid A,B)}$$

$$p(A \mid C) = \frac{p(A)\sum_{B} p(B)\,p(C \mid A,B)}{\sum_{A,B} p(A)\,p(B)\,p(C \mid A,B)}$$

How to predict behavioral data?

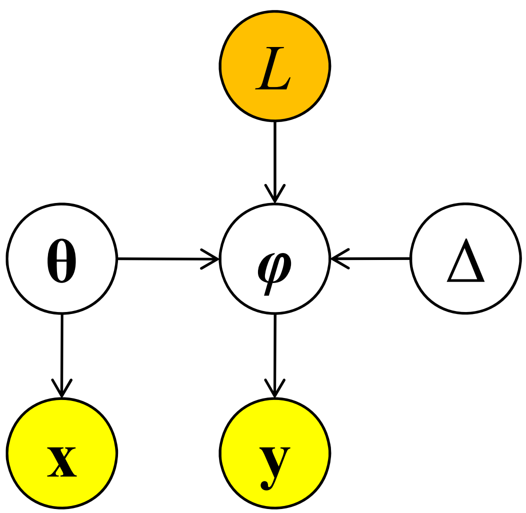
Example: change localization



Step 1: What are the parameters?

- Number of items N (assumed known)
- Where did the change occur? $L = 1,\ldots,N$
- How big was the change? Δ
- What were the original features? $\theta_1,\ldots, \theta_N$
- What were the new features? $\varphi_1,\ldots, \varphi_N$
- Internal representations of original features: $x_1,\ldots, x_N$
- Internal representations of new features: $y_1,\ldots, y_N$

Step 2: Draw generative model, write down prior and conditional probabilities



$$p(L) = \frac{1}{N}$$

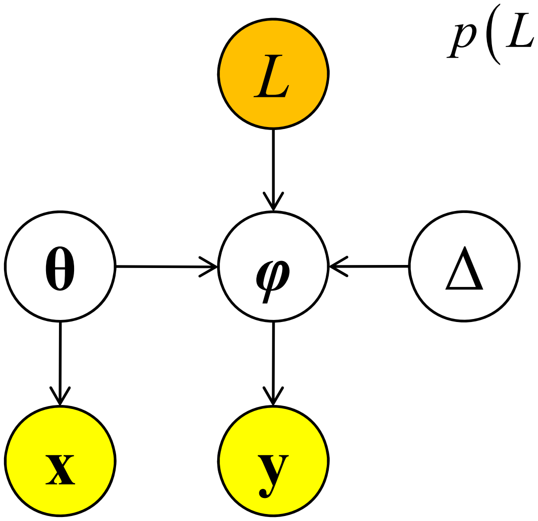
$$p(\theta_i) = p(\Delta) = \text{constant}$$

$$p(\boldsymbol{\varphi} \mid \boldsymbol{\theta}, L, \Delta) = \delta(\boldsymbol{\varphi} - \boldsymbol{\theta} - \Delta \mathbf{1}_L)$$

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i \mid \theta_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{x,i}^2}} e^{-\frac{(x_i - \theta_i)^2}{2\sigma_{x,i}^2}}$$

$$p(\mathbf{y} \mid \boldsymbol{\varphi}) = \prod_{i=1}^{N} p(y_i \mid \varphi_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{y,i}^2}} e^{-\frac{(y_i - \varphi_i)^2}{2\sigma_{y,i}^2}}$$

Step 3: Compute the posterior over the task variable using probability calculus



$$p(L,\mathbf{x},\mathbf{y},\boldsymbol{\theta},\boldsymbol{\varphi}) = p(L)\,p(\boldsymbol{\theta})\,p(\Delta)\,p(\boldsymbol{\varphi}\,|\,\boldsymbol{\theta},L,\Delta)\,p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\mathbf{y}\,|\,\boldsymbol{\varphi})$$

$$p(L\,|\,\mathbf{x},\mathbf{y}) \propto p(L,\mathbf{x},\mathbf{y}) = \iiint p(L,\mathbf{x},\mathbf{y},\boldsymbol{\theta},\boldsymbol{\varphi})\,d\Delta\,d\boldsymbol{\theta}\,d\boldsymbol{\varphi}$$

$$= \iiint p(L)\,p(\boldsymbol{\theta})\,p(\Delta)\,p(\boldsymbol{\varphi}\,|\,\boldsymbol{\theta},L,\Delta)\,p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\mathbf{y}\,|\,\boldsymbol{\varphi})\,d\Delta\,d\boldsymbol{\theta}\,d\boldsymbol{\varphi}$$

$$\propto \int\left(\int p(\mathbf{x}\,|\,\boldsymbol{\theta})\left(\int p(\boldsymbol{\varphi}\,|\,\boldsymbol{\theta},L,\Delta)\,p(\mathbf{y}\,|\,\boldsymbol{\varphi})\,d\boldsymbol{\varphi}\right)d\boldsymbol{\theta}\right)d\Delta$$

$$= \int\left(\int p(\mathbf{x}\,|\,\boldsymbol{\theta})\left(\int \delta(\boldsymbol{\varphi}-\boldsymbol{\theta}-\Delta\mathbf{1}_L)\,p(\mathbf{y}\,|\,\boldsymbol{\varphi})\,d\boldsymbol{\varphi}\right)d\boldsymbol{\theta}\right)d\Delta$$

$$= \int\left(\int p(\mathbf{x}\,|\,\boldsymbol{\theta})\,p(\mathbf{y}\,|\,\boldsymbol{\varphi}=\boldsymbol{\theta}+\Delta\mathbf{1}_L)\,d\boldsymbol{\theta}\right)d\Delta$$

$$= \int\left(\prod_{i=1}^{N}\int \frac{1}{\sqrt{2\pi\sigma_{x,i}^2}}e^{-\frac{(x_i-\theta_i)^2}{2\sigma_{x,i}^2}}\frac{1}{\sqrt{2\pi\sigma_{y,i}^2}}e^{-\frac{(y_i-\theta_i-\Delta\mathbf{1}_{L,i})^2}{2\sigma_{y,i}^2}}\,d\theta_i\right)d\Delta$$

$$= \ldots \propto \sqrt{2\pi\left(\sigma_{x,L}^2+\sigma_{y,L}^2\right)}\,e^{\frac{(x_L-y_L)^2}{2\left(\sigma_{x,L}^2+\sigma_{y,L}^2\right)}}$$

