ECSE-626

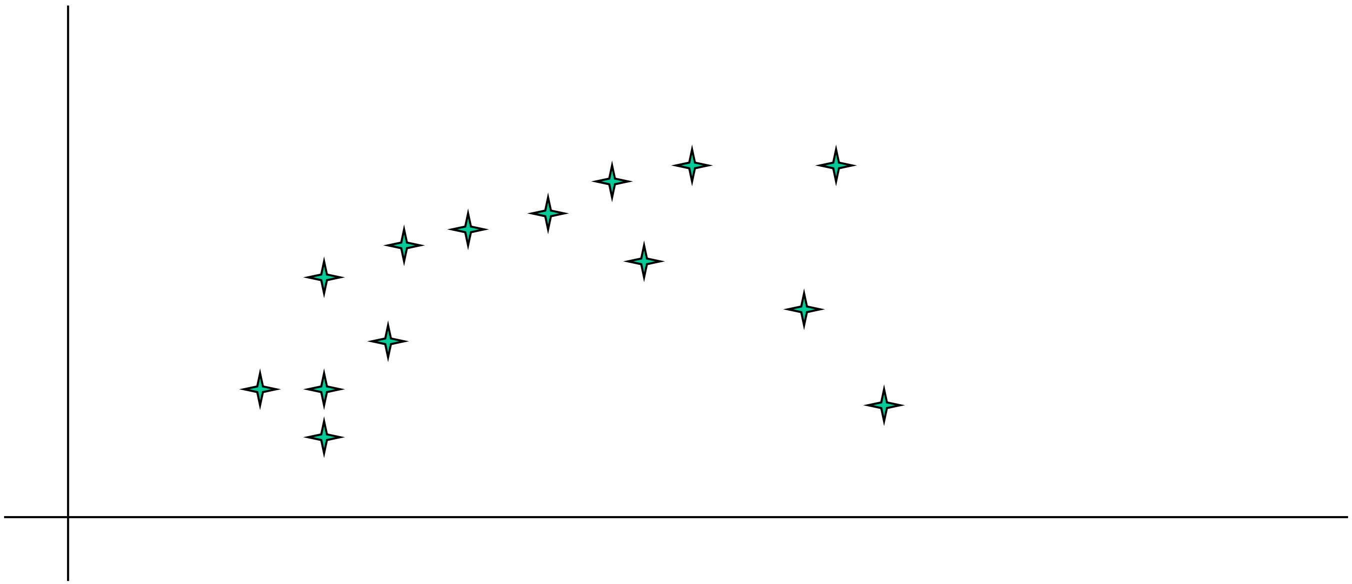
Statistical Computer Vision

Bayesian Inference and Data Modelling

Bayesian Inference

- Let us now examine Bayesian inference in the context of real data modeling problems, such as those found in computer vision.
- We begin with the simple problem of fitting data to a model.

Data Modelling

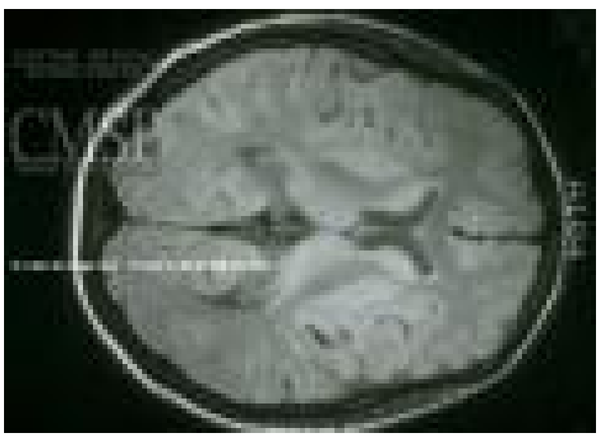


Problem: Given a set of data points acquired from a set of measurements, infer the underlying model that generated the points.

