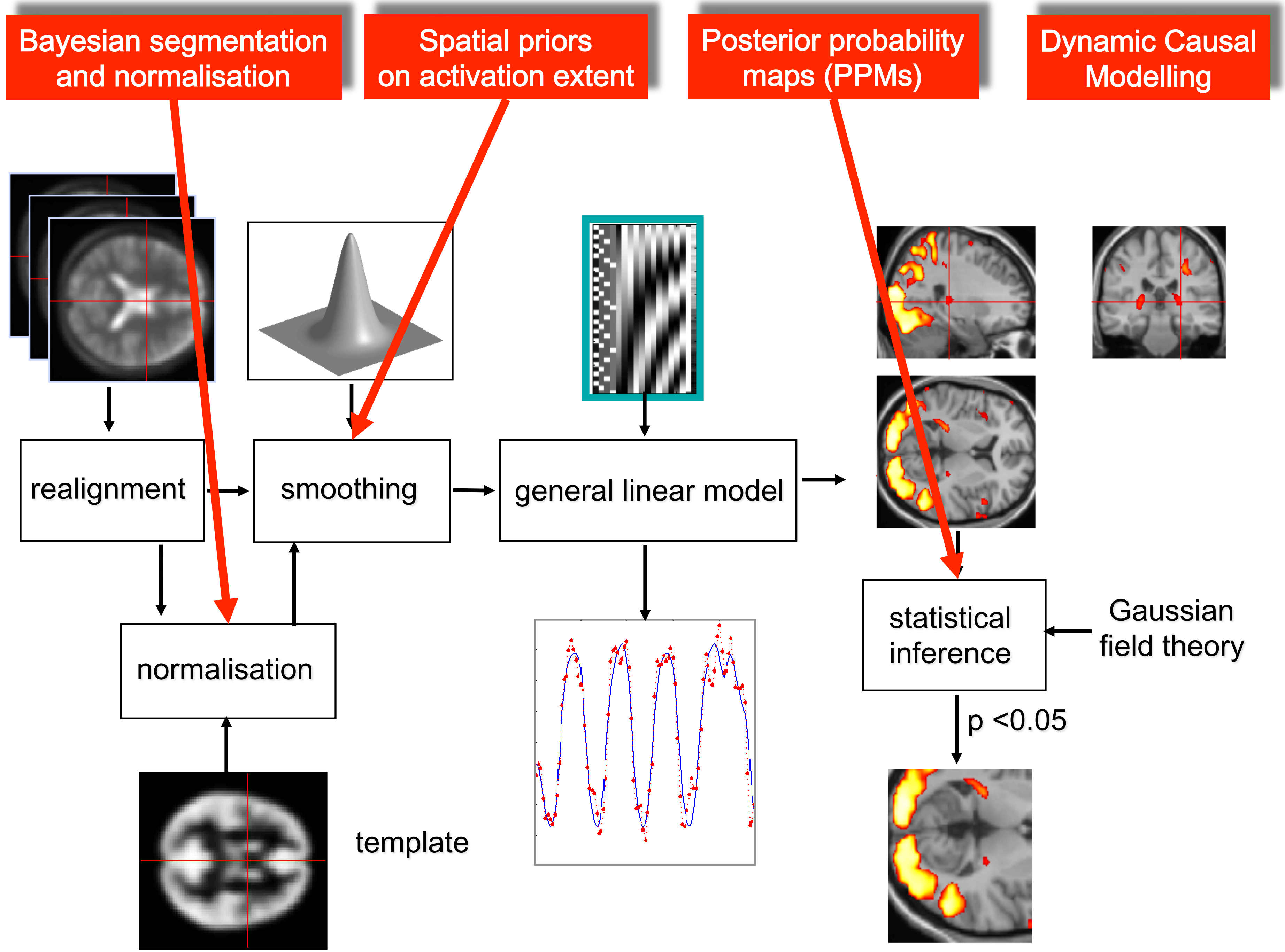
Bayesian Inference

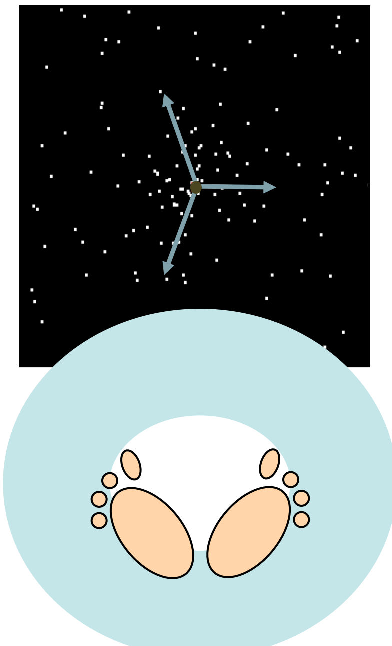
Thomas Nichols

With thanks
Lee Harrison

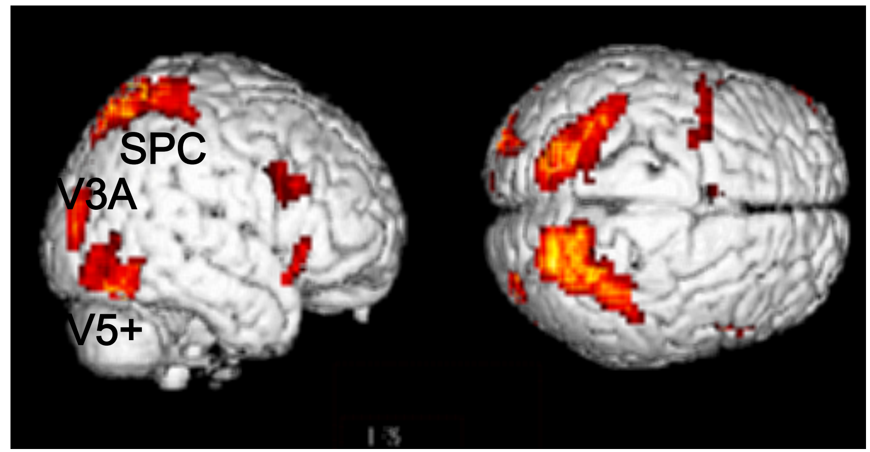

Bayesian segmentation and normalisation
Spatial priors on activation extent
Posterior probability maps (PPMs)
Dynamic Causal Modelling
realignment
smoothing
general linear model
normalisation
template
statistical inference
Gaussian field theory
p <0.05

Attention to Motion

Paradigm



Results



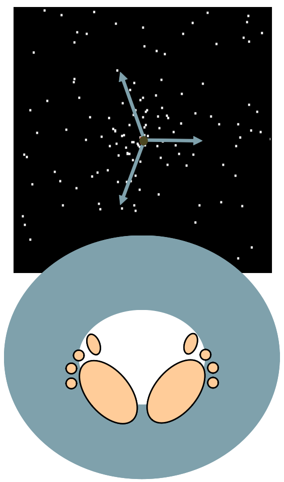
Attention – No attention

Büchel & Friston 1997, Cereb. Cortex
Büchel et al. 1998, Brain

- fixation only		
- observe static dots	+ photic	→ V1
- observe moving dots	+ motion	→ V5
- task on moving dots	+ attention	→ V5 + parietal cortex

Attention to Motion

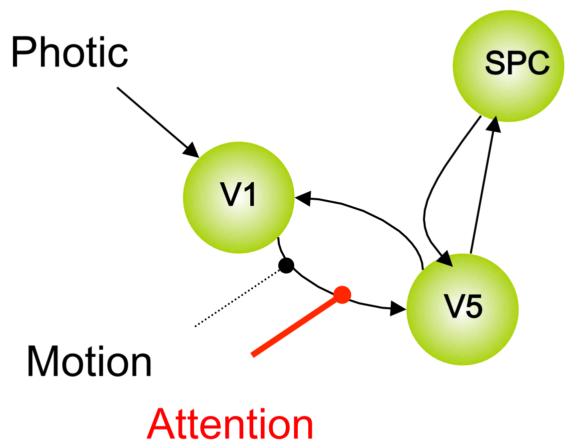
Paradigm



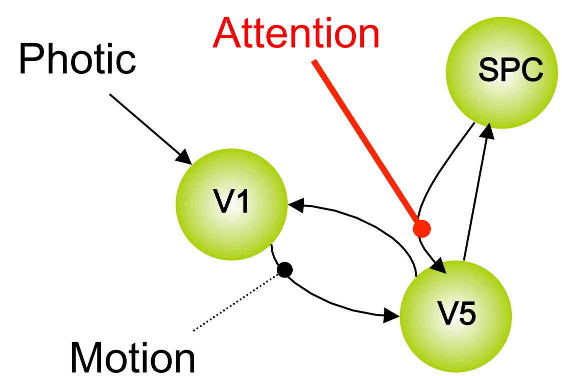
- fixation only
- observe static dots
- observe moving dots
- task on moving dots