Step 4: Pick a decoder (e.g. MAP)

$$\hat{L}(\mathbf{x},\mathbf{y}) = \underset{L}{\operatorname{argmax}} \sqrt{\sigma_{x,L}^2 + \sigma_{y,L}^2}\, e^{\frac{(x_L - y_L)^2}{2\left(\sigma_{x,L}^2 + \sigma_{y,L}^2\right)}}$$

Step 5: Monte Carlo simulation

Draw many sets of **x**, **y** (trials) from generative model but with priors given by experiment, in each experimental condition separately.

Compute $\hat{L}(\mathbf{x},\mathbf{y})$ on each trial.

→ Histograms $p\left(\hat{L} \mid \text{experimental condition}\right)$

How to compare models to data?

What makes model A better than model B?

→ If it describes the data better…

→ What do we mean by “describing better”?

→ Lower error, higher goodness-of-fit…

→ What is the right error or goodness-of-fit measure to use?

→ Look up in statistics book / pull out of hat (t-test, R^2, χ^2, SSE, …)

Maximum-likelihood fitting

- Data D
- Model M

$$p(M \mid D) \propto \underbrace{p(D \mid M)}_{\text{Model likelihood}} \underbrace{p(M)}_{\text{Flat model prior}}$$

- Find model with highest likelihood

$$\operatorname*{argmax}_{M} \; p(D \mid M)$$

Maximum-likelihood fitting

- Model parameters θ
- Find parameters that work best for given model

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, p\left(D \mid M, \theta\right)$$

$$p\left(D \mid M\right) = p\left(D \mid M, \hat{\theta}_{\mathrm{ML}}\right)$$

- Repeat for all candidate models

Example: linear regression

- Data: $D = (X,Y)$
- Model M:
  $y = ax + b$ + Gaussian noise with fixed variance

$$\begin{aligned} p(D \mid M,\theta) &= p(X,Y \mid a,b,\sigma) \\ &= p(Y \mid X,a,b,\sigma)\, p(X) \\ &= p(X) \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - aX_i - b)^2}{2\sigma^2}} \end{aligned}$$

$$\left(\hat{a},\hat{b}\right) = \underset{a,b}{\operatorname{argmin}} \sum_i \left(Y_i - aX_i - b\right)^2$$

Example: probability distributions

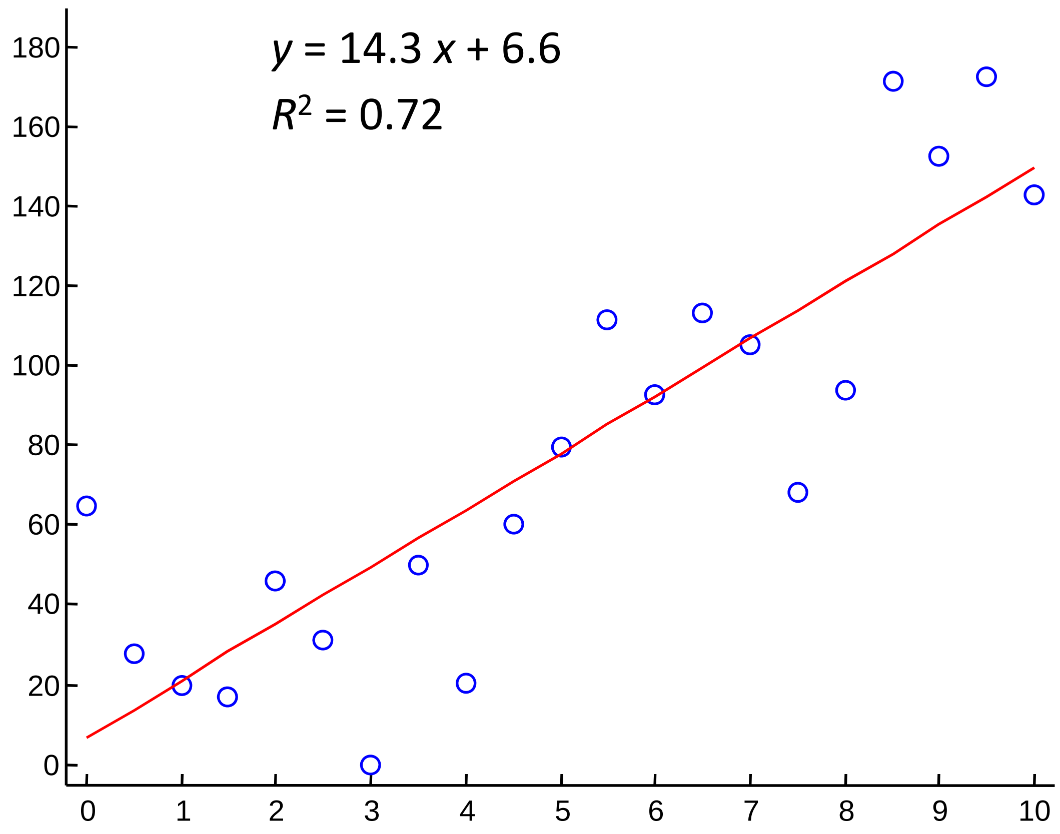
- Data: histogram (n_1,n_2,..,n_B)
- Model M: n_i drawn from multinomial with probabilities $p_i(\theta)$