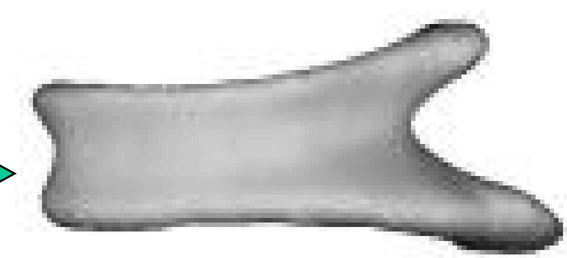
Data Modelling

- This problem is typical in computer vision:

Brain ventricle



catalog.cmsp.com

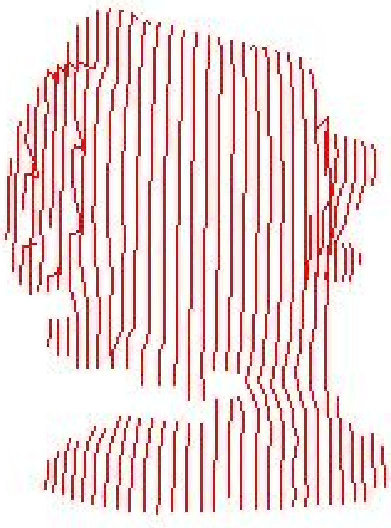


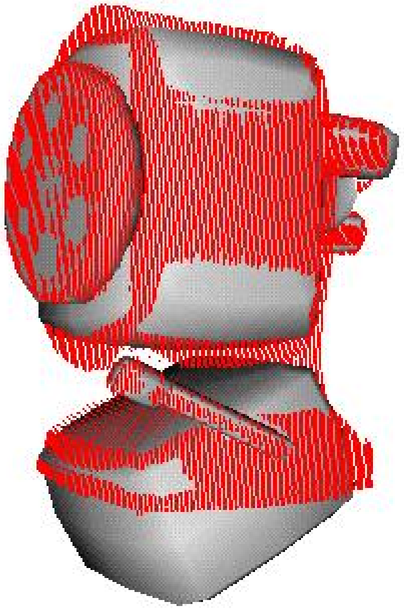
homepages.inf.ed.ac.uk

Data Modelling

Range image

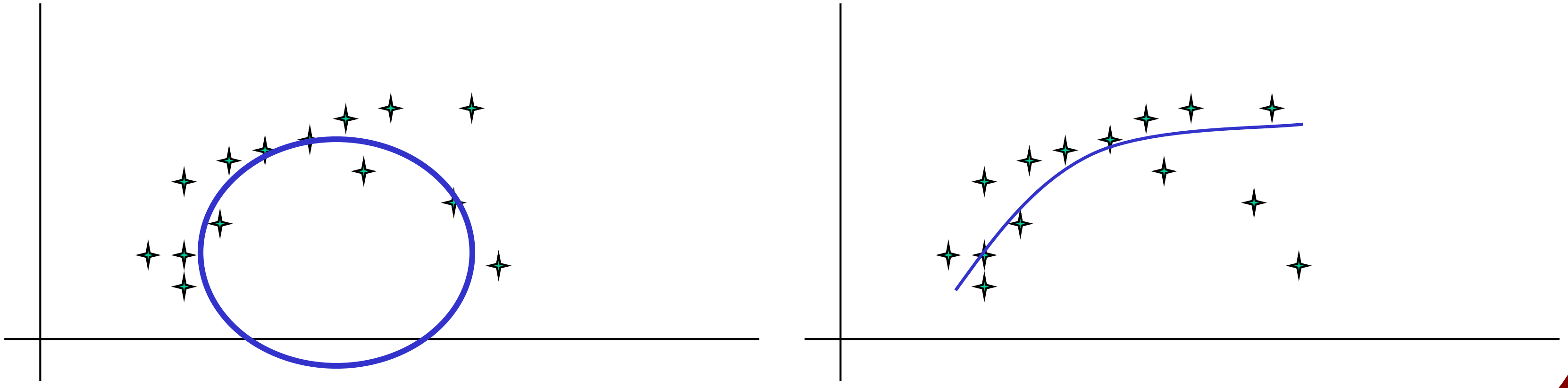
Parametric model





Data Modelling

- This problem is ill-posed in that many different models could lead to the same measurements. The uniqueness constraint is violated.