Dynamic Causal Models

<u>Model 1 (forward):</u>
attentional modulation
of V1→V5: forward

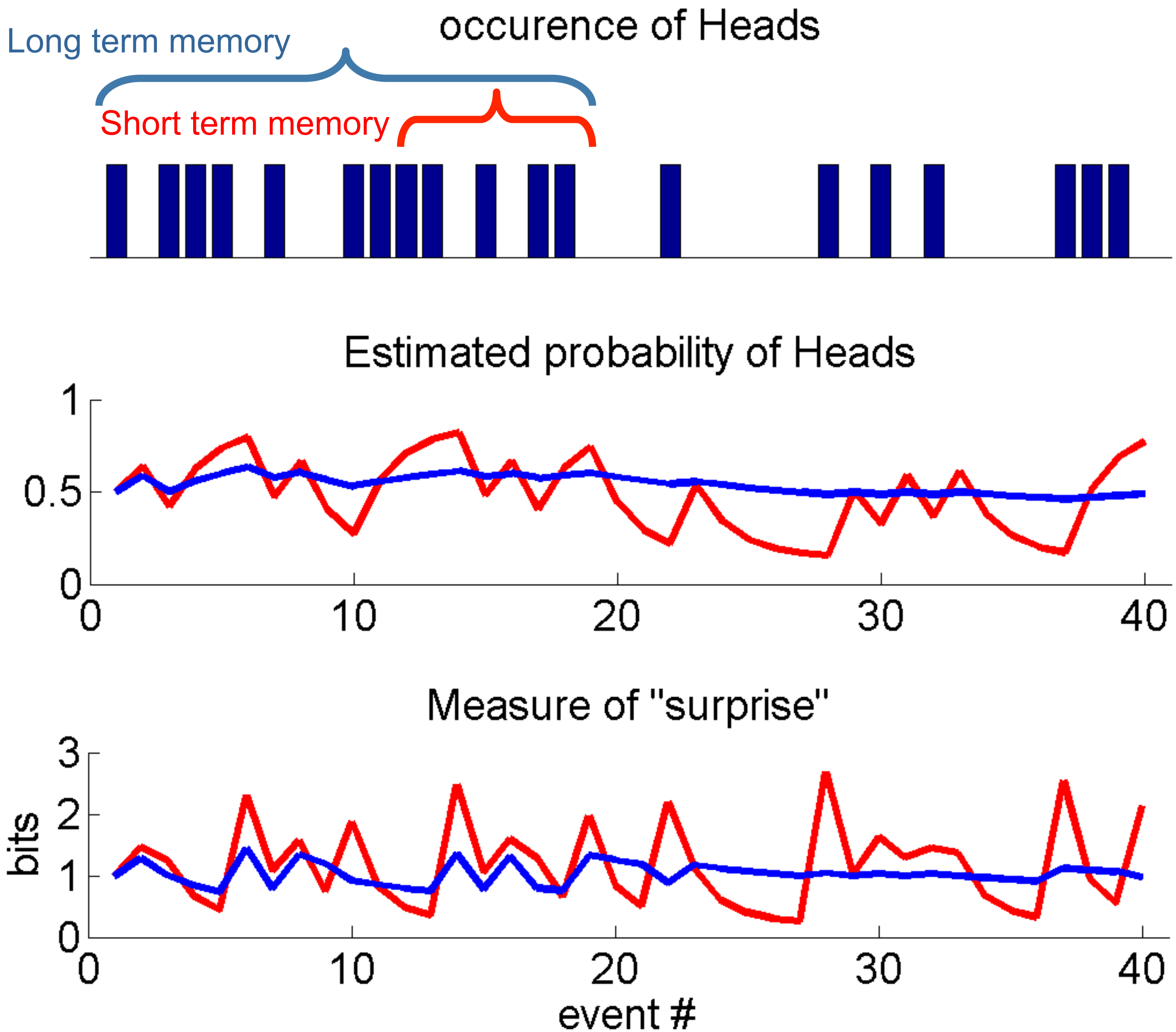


<u>Model 2 (backward):</u>
attentional modulation
of SPC→V5: backward



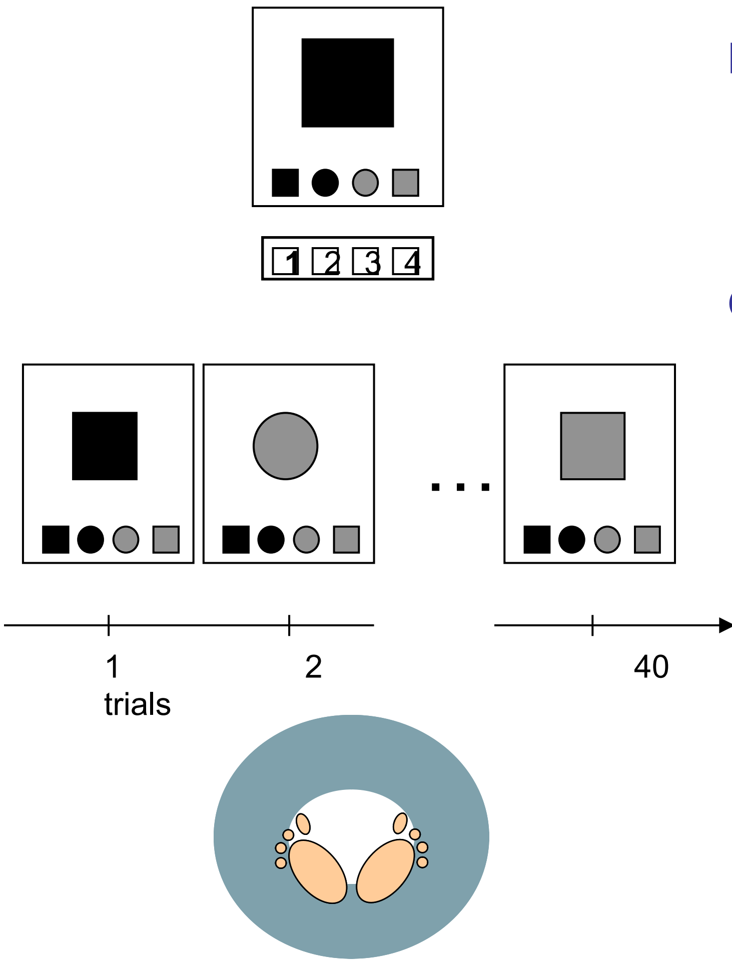
Bayesian model selection: Which model is optimal?

Responses to Uncertainty



Responses to Uncertainty

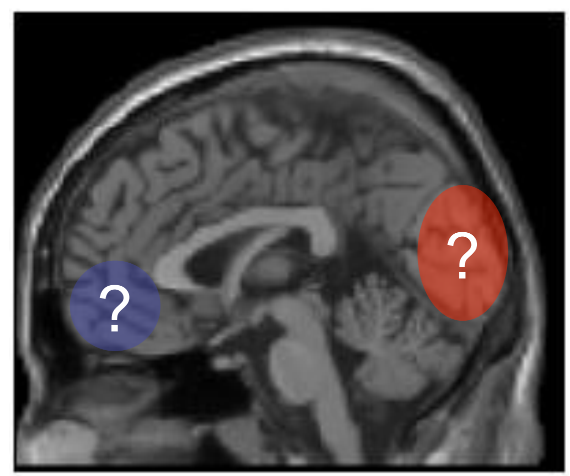
Paradigm



Stimuli sequence of randomly sampled discrete events

Model simple computational model of an observers response to uncertainty based on the number of past events (extent of memory)

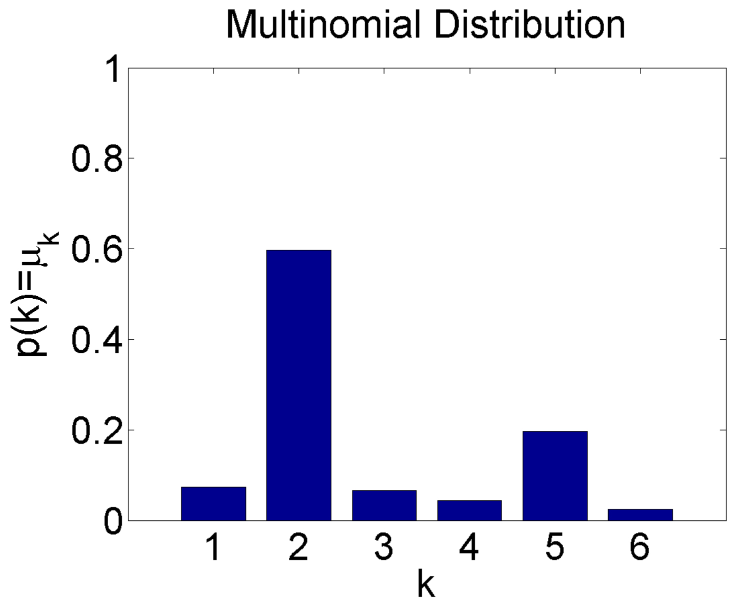
Question which regions are best explained by short / long term memory model?