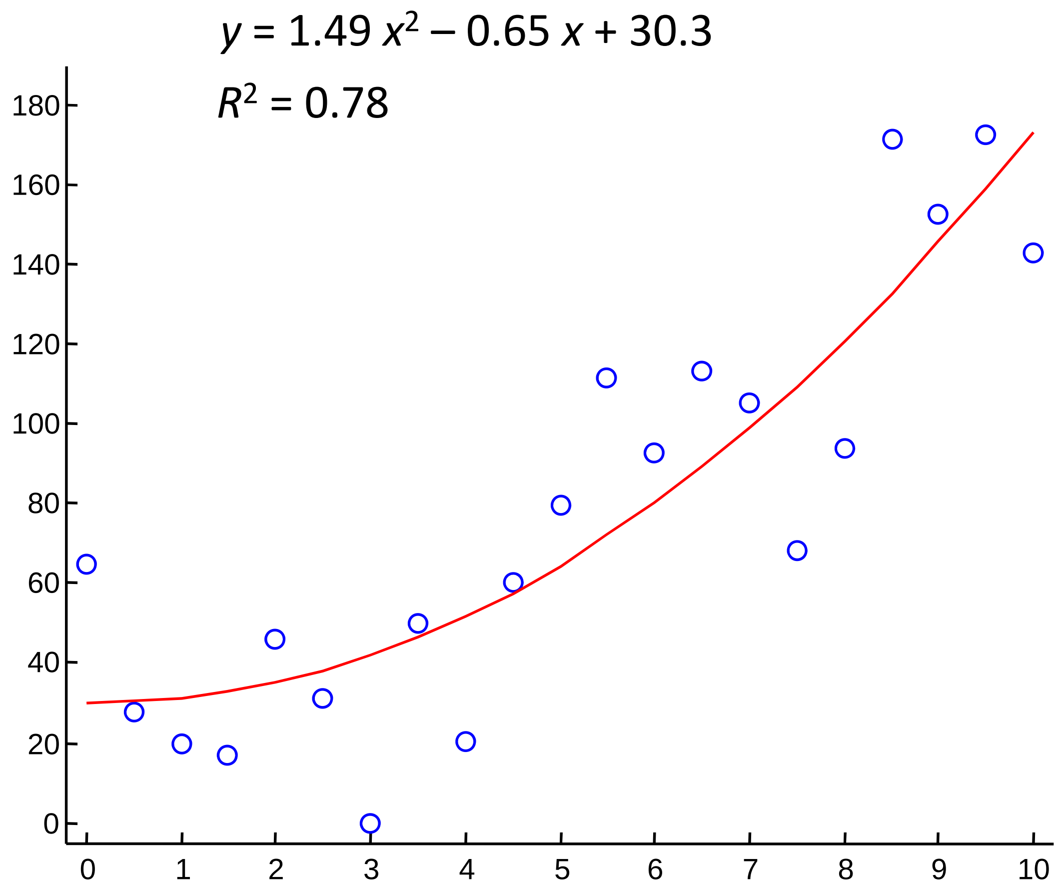
$$p\left(D \mid M, \theta\right) = p\left(\mathbf{n} \mid \mathbf{p}\left(\theta\right)\right)$$

$$= \frac{\left(n_1 + \cdots + n_B\right)!}{n_1! \cdots n_B!} p_1\left(\theta\right)^{n_1} \cdots p_B\left(\theta\right)^{n_B}$$

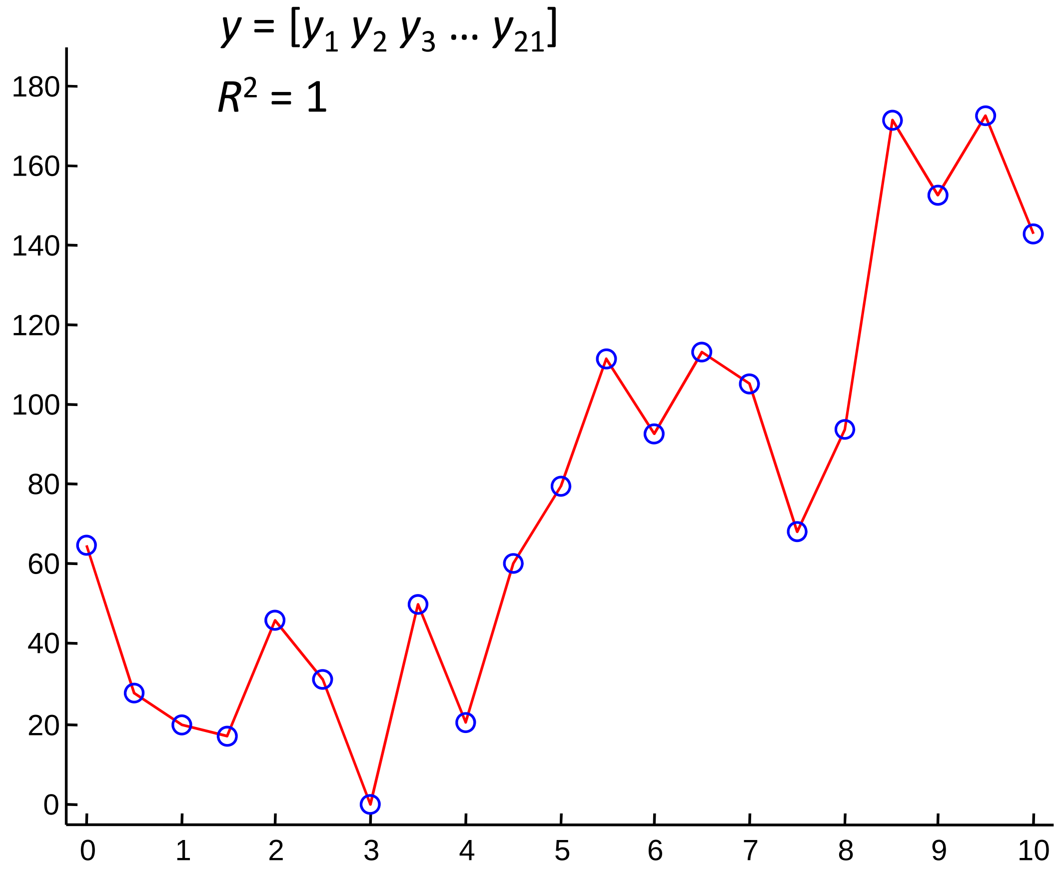
$$\log p\left(D \mid M, \theta\right) = \sum_{i=1}^{B} n_i \log p_i\left(\theta\right) + \text{constant}$$

Is a better fit always better?





y = 1.49 x2 – 0.65 x + 30.3
R2 = 0.78
0
20
40
60
80
100
120
140
160
180
0
1
2
3
4
5
6
7
8
9
10



Why is this not a good model?