Data Modelling

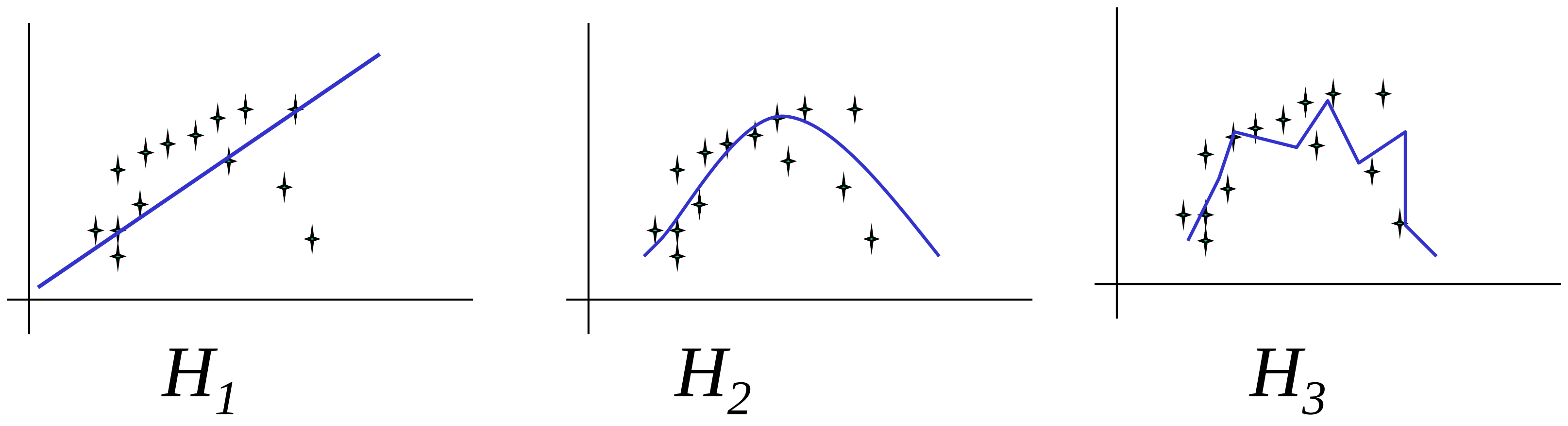
What is the difference between a frequentist and a Bayesian approach to this problem?

Data Modelling

- Frequentist approach: Try to estimate the "true" underlying model through optimization.
- Bayesian approach: There is no *true* model. Even if there were, we will never know what it is due to the uncertainty in the measurement process. The best we can do is try to infer the probability of competing hypotheses given the measurements.

Data Modelling

- Let H_i denote the different hypotheses about the types of models possible:



Data Modelling

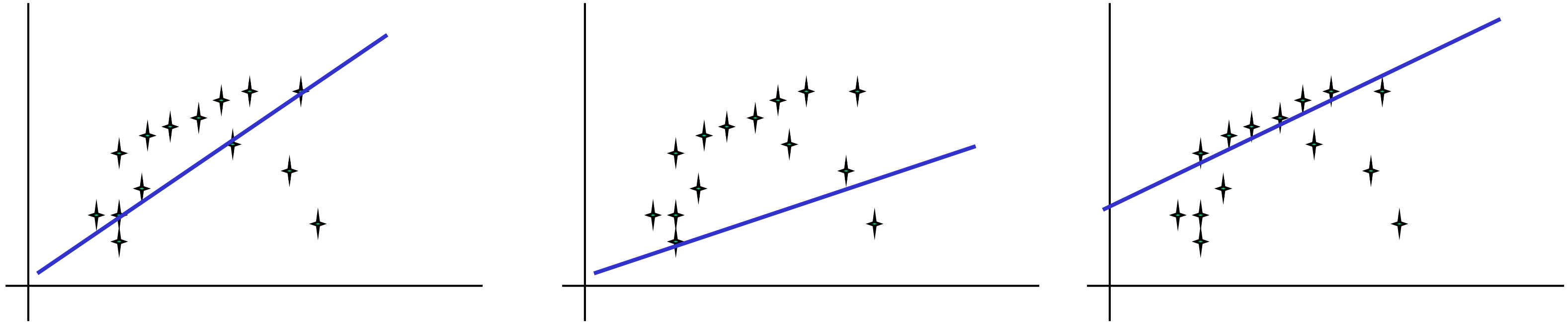
- There are two stages of inference to this problem:
  1. Model fitting
  2. Model comparison

Data Modelling

- There are two stages of inference to this problem:
  1. Model fitting
  2. Model comparison

Model Fitting

- At the *first* level of inference, we assume a particular model hypothesis is true (e.g. H_1). We infer what the model parameters $\boldsymbol{w}$ might be given the data:



Model Fitting

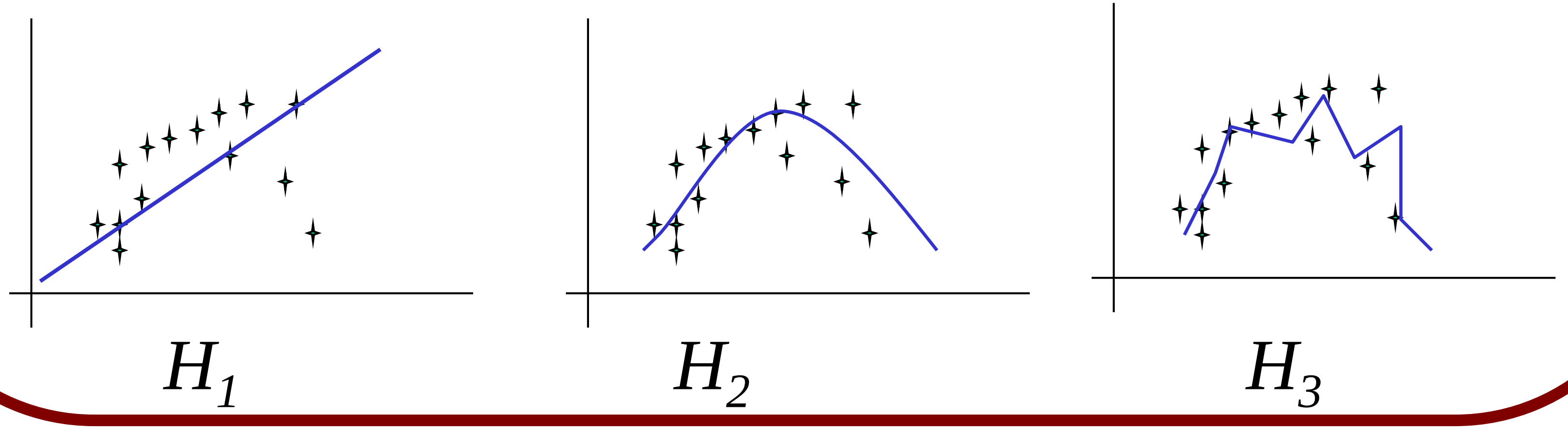
- This process is repeated for each model.
- The results of this inference stage are often (but not always) summarized by the most probable parameter values, $\mathbf{w}_{MP}$, and error bars on these parameters.
- If the result is in this format, applying Bayesian methods to this problem is little different from the solution given by orthodox statistics.

Data Modelling

- There are two stages of inference to this problem:
  1. Model fitting
  2. Model comparison

Model Comparison

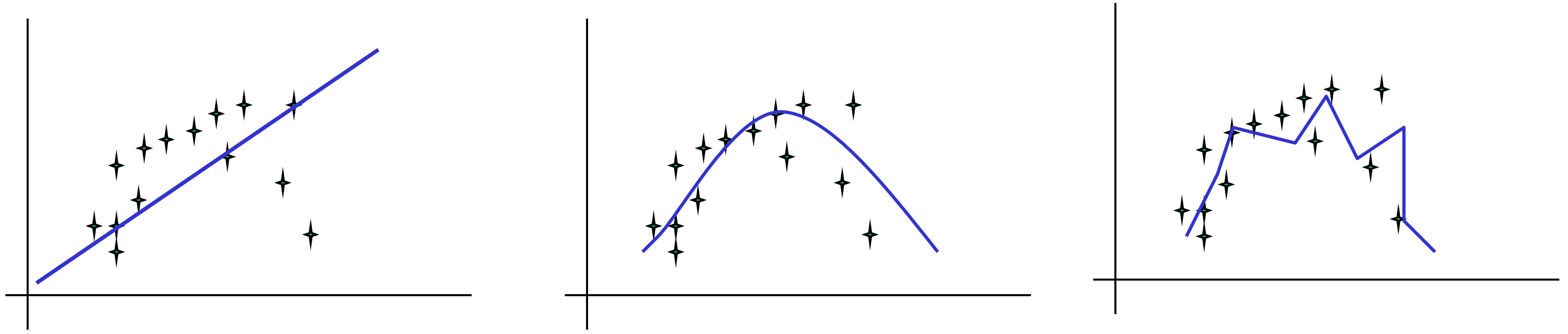
- At the *second* level of inference, we wish to infer which model is most plausible given the data.
- We wish to compare the models in light of the data, and assign rankings to the alternatives.



Model Comparison

- It is here that Bayesian methods are different from maximum likelihood methods.