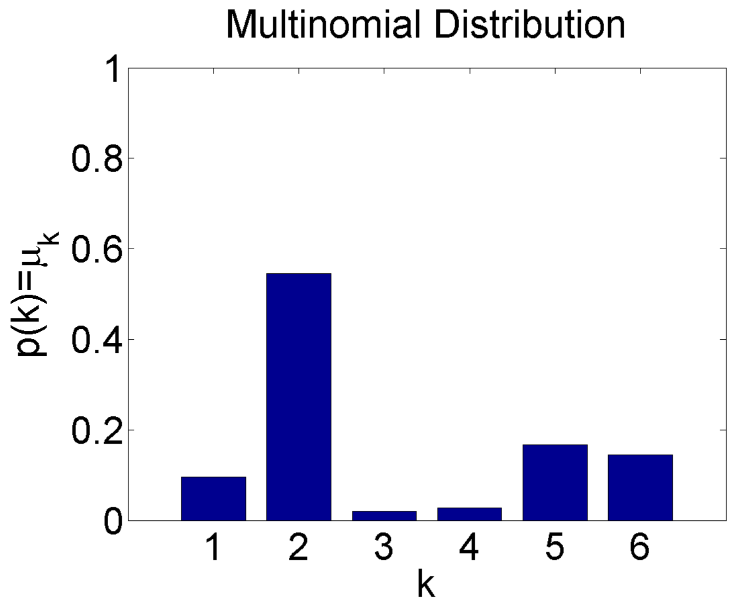
Overview

- Introductory remarks
- Some probability densities/distributions
- Probabilistic (generative) models
- Bayesian inference
- A simple example – Bayesian linear regression
- SPM applications
  - Segmentation
  - Dynamic causal modeling
  - Spatial models of fMRI time series