Occam's razor (parsimony)

- "Simpler models are better"
- Simpler: fewer assumptions, fewer parameters
- But not a rigorous formulation
- Can only decide between two models that fit the data equally well
- Balance between complexity and power

→ Bayesian model comparison

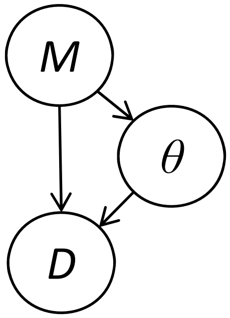
Bayesian model comparison

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} p(D \mid M, \theta)$$

$$p(D \mid M) = p(D \mid M, \hat{\theta}_{\mathrm{ML}})$$

$$p(\theta \mid D, M) \propto p(D \mid M, \theta)\, p(\theta \mid M)$$

$$p(D \mid M) = \int p(D \mid M, \theta)\, p(\theta \mid M)\, d\theta$$



goodness of fit averaged over all possible parameter combinations

How does this help?

Assume $p(\theta|M)$ is flat

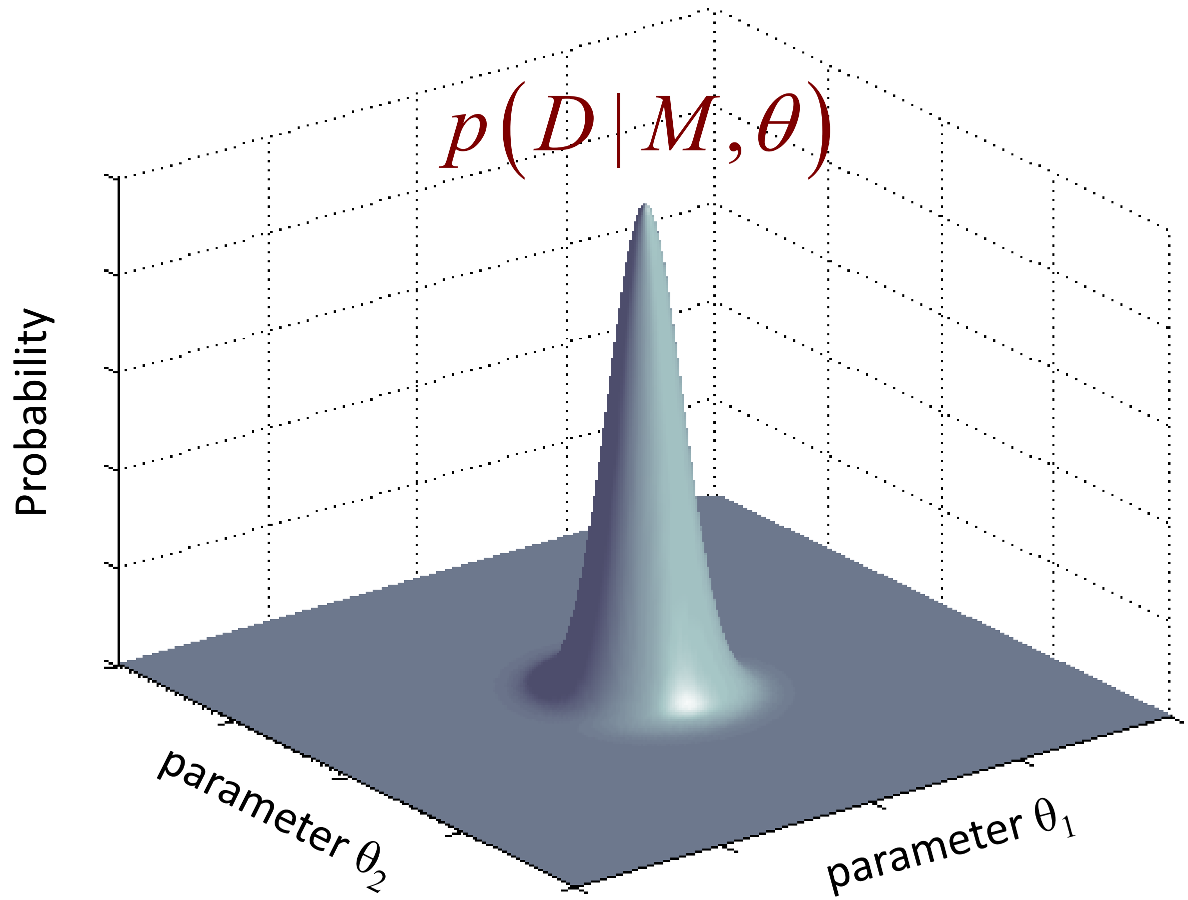
$$p(\theta \mid M) = \frac{1}{\text{Volume of parameter space}}$$

$$p(\theta \mid D, M) \propto p(D \mid M, \theta)$$

$$p(D \mid M) = \frac{1}{\text{Volume of parameter space}} \int p(D \mid M, \theta)\, d\theta$$