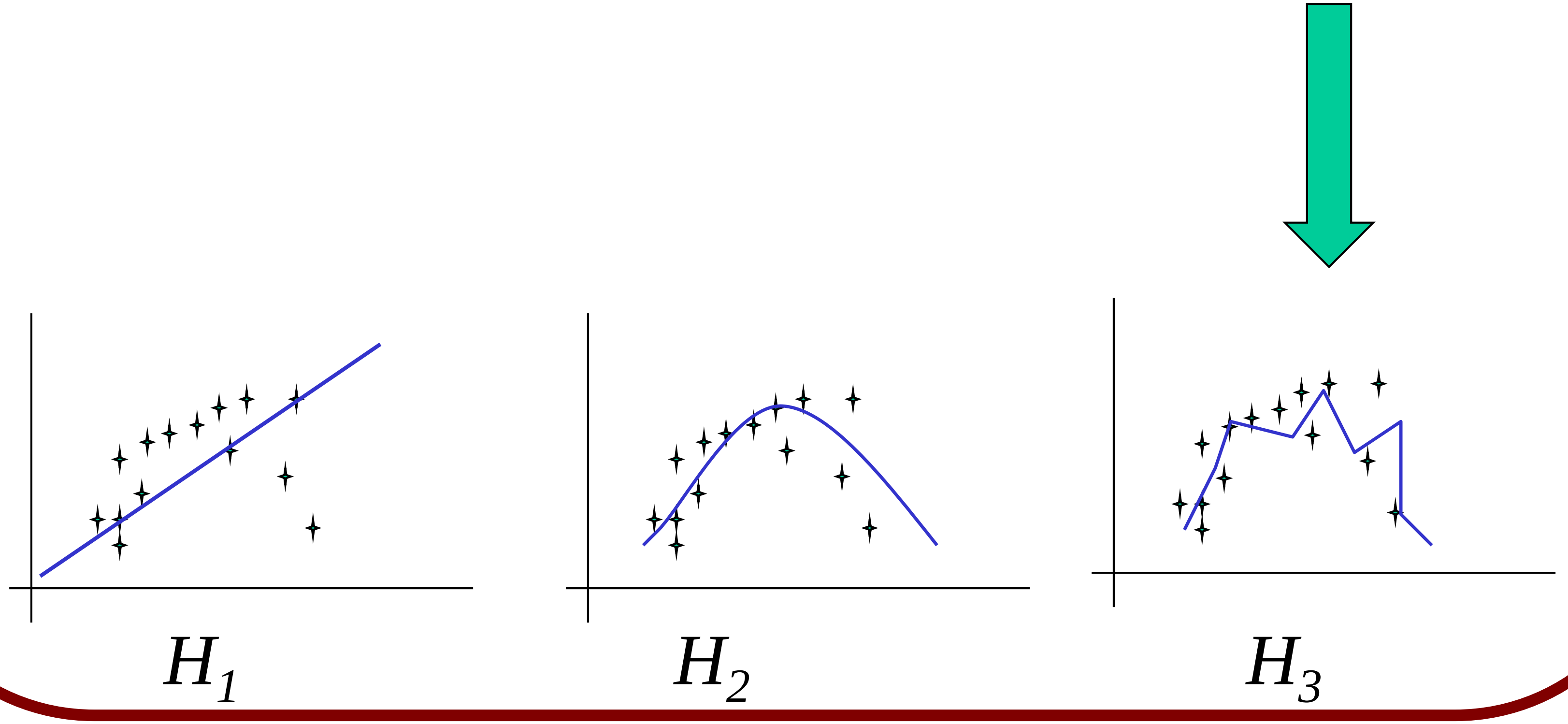
- Why don't we simply choose the model that fits the data best?



H_1 H_2 H_3

Model Comparison

- It is not simply a matter of choosing the model that fits the data best – complex models will usually fit the data better



Model Comparison

- Maximum likelihood models would lead to over-parametrized models that don’t generalize well.

Model Comparison

- Recall that the ill-posedness of the problem is related to the solution space.

- Intuitively, one can see that the complex model defines a solution subspace in F which renders the problem more ill-posed than the case of the simplest model.

Why?

Model Comparison

- Recall that the ill-posedness of the problem is related to the solution space.
- Intuitively, one can see that the complex model defines a solution subspace in F which renders the problem more ill-posed than the case of the simplest model.
- This is because of the *instability* caused by the model complexity.

Model Comparison

- How do both the Tikhonov and the Bayesian approaches render the problem well-posed?