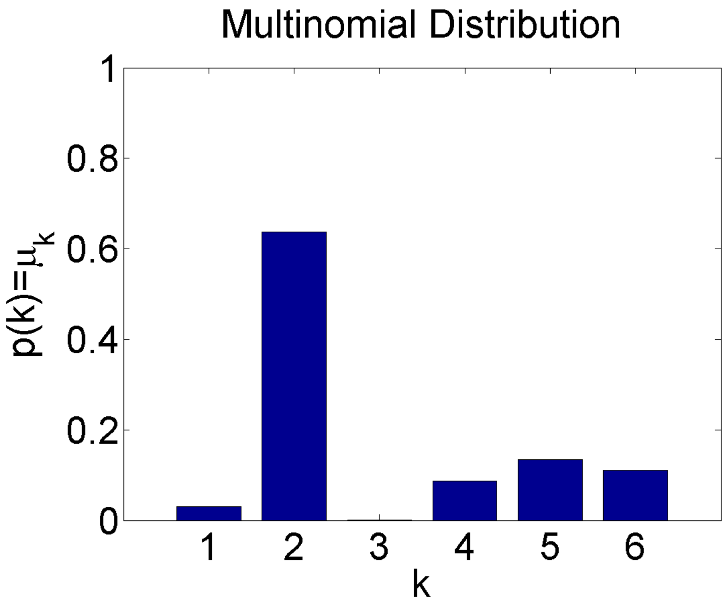
Probability distributions and densities



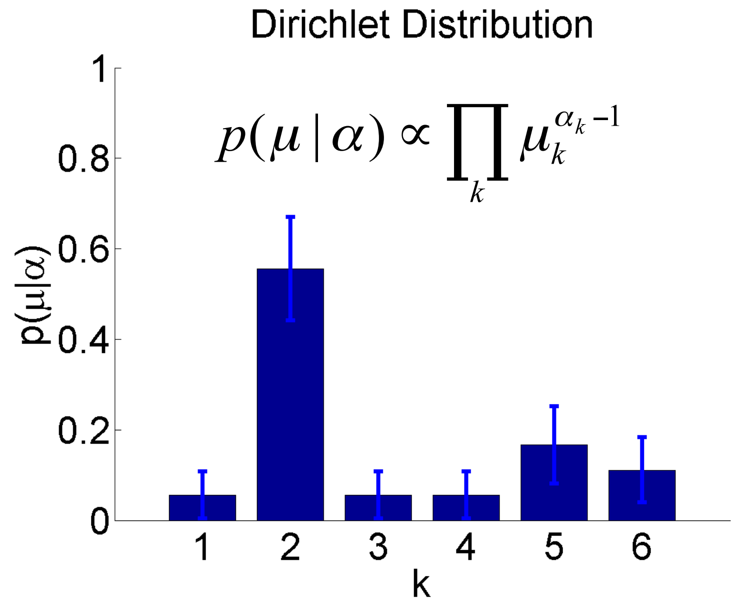
Probability distributions and densities

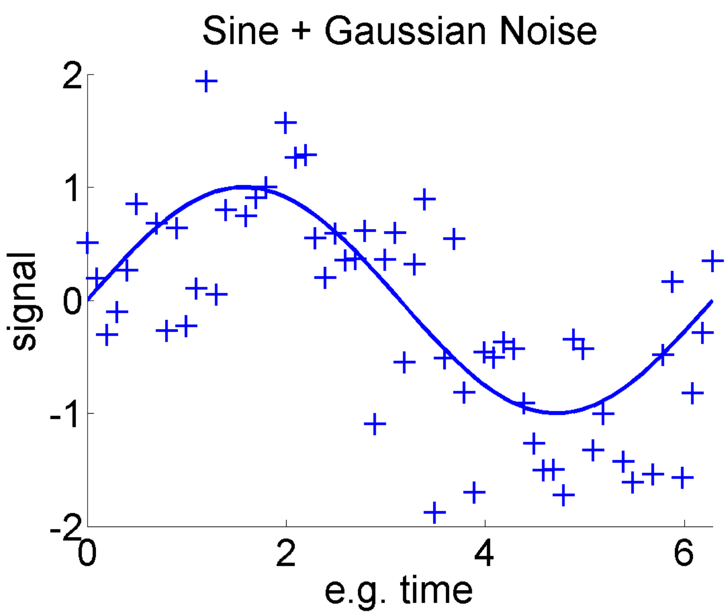


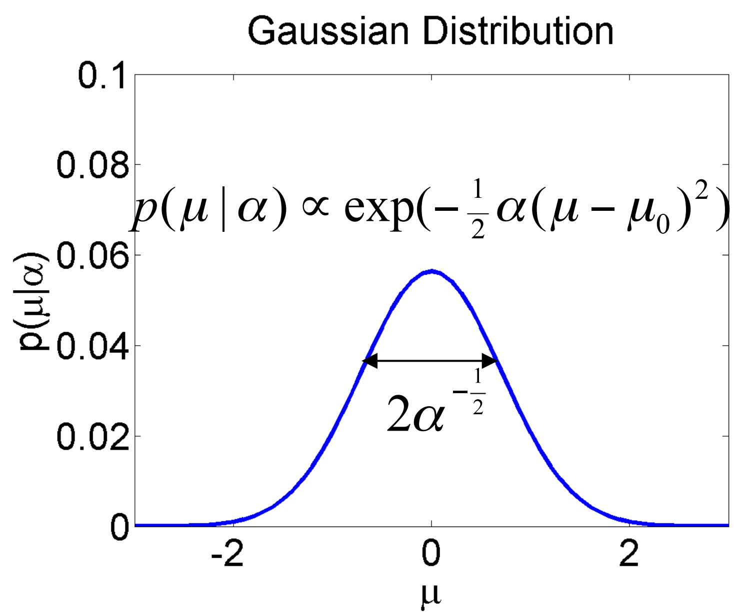
Probability distributions and densities



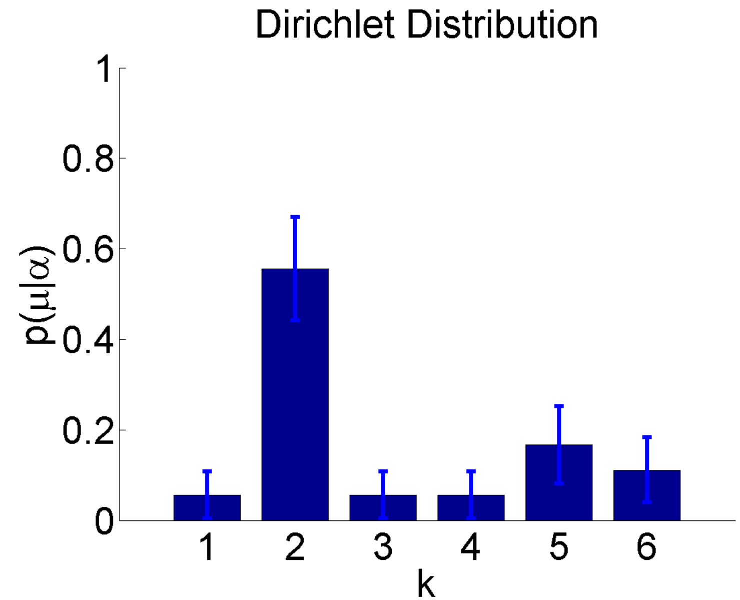
Probability distributions and densities

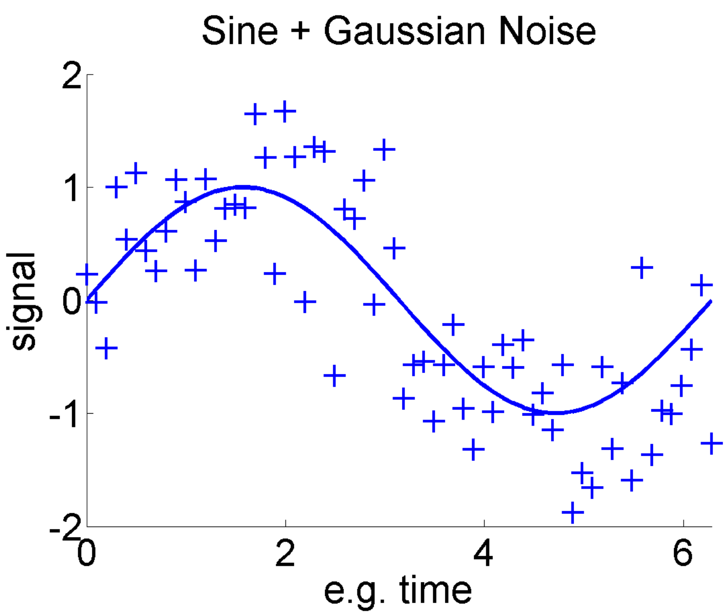


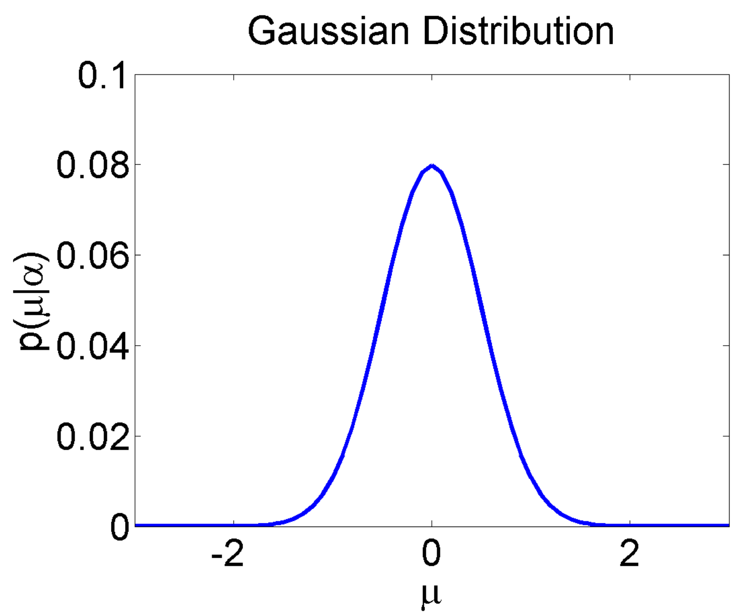




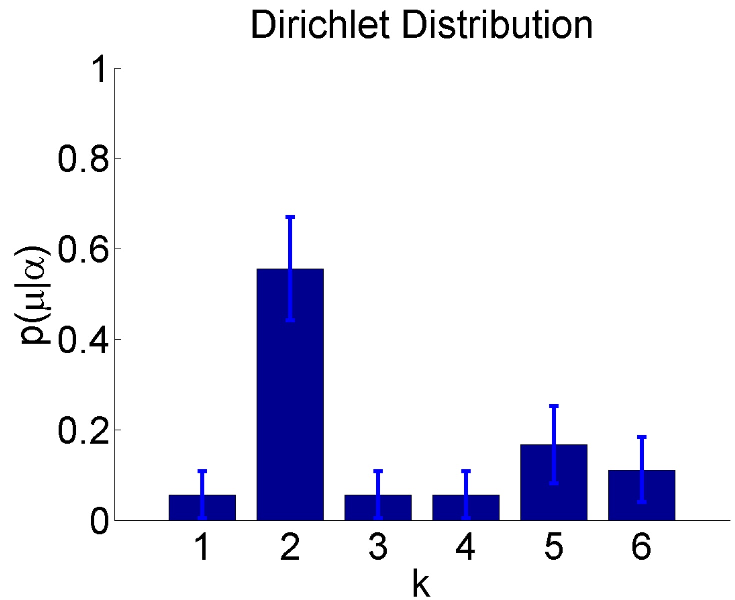
Probability distributions and densities

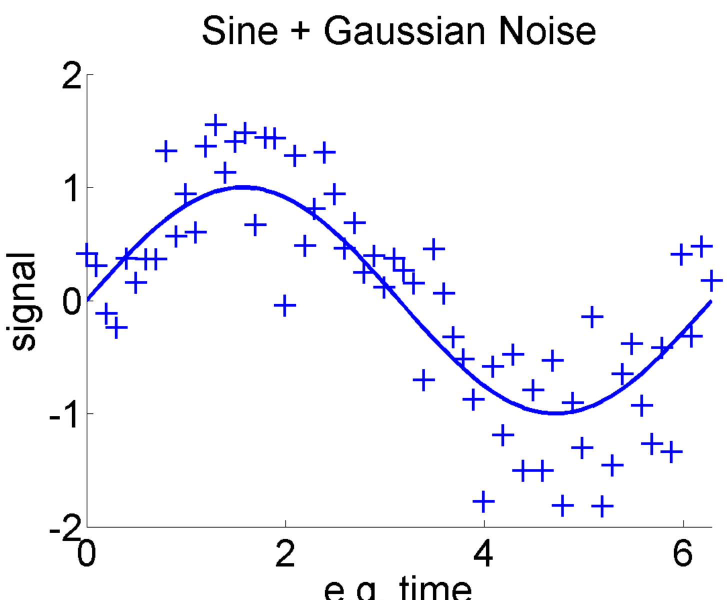


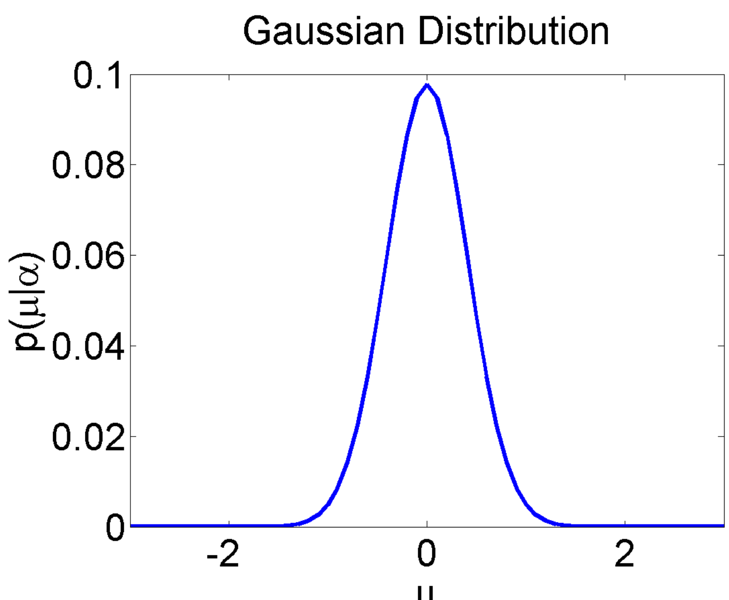




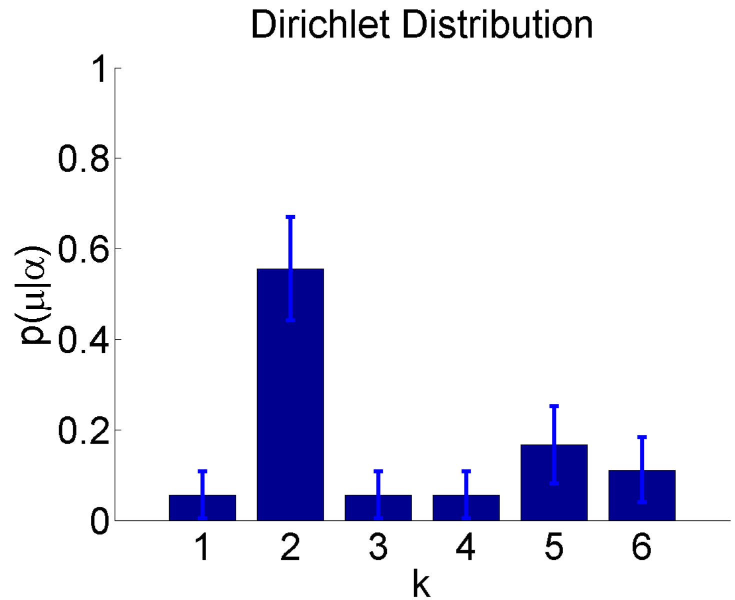
Probability distributions and densities

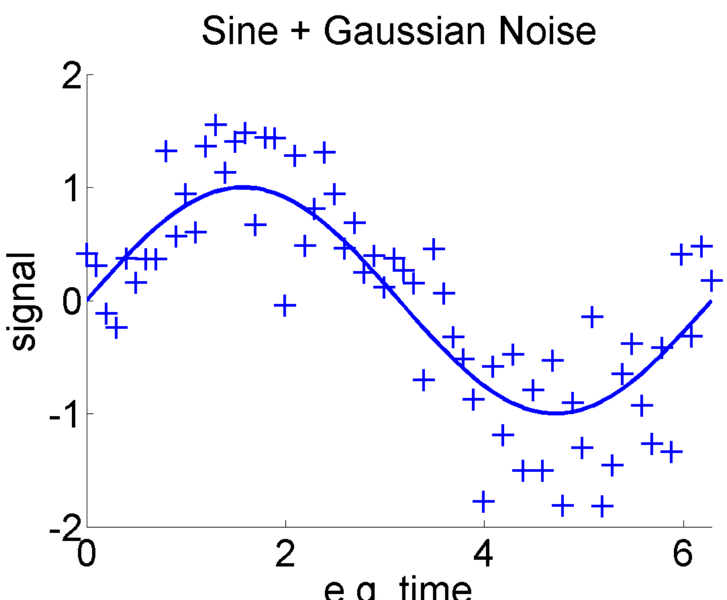


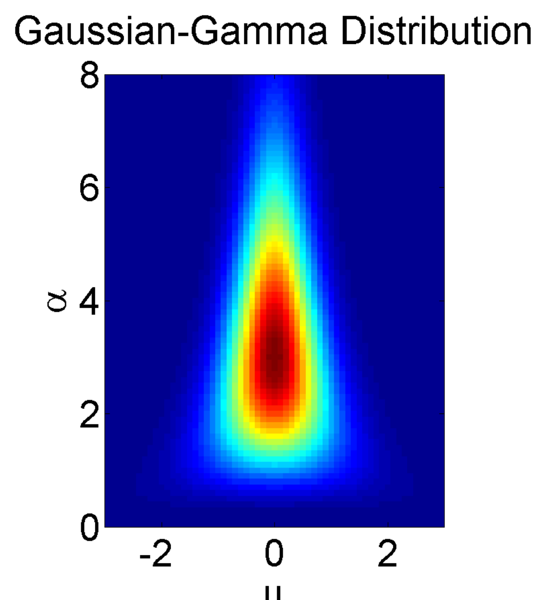




Probability distributions and densities





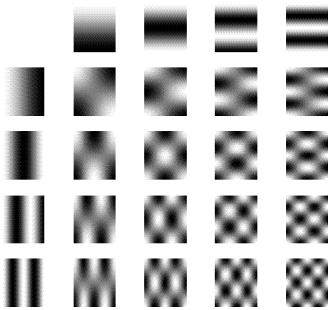


Generative models

$q(\theta)?$

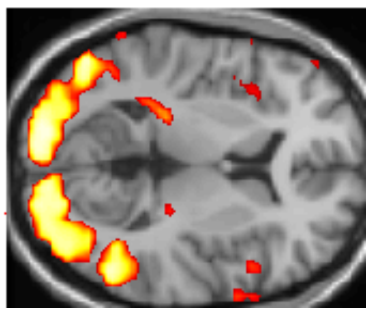
estimation

β_2



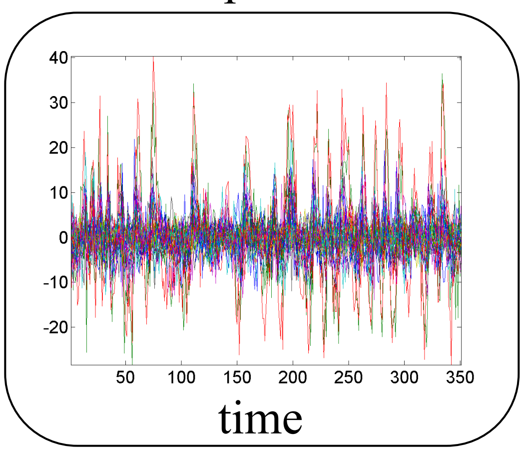