Many parameters → large volume

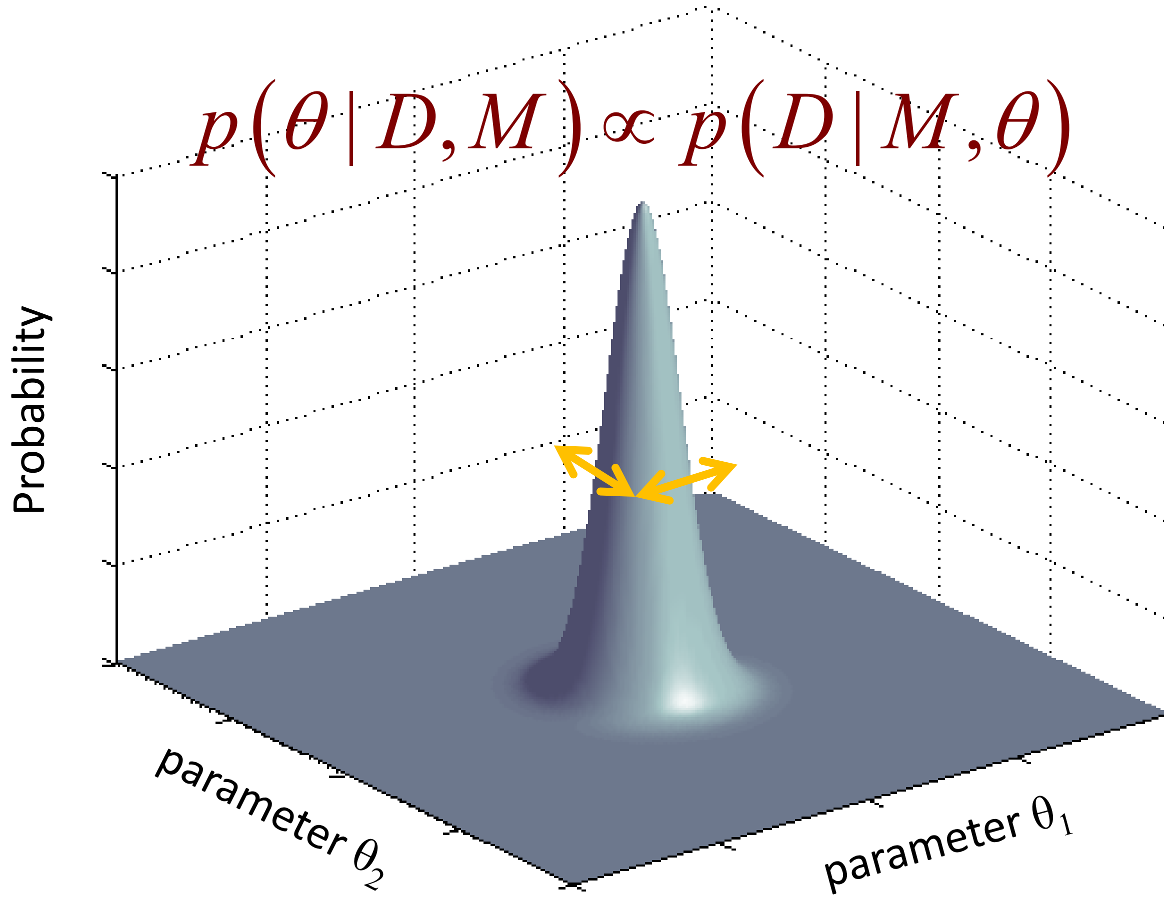
Likelihood landscape



p(*D*|*M*,θ) is high if the data are fit well *compared to other possible data*

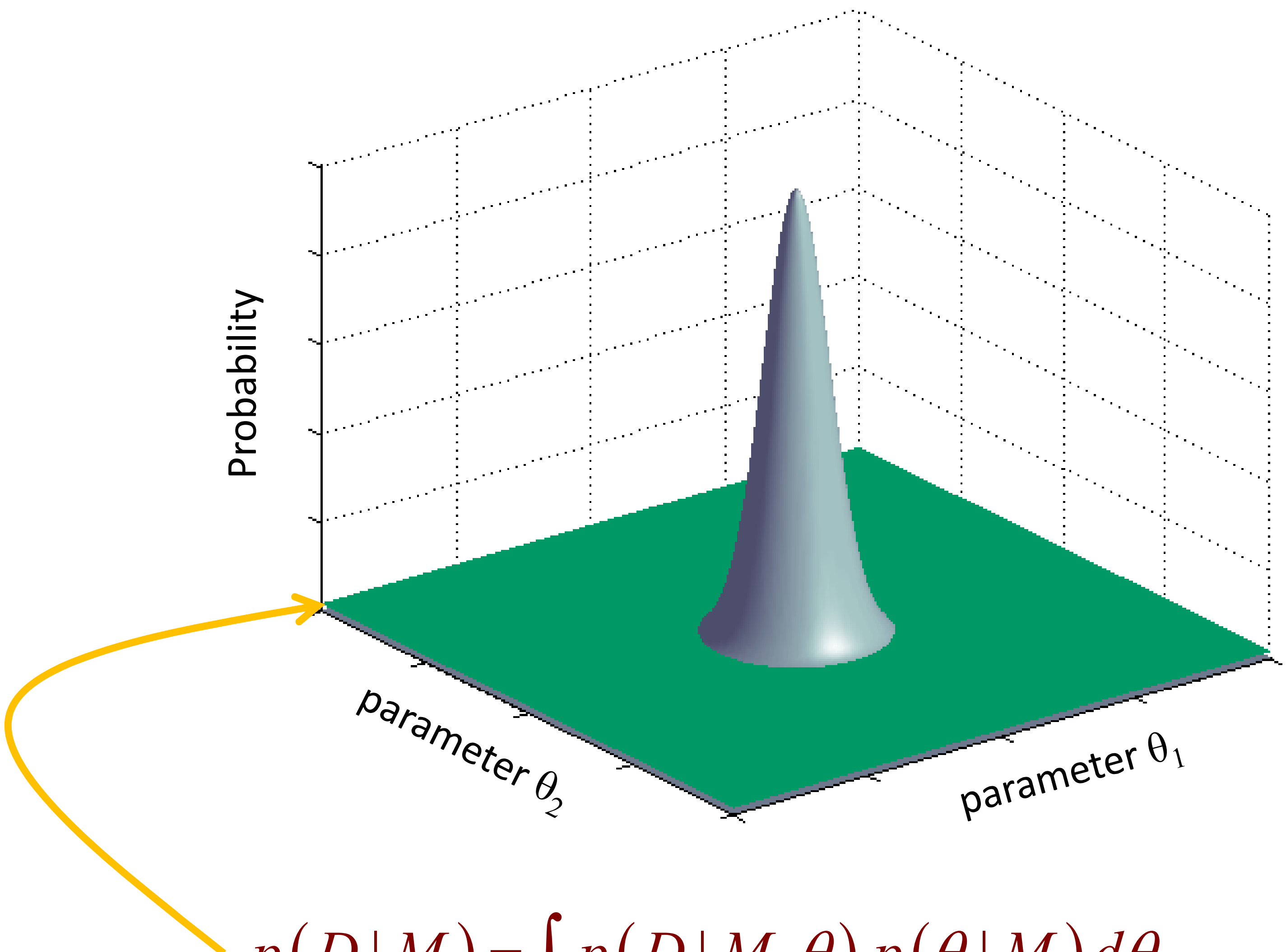
Normalized

$$p(\theta | D, M) \propto p(D | M, \theta)$$



Error bars on parameters

Unnormalized but averaged



$$p(D|M) = \int p(D|M,\theta)\, p(\theta|M)\, d\theta$$

Bayesian model comparison

$$p(D|M) = \int p(D|M,\theta)\, p(\theta|M)\, d\theta$$

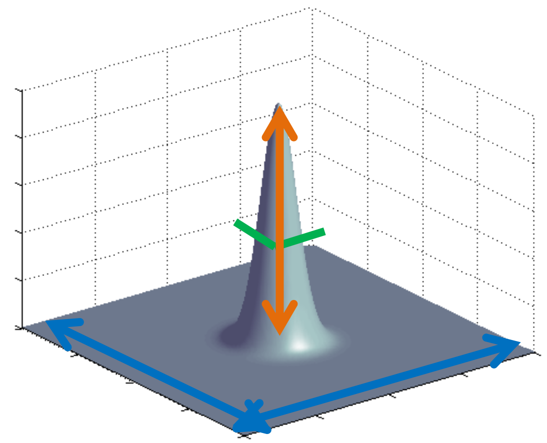
- Penalizes poorly fitting models ($p(D|M,\theta)$ low overall)
- Penalizes non-specific models (peak of $p(D|M,\theta)$ is low, since it is normalized over D)
- Penalizes models that have to be finely tuned (width of $p(D|M,\theta)$ is low)
- Penalizes models with many parameters (low $p(\theta|M)$)
- Penalizes models with poor choice of prior range of parameters ($p(\theta|M)$ doesn't overlap with $p(D|M,\theta)$)

How to compute the integral?

$$p(D|M)=\int p(D|M,\theta)\,p(\theta|M)\,d\theta$$

- Sum over all possible parameter combinations?
- Say 4 parameters, each parameter takes 50 values, each model simulation takes 10 ms → 17 hours
- Approximation would be useful!

Approximating it..



- Peak of $p(D\,|\,M,\theta)$ is $p\left(D\,|\,M,\hat{\theta}_{\mathrm{MAP}}\right)$
- Width of $p(D\,|\,M,\theta)$ is $\sigma_{\theta|D}$