Model Comparison

- Tikhonov regularization methods:
  - Define a “smoothing” stabilizer on a subset of F.
  - Perform unconstrained optimization to find solution, z, which minimizes the stabilizer subject to consistency with the data.

- What would an equivalent Bayesian regularization scheme involve?

Model Comparison

- An equivalent Bayesian regularization scheme would map to developing a strong prior which favors the "smoother" solution.

Recall:

$$P(u \mid z) = \frac{\exp(-\beta\rho_U(Az,u))}{Z_1}$$

$$P(z) = \frac{\exp(-\beta\alpha\Omega[z])}{Z_2}$$

$$P(u) = \frac{Z}{Z_1 Z_2}$$

Model Comparison

- An equivalent Bayesian regularization scheme would map to developing a strong prior which favors the "smoother" solution.

$$P(H_i \mid D) \propto P(D \mid H_i)P(H_i)$$

Model Comparison

- Is this necessary?
- Do we need to add in a strong prior to favor the simpler model?
- What if we have no such predisposition?

Model Comparison

- It turns out that we do not!
- Bayesian methods embed *Occam's razor* that naturally penalizes overly complex models:

Occam's Razor

If several explanations are compatible with a set of observations, Occam's razor advises us to choose the simplest.

"Coherent inference (as embodied by Bayesian probability) automatically embodies Occam's razor, quantitatively."[1]

[1]Mackay, Chap. 28, pg. 344.

Model Comparison & Occam's Razor

- Suppose we wish to evaluate the plausibility of H_1 (simpler) and H_2 (more complex) in light of data D using Bayes' Theory:

$$\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{P(H_1)}{P(H_2)} \frac{P(D | H_1)}{P(D | H_2)}$$

$\frac{P(H_1)}{P(H_2)}$ Measures how much our initial beliefs favor H_1 over H_2.

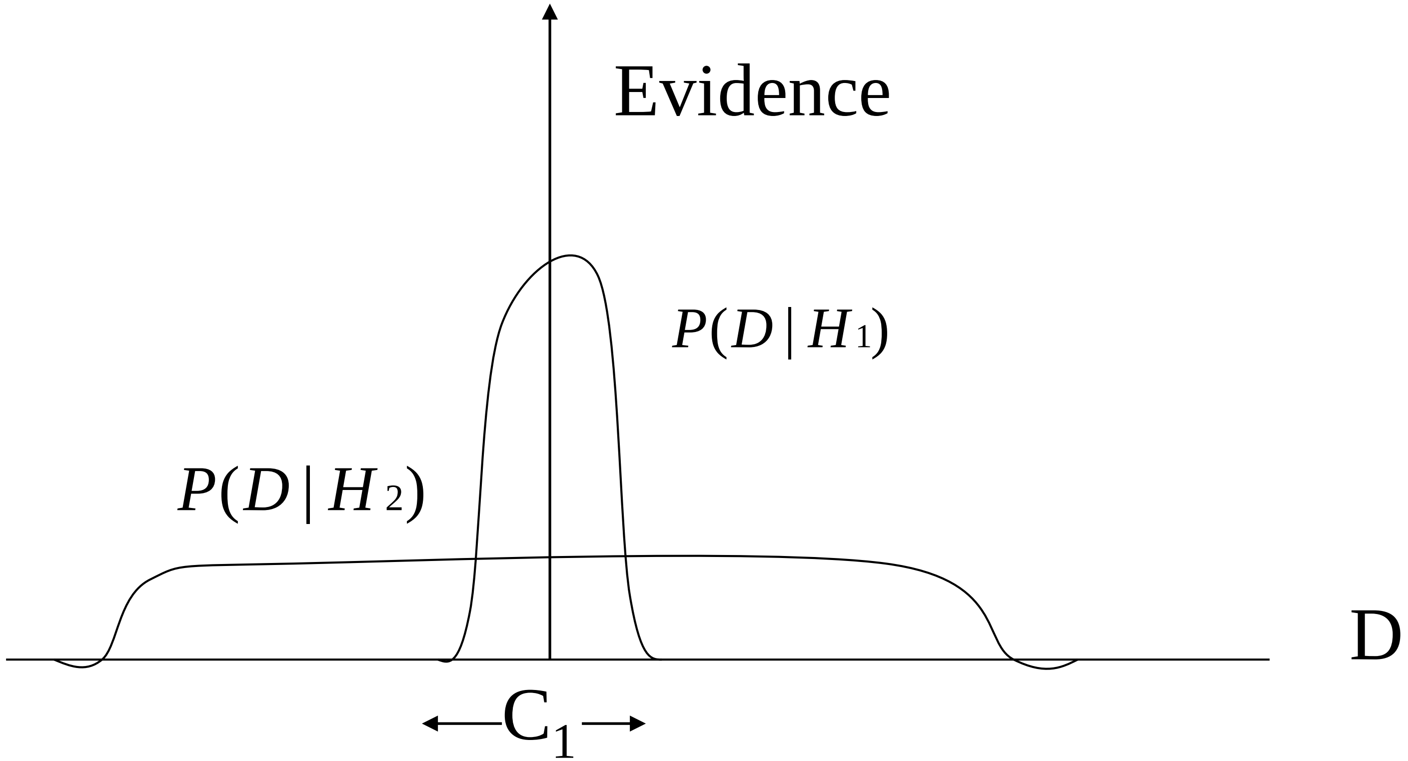
$\frac{P(D | H_1)}{P(D | H_2)}$ Measures how well observed data were predicted by H_1 over H_2.