space

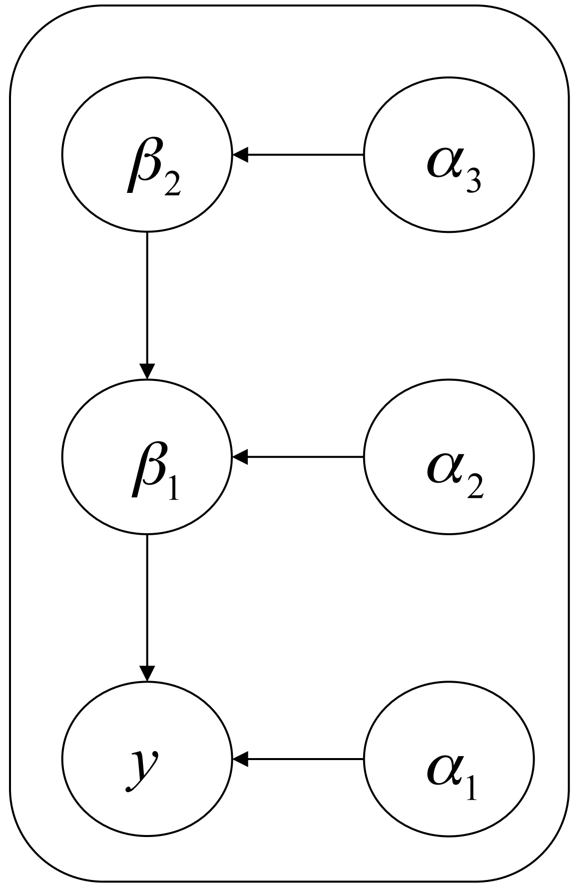
β_1



space

y





$$\beta_2 = e_3 \sim N(0, \alpha_3^{-1})$$

$$\beta_1 = X_2 \beta_2 + e_2$$

$$y = X_1 \beta_1 + e_1$$

generation

$$\theta = \{\beta, \alpha\}$$

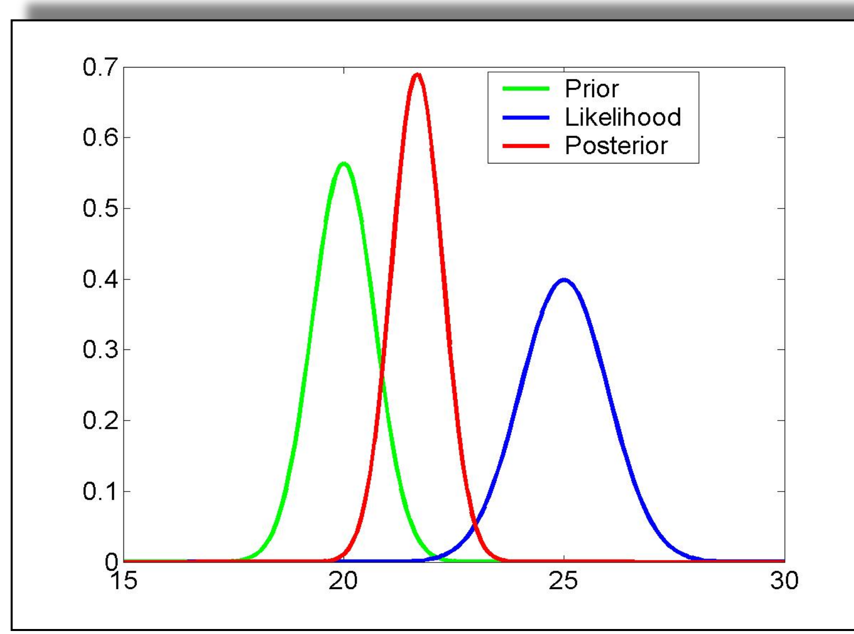
Bayesian statistics

new data $\quad$ prior knowledge

$$\boxed{p(y \mid \theta)} \qquad \boxed{p(\theta)}$$

$$\boxed{p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)}$$

posterior $\propto$ likelihood • prior



Bayes theorem allows one to formally incorporate prior knowledge into computing statistical probabilities.

The "posterior" probability of the parameters given the data is an optimal combination of prior knowledge and new data, weighted by their relative precision.

Bayes' rule

Given data *y* and parameters θ, their joint probability can be written in 2 ways:

$$p(\theta \mid y)\,p(y) = p(y,\theta) \qquad\qquad p(y,\theta) = p(y \mid \theta)\,p(\theta)$$

Eliminating $p(y,\theta)$ gives Bayes' rule:

Likelihood

Prior

Posterior

$$p(\theta \mid y) = \frac{p(y \mid \theta)\,p(\theta)}{p(y)}$$

Evidence

Principles of Bayesian inference

⇨ Formulation of a generative model

likelihood $p(y|\theta)$

prior distribution $p(\theta)$

⇨ Observation of data

y

⇨ Update of beliefs based upon observations, given a prior state of knowledge

$$p(\theta \mid y) \propto p(y \mid \theta) p(\theta)$$

Univariate Gaussian

Normal densities

$$p(\beta) = N(\beta; \mu_p, \alpha_p^{-1})$$

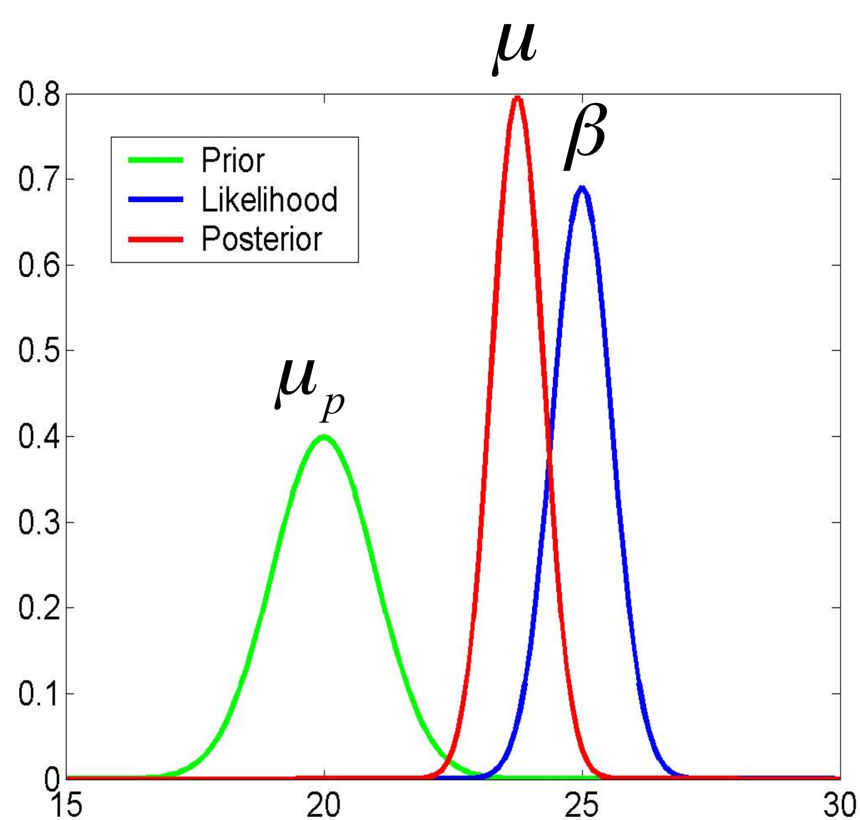
$$y = \beta + e$$

$$p(y \mid \beta) = N(y; \beta, \alpha_e^{-1})$$

$$p(\beta \mid y) = N(\beta; \mu, \alpha^{-1})$$

$$\alpha = \alpha_e + \alpha_p$$

$$\mu = \alpha^{-1}\left(\alpha_e y + \alpha_p \mu_p\right)$$



Posterior mean = precision-weighted combination of prior mean and data mean

Bayesian GLM: univariate case

Normal densities

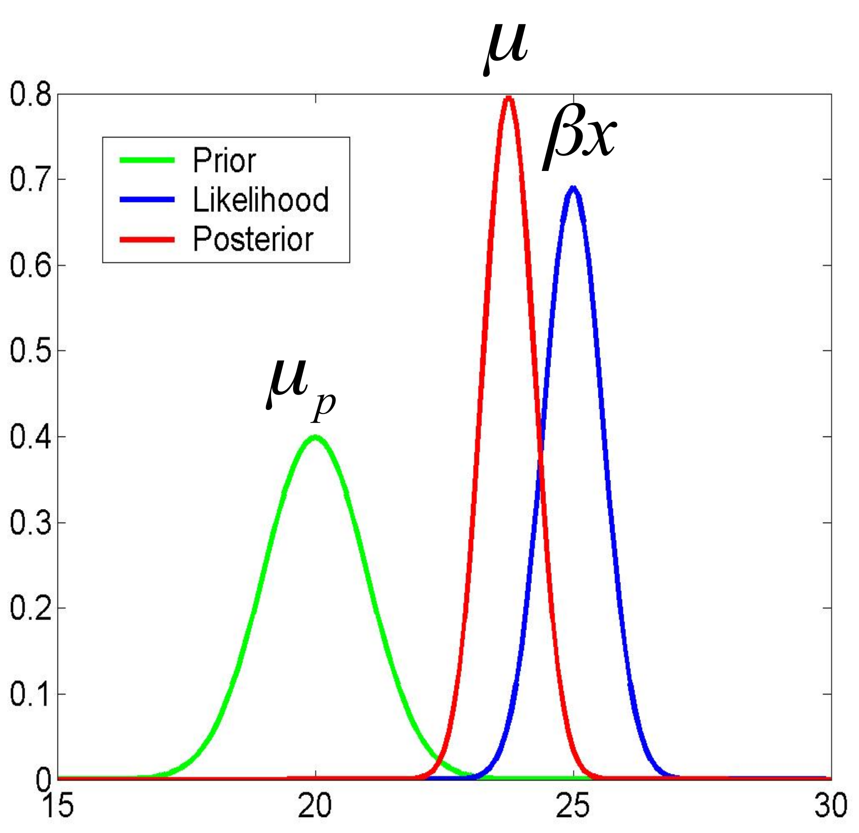
$$p(\beta) = N(\beta; \mu_p, \alpha_p^{-1})$$

$$p(y \mid \beta) = N(y; \beta x, \alpha_e^{-1})$$

$$p(\beta \mid y) = N(\beta; \mu, \alpha^{-1})$$

$$\alpha = \alpha_e x^2 + \alpha_p$$
$$\mu = \alpha^{-1}(\alpha_e xy + \alpha_p \mu_p)$$

$$y = \beta x + e$$



Bayesian GLM: multivariate case

Normal densities

$$p(\beta) = N(\beta; \mu_p, C_p)$$

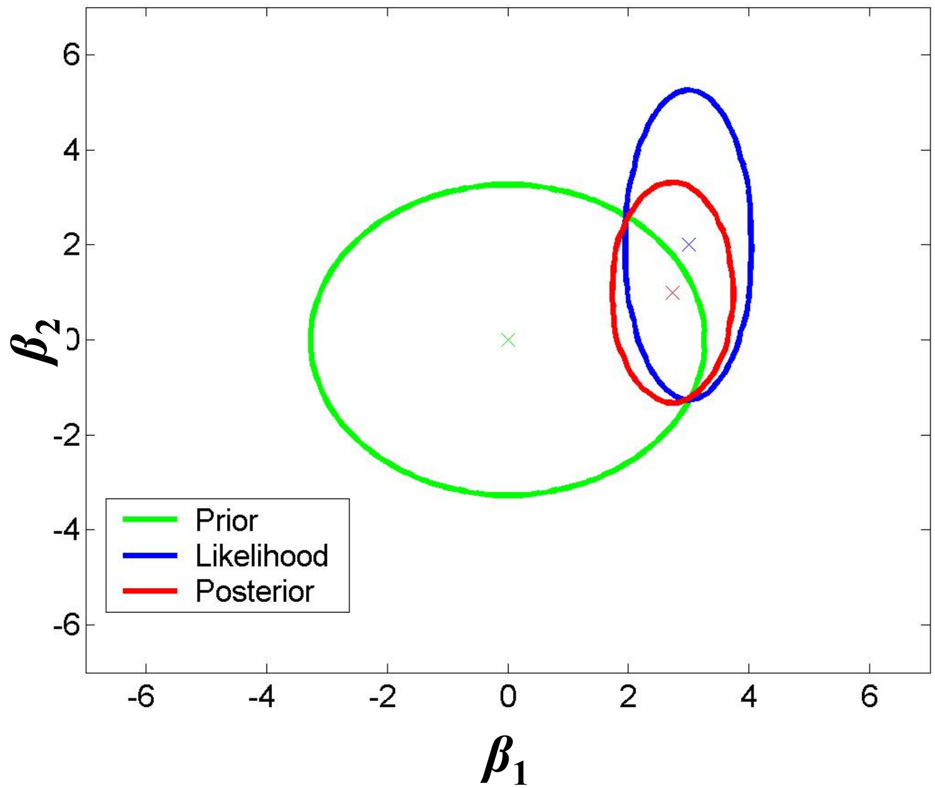
$$p(y \mid \beta) = N(y; X\beta, C_e)$$

$$p(\beta \mid y) = N(\beta; \mu, C)$$

$$C^{-1} = X^T C_e^{-1} X + C_p^{-1}$$

$$\mu = C\left(X^T C_e y + C_p^{-1} \mu_p\right)$$

$$y = X\beta + e$$

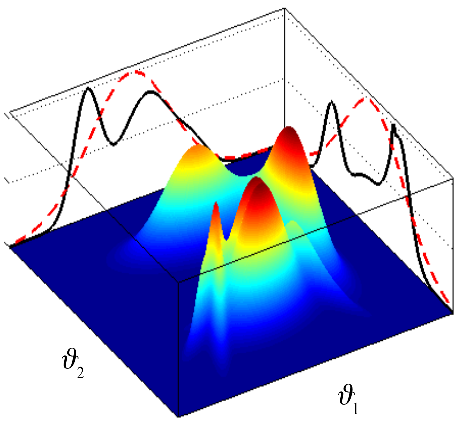


One step if $\mathbf{C}_e$ and $\mathbf{C}_p$ are known.
Otherwise iterative estimation.

Approximate inference: optimization

True posterior $$p(\theta \mid y, m) = \frac{p(y, \theta \mid m)}{p(y|m)}$$

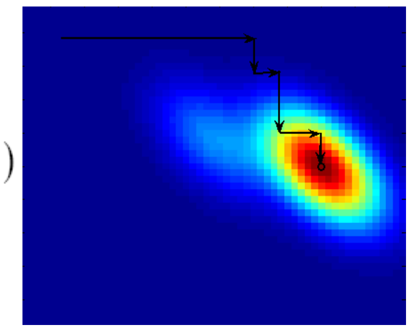
mean-field approximation



Approximate posterior $$q(\theta) = \prod_i q(\theta_i)$$

iteratively improve

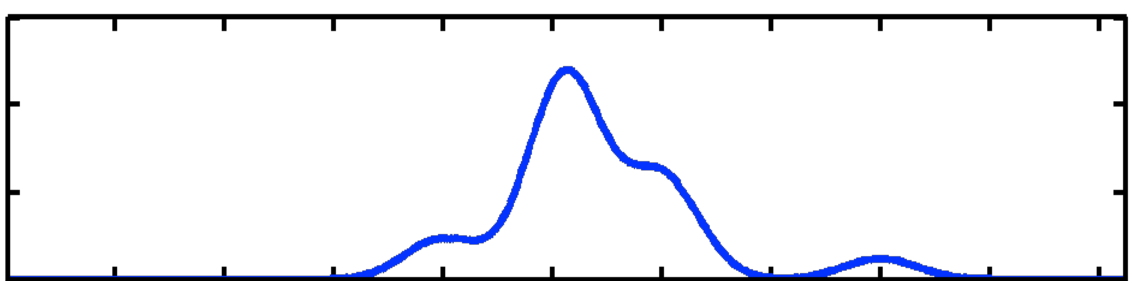
$q(\theta_1)$



$q(\theta_2)$

$$\log p(y|m) = \overbrace{\int_\theta q(\theta) \log\left(\frac{p(y, \theta \mid m)}{q(\theta)}\right)}^{\text{free energy}} + \int_\theta q(\theta) \log\left(\frac{q(\theta)}{p(\theta \mid y, m)}\right)$$