- Width of $p(\theta\,|\,M)$ is σ_{θ}

Then

$$p(D\,|\,M) = \int p(D\,|\,M,\theta)\,p(\theta\,|\,M)\,d\theta$$

$$\approx p\left(D\,|\,M,\hat{\theta}_{\mathrm{MAP}}\right)p\left(\hat{\theta}_{\mathrm{MAP}}\,|\,M\right)\sigma_{\theta|D}$$

$$\approx p\left(D\,|\,M,\hat{\theta}_{\mathrm{MAP}}\right)\frac{\sigma_{\theta|D}}{\sigma_{\theta}}$$

Occam factor

Compare

$$p(D\,|\,M) = p\left(D\,|\,M,\hat{\theta}_{\mathrm{ML}}\right)$$

Laplace approximation

$$p(D \mid M) \approx p\left(D \mid M, \hat{\theta}_{\text{MAP}}\right) p\left(\hat{\theta}_{\text{MAP}} \mid M\right) \frac{1}{\sqrt{\det \frac{\mathbf{H}}{2\pi}}}$$

Hessian of the -log posterior:

$$\mathbf{H} = -\nabla\nabla \log p\left(\theta \mid D, M\right)\Big|_{\theta=\hat{\theta}_{\text{MAP}}}$$

Exercises:

- Prove this.
- What is **H** when the posterior is a multivariate Gaussian centered at $\hat{\theta}_{\text{MAP}}$?

Goodness of a model

$$p(M \mid D) \propto p(D \mid M)\, p(M)$$

$$p(D \mid M) = \int p(D \mid M, \theta)\, p(\theta \mid M)\, d\theta$$

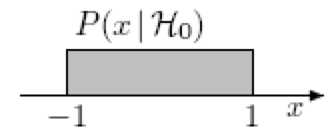
Relative goodness of two models:

$$\log \frac{p(D \mid M_1)\, p(M_1)}{p(D \mid M_2)\, p(M_2)} = \log \frac{p(M_1)}{p(M_2)} + \log \frac{\int p(D \mid M_1, \theta)\, p(\theta \mid M_1)\, d\theta}{\int p(D \mid M_2, \theta)\, p(\theta \mid M_2)\, d\theta}$$