Model Comparison & Occam's Razor

$$\frac{P(H_1)}{P(H_2)}$$

Gives us the opportunity to insert prior bias based on experience, aesthetic reasons.
Here we can state explicit constraint as in regularization.
This prior bias is not necessary in that the data-dependent factor embodies *Occam's razor* automatically.

Model Comparison & Occam's Razor



Model Comparison & Occam's Razor

- Simple models tend to make more precise predictions. Complex models capable of making greater variety of predictions – $P(D|H_2)$ more thinly spread out.
- H_2 does not predict the data sets in C_1 as strongly as H_1.

Model Comparison & Occam's Razor

- When:
  - Data are compatible in with both models,
  - Equal prior probably probabilities assigned to both models,

  Then:
  - *Simpler* model will be more probable than *complex* model, if data falls in region C_1.

Model Fitting

- At the *first* level of inference, we assume a particular model hypothesis is true (e.g. H_1). We infer what the model parameters $\boldsymbol{w}$ might be given the data:

$$P(\mathbf{w} \mid D, H_i) = \frac{P(D \mid \mathbf{w}, H_i) P(\mathbf{w} \mid H_i)}{P(D \mid H_i)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Model Fitting

- Common practice to use gradient methods to find maximum of posterior: $\mathbf{w}_{MP}$.
- Summarize posterior by value of $\mathbf{w}_{MP}$, and confidence intervals on best fit parameters, attained from curvature of posterior.
- Evaluating the Hessian $\mathbf{A}$ at $\mathbf{w}_{MP}$, and Taylor expanding the log posterior probability with

  $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{MP}$:

$$P(\mathbf{w} \mid D, H_i) \simeq P(\mathbf{w}_{MP} \mid D, H_i) \exp(-\frac{1}{2}\Delta\mathbf{w}\mathbf{A}\Delta\mathbf{w})$$

Model Fitting

- Posterior locally approximated by Gaussian with covariance matrix $\mathbf{A}^{-1}$
- Whether or this approximation is good or not depends on problem we are solving.

Model Comparison

- We wish to infer the most plausible model given the data. The posterior probability for each model is:

$$P(H_i \mid D) \propto P(D \mid H_i) P(H_i)$$

$P(D \mid H_i)$ Evidence for H_i

$P(H_i)$ Subjective prior over hypothesis space: How plausibility of alternative models before data arrived.

Model Comparison

- Assuming equal priors, rank hypotheses by evaluating the *evidence*, which embodies Occam's razor.
- Evidence can be evaluated for parametric, non-parametric models, any data modelling problem: regression, density estimation, classification.