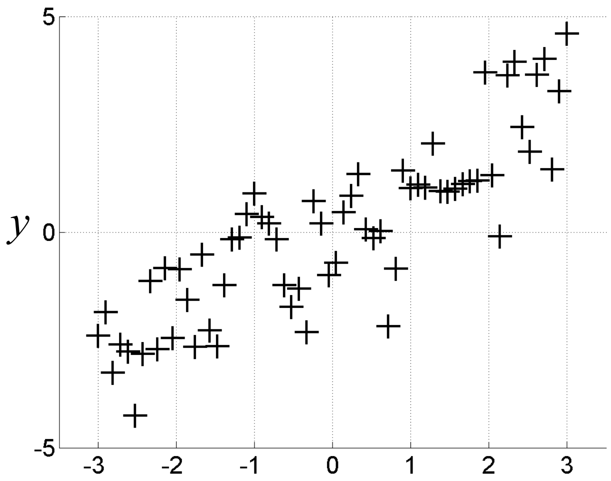
Objective function



Value of parameter

Simple example – linear regression

Data



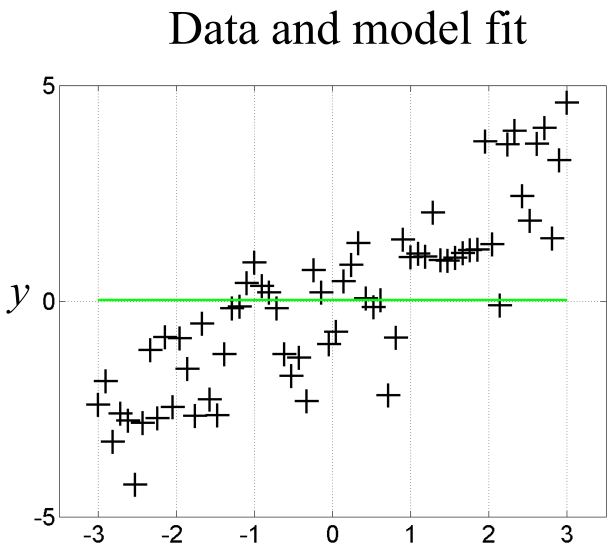
Ordinary least squares

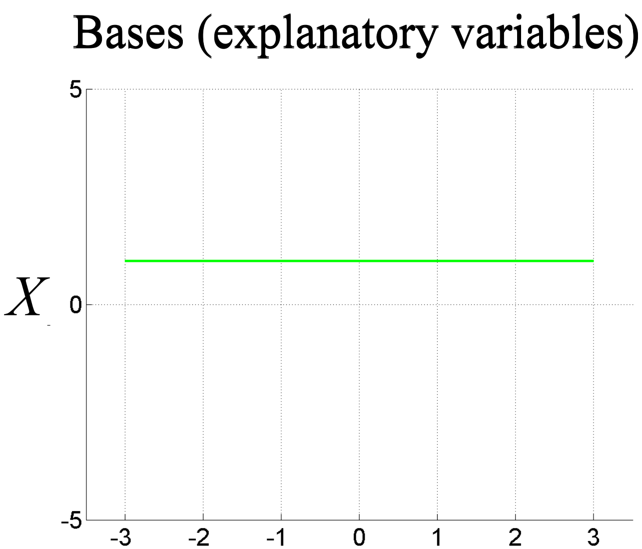
$$y = X\beta$$

$$E_D = (y - X\beta)^T (y - X\beta)$$

$$\frac{\partial E_D}{\partial \beta} = 0 \Rightarrow \hat{\beta}_{ols} = (X^T X)^{-1} X^T y$$

Simple example – linear regression



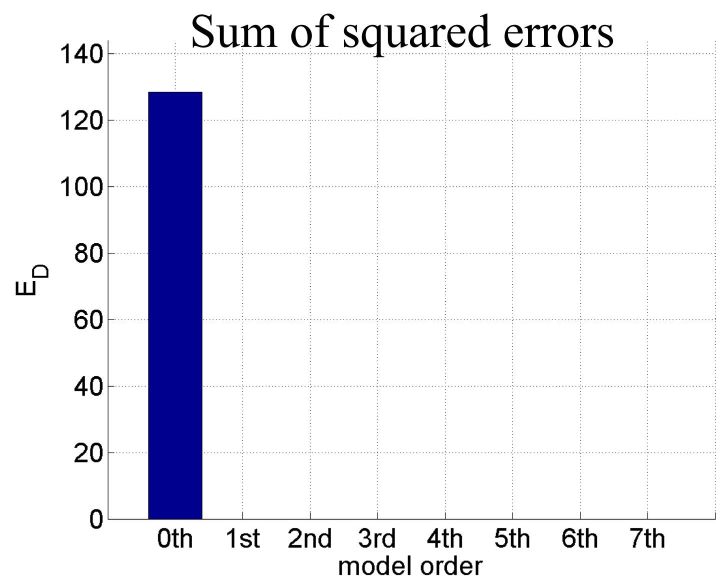


Ordinary least squares

$$y = X\beta$$

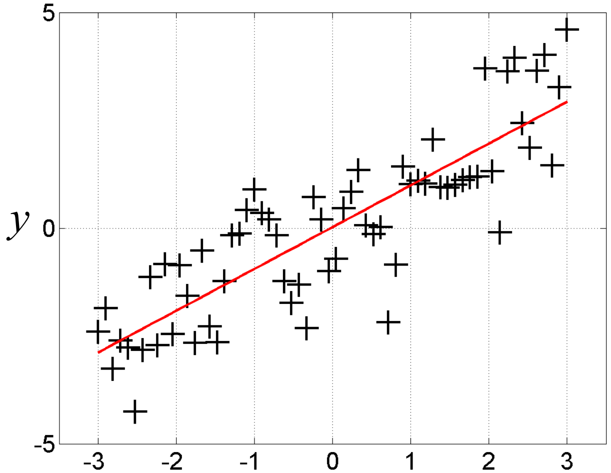
$$E_D = (y - X\beta)^T (y - X\beta)$$

$$\frac{\partial E_D}{\partial \beta} = 0 \Rightarrow \hat{\beta}_{ols} = (X^T X)^{-1} X^T y$$



Simple example – linear regression

Data and model fit



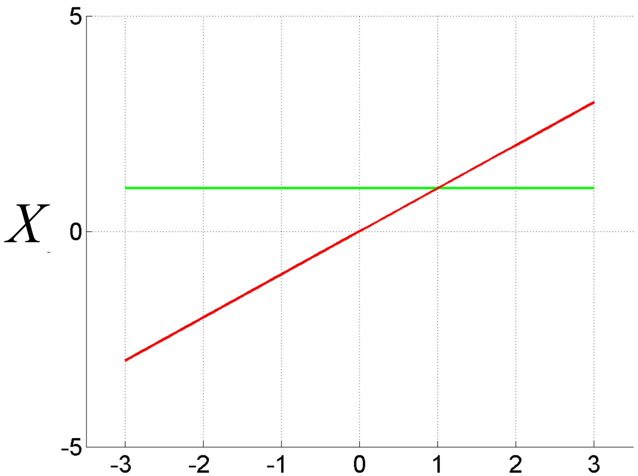
Ordinary least squares

$$y = X\beta$$

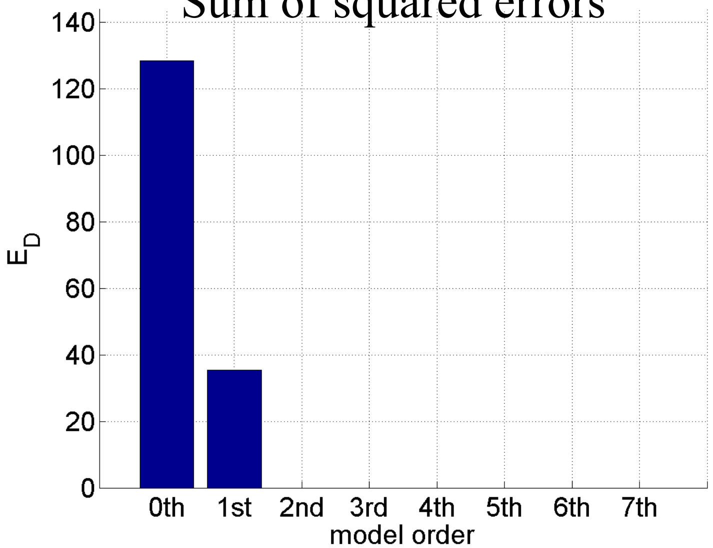
$$E_D = (y - X\beta)^T (y - X\beta)$$

$$\frac{\partial E_D}{\partial \beta} = 0 \Rightarrow \hat{\beta}_{ols} = (X^T X)^{-1} X^T y$$

Bases (explanatory variables)