Exercises

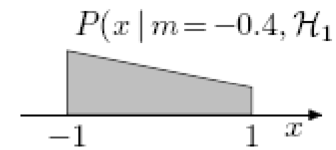
Exercise 28.1.[3] Random variables x come independently from a probability distribution $P(x)$. According to model $\mathcal{H}_0$, $P(x)$ is a uniform distribution

$$P(x \mid \mathcal{H}_0) = \frac{1}{2} \qquad x \in (-1, 1). \tag{28.20}$$



According to model $\mathcal{H}_1$, $P(x)$ is a nonuniform distribution with an unknown parameter $m \in (-1, 1)$:

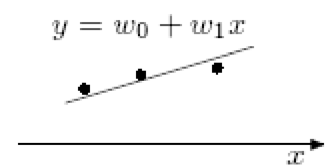
$$P(x \mid m, \mathcal{H}_1) = \frac{1}{2}(1 + mx) \qquad x \in (-1, 1). \tag{28.21}$$



Given the data $D = \{0.3, 0.5, 0.7, 0.8, 0.9\}$, what is the evidence for $\mathcal{H}_0$ and $\mathcal{H}_1$?

Exercise 28.2.[3] Datapoints (x, t) are believed to come from a straight line. The experimenter chooses x, and t is Gaussian-distributed about

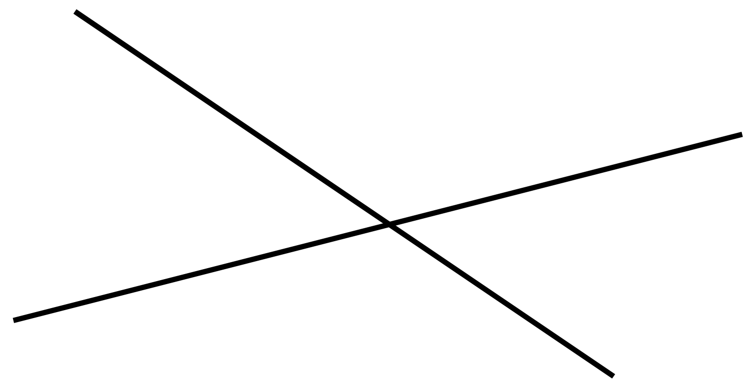
$$y = w_0 + w_1 x \tag{28.22}$$



with variance σ_ν^2. According to model $\mathcal{H}_1$, the straight line is horizontal, so $w_1 = 0$. According to model $\mathcal{H}_2$, w_1 is a parameter with prior distribution Normal$(0,1)$. Both models assign a prior distribution Normal$(0,1)$ to w_0. Given the data set $D = \{(-8, 8), (-2, 10), (6, 11)\}$, and assuming the noise level is $\sigma_\nu = 1$, what is the evidence for each model?

David MacKay, *Information theory, inference, and learning algorithms* (2003)

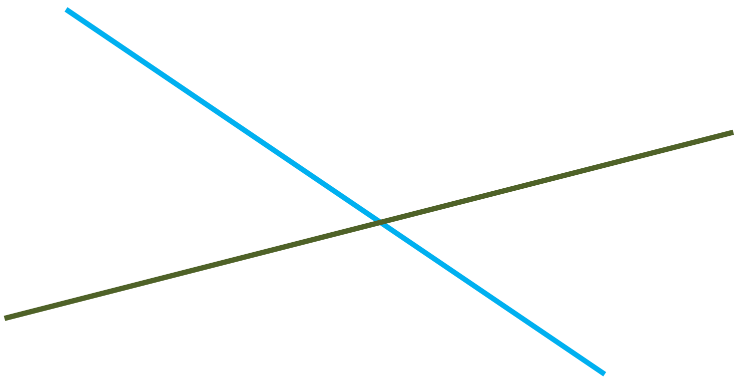
Bayesian model comparison and Gestalt laws



"Law of continuity"

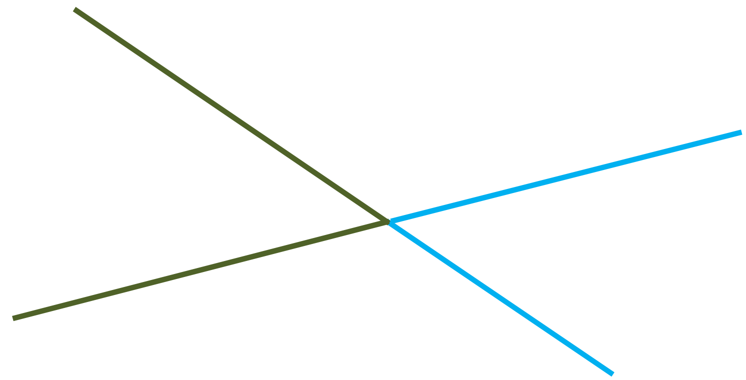
Bayesian model comparison

Model 1



2 lines
Each line 2 free parameters
→ 4 free parameters
Assume each takes 50 values
Uniform priors

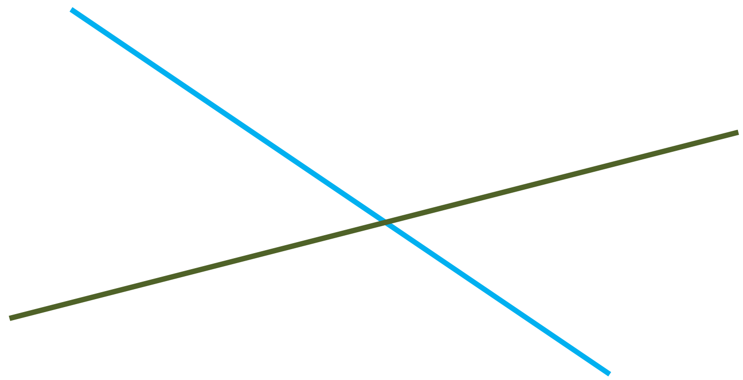
Model 2



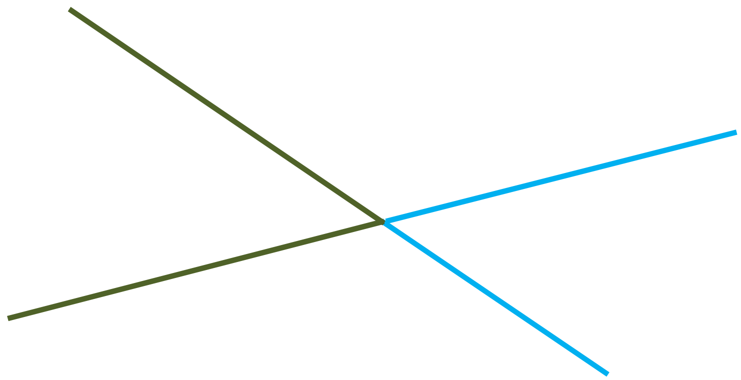
2 angles
Each angle 4 free parameters
→ 8 free parameters
Assume each takes 50 values
Uniform priors