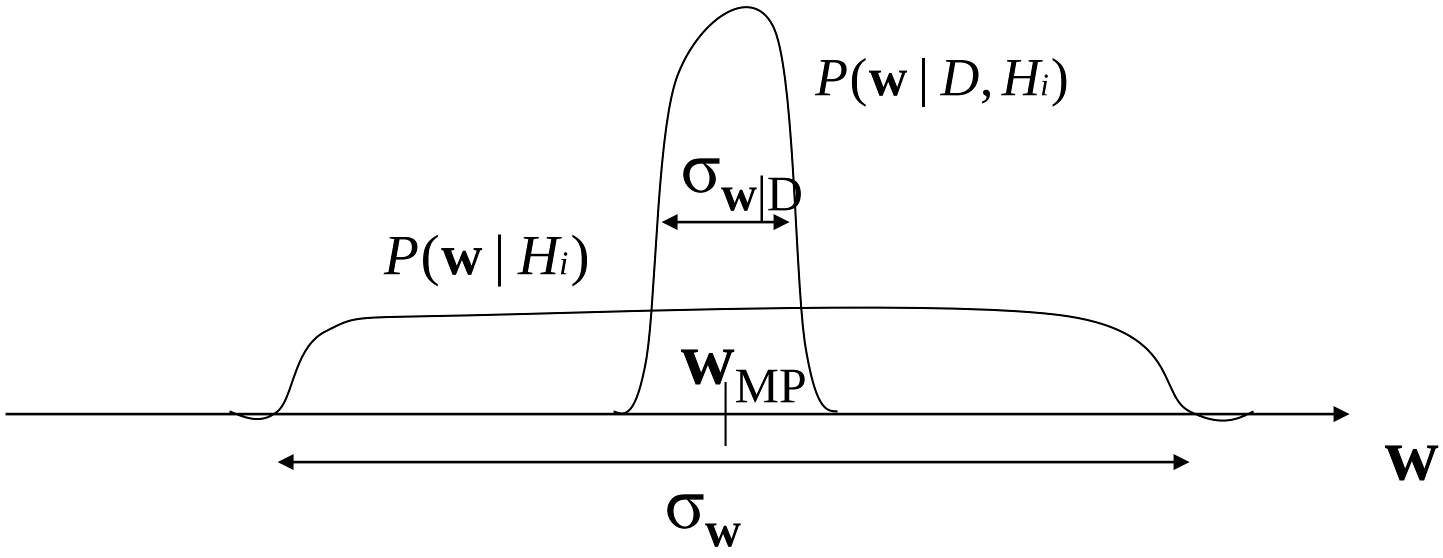
Occam's Razor

$$P(\mathbf{w} \mid D, H_i) = \frac{P(D \mid \mathbf{w}, H_i) P(\mathbf{w} \mid H_i)}{P(D \mid H_i)}$$

where $P(D \mid H_i) = \int P(D \mid \mathbf{w}, H_i) P(\mathbf{w} \mid H_i) d\mathbf{w}$

For many problems, posterior has strong peak at $\mathbf{w}_{\mathrm{MP}}$.

Occam's Razor



Occam's Razor

Evidence can then be approximated using Laplace's method –
height of peak of integrand $P(D \mid \mathbf{w}, H_i)P(\mathbf{w} \mid H_i)$
times width: $\sigma_{\mathbf{w}|D}$

$$P(D \mid H_i) \simeq \underbrace{P(D \mid \mathbf{w}_{MP}, H_i)} \times \underbrace{P(\mathbf{w}_{MP} \mid H_i)\sigma_{w \mid D}}$$

$$\text{Evidence} \simeq \text{Best fit likelihood} \times \text{Occam factor}$$

Occam Factor

If $P(\mathbf{w} \mid H_i)$ is uniform on some large interval, σ_w,

$$P(\mathbf{w}_{\mathrm{MP}} \mid H_i) = \frac{1}{\sigma_w}$$

$$\text{Occam factor} = \frac{\sigma_{w \mid D}}{\sigma_w}$$

Occam factor is equal to the ratio of the posterior accessible volume of H_i's parameter space to the prior accessible volume.

Occam Factor

- Logarithm of Occam factor measures amount of information gain about model parameters after data arrives.
- Complex model penalized more by Occam factor than simpler model (due to larger $\sigma_{\mathbf{w}}$).
- Occam factor also penalizes models that have to be fine tuned to fit data (smaller $\sigma_{\mathbf{w}|D}$).
- Model that achieves greatest evidence: tradeoff between minimizing natural complexity measure and minimizing data misfit.

Occam Factor

- Occam factor for a model easy to evaluate – depends on error bars on parameters, which were already evaluated when fitting the model to the data.

Inference vs. Decision Theory

- Inference is different from decision theory.
- Goal of inference: Given defined hypothesis space and particular data set, assign probabilities to hypotheses.
- Goal of decision theory: choose between alternative *actions* on basis of these probabilities so as to minimize expectation of “loss function”.

Inference vs. Decision Theory

- Model comparison does not imply *model choice* – making predictions based on all information.
- Decision theory would involve:
  - Choosing future actions
  - Deciding whether to create new models
  - Deciding what data to gather next (active vision)