Bayesian model comparison

Model 1



Model 2



$$p(D|M_1) = \int p(D|M_1,\theta)\,p(\theta|M_1)\,d\theta \approx 1 \cdot \left(\frac{1}{50}\right)^4$$

$$p(D|M_2) \approx \left(\frac{1}{50}\right)^8$$

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)\,p(M_1)}{p(D|M_2)\,p(M_2)} = 50^4 \approx 6250000$$

Open questions

- Can the Gestalt laws be written as outcomes of Bayesian model comparison?
- Can such Bayesian models be tested by changing parameters and measuring human behavior?
- How is hierarchical inference implemented in neural networks?

Small project

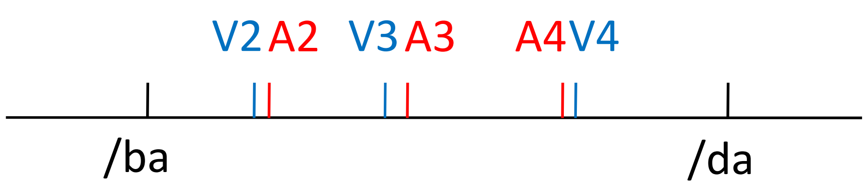
- Auditory-visual speech perception data
- Identify a syllable as /ba/ or /da/
- Factorial design
- In each condition, % responses "/ba/" and "/da/"

		Auditory					
		BA	2	3	4	DA	none
Visual	BA						
	2						
	3						
	4						
	DA						
	none						

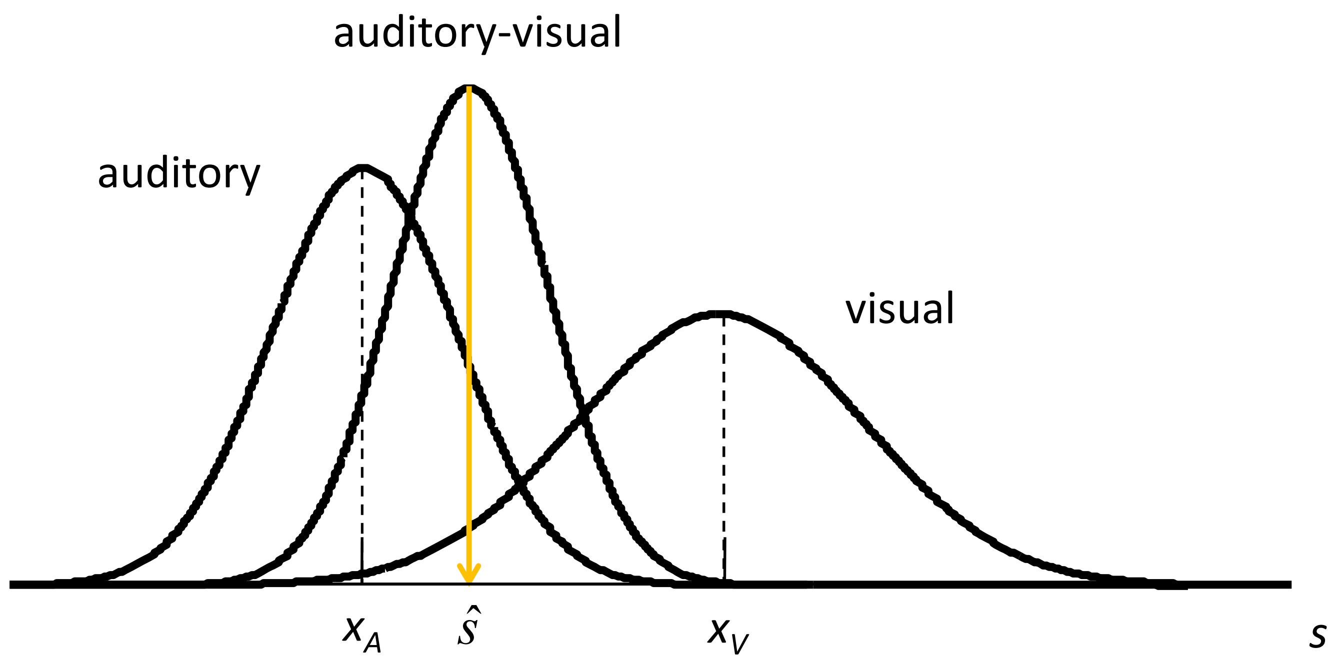
Massaro et al., 1993
http://mambo.ucsc.edu/psl/data/mass93a.html

Approach

1. Model structure
   a) Inference model vs modeler's model
   b) What are the free parameters?
   c) First pass: fix feature values of intermediates (equidistant, equal between modalities)



2. Predict responses using Bayesian model
   a) Assume conditional independence
   b) Collapse onto two categories
   c) Assume variances independent of *s*
   d) Make other assumptions if necessary



3. Is the Bayesian model better than the established model?
   a) Work out alternative model (FLMP; multiplies response frequencies)
   b) Maximum-likelihood fitting
   c) Bayesian comparison (integrate over free parameters; approximate where necessary)
4. Discuss results and caveats

Due by Saturday, April 11