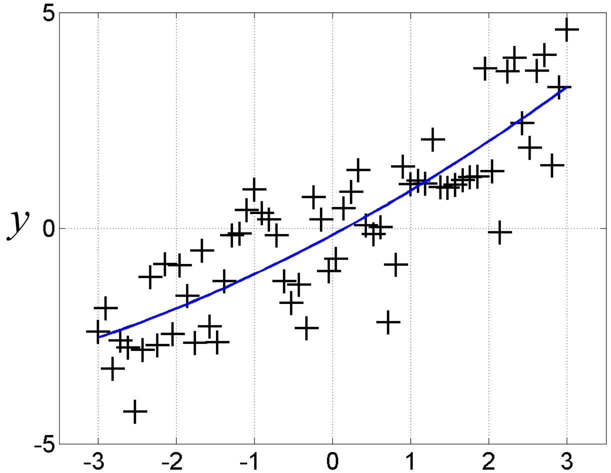


Sum of squared errors



Simple example – linear regression

Data and model fit



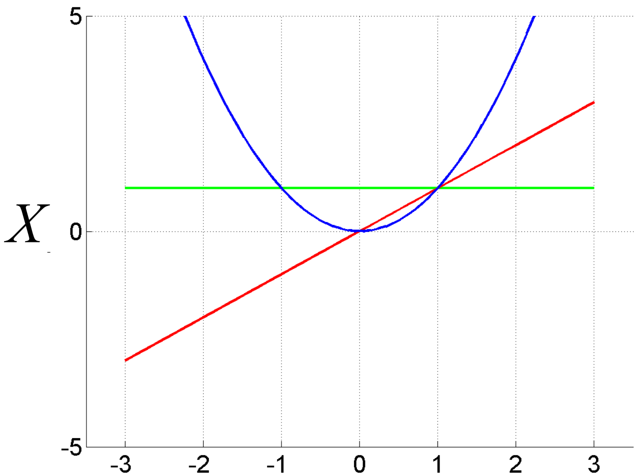
Ordinary least squares

$$y = X\beta$$

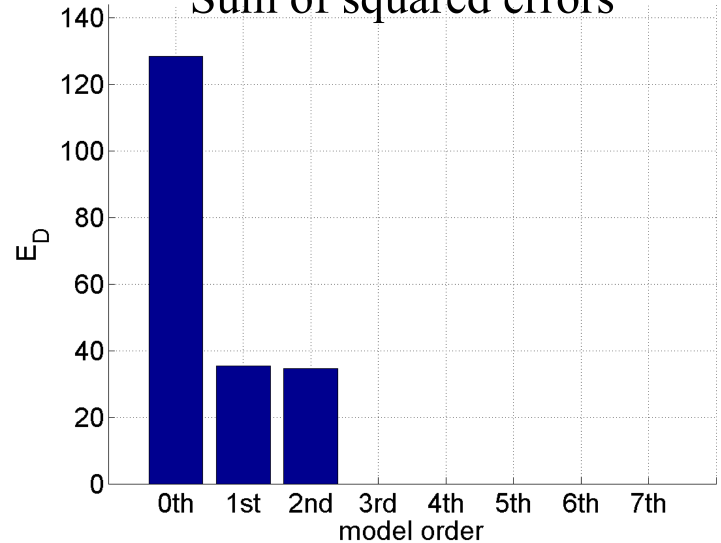
$$E_D = (y - X\beta)^T (y - X\beta)$$

$$\frac{\partial E_D}{\partial \beta} = 0 \Rightarrow \hat{\beta}_{ols} = (X^T X)^{-1} X^T y$$

Bases (explanatory variables)

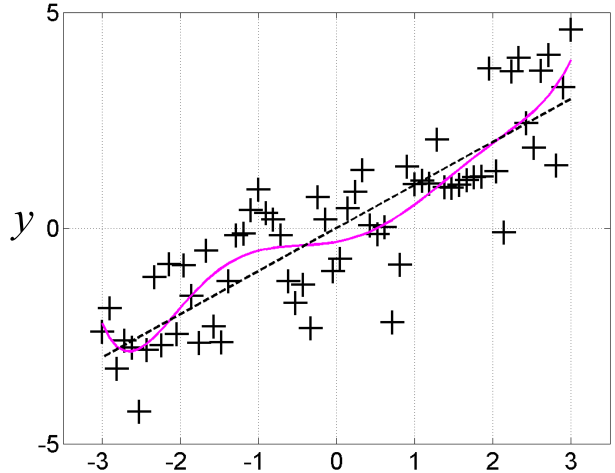


Sum of squared errors



Simple example – linear regression

Data and model fit



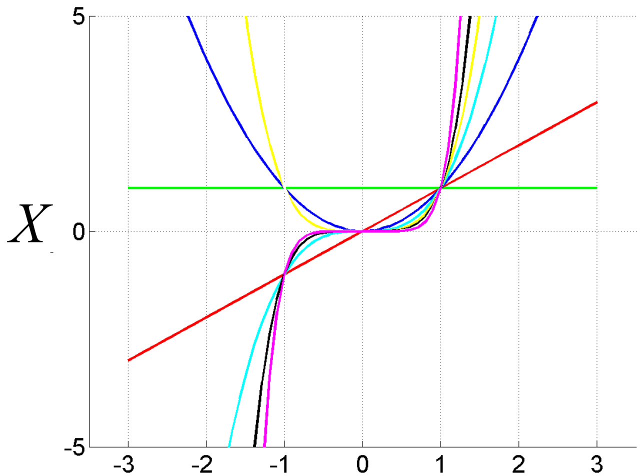
Ordinary least squares

Over-fitting: model fits noise

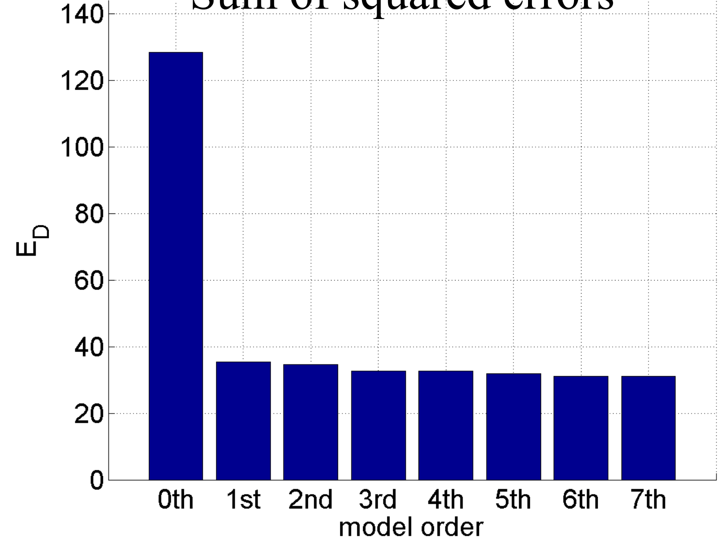
Inadequate cost function: blind to overly complex models

Solution: include uncertainty in model parameters

Bases (explanatory variables)



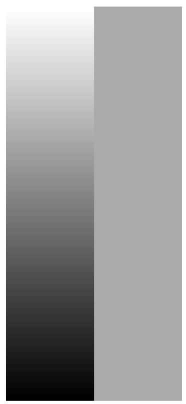
Sum of squared errors



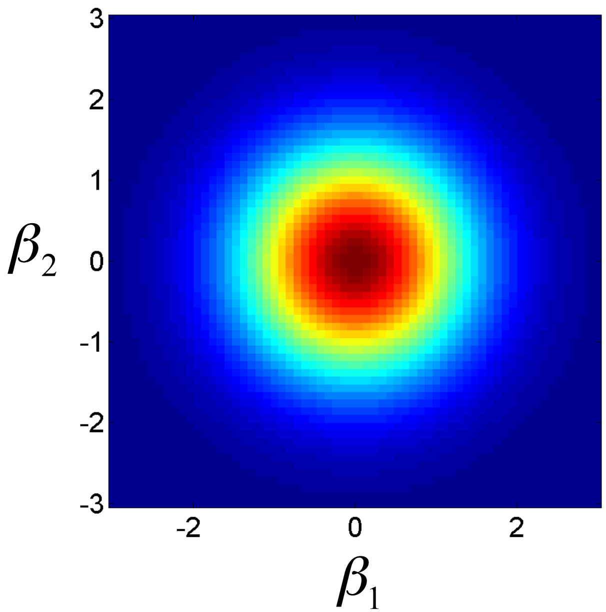
Bayesian linear regression:
priors and likelihood

Model: $y = X\beta + e$

$X =$



Bayesian linear regression: *priors and likelihood*

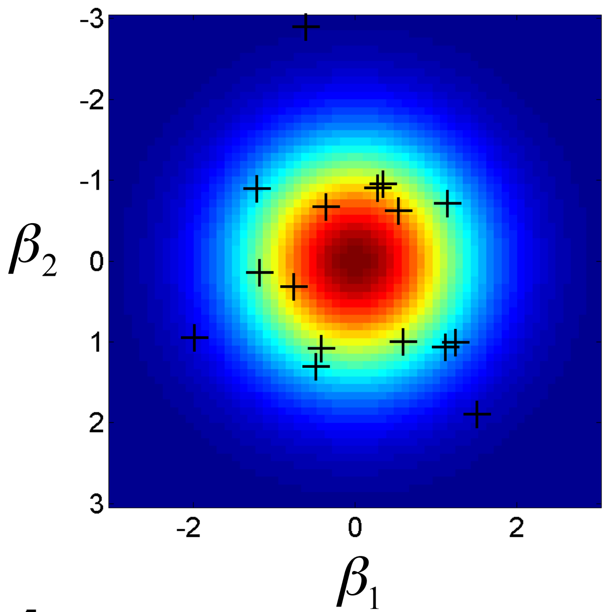


Model: $y = X\beta + e$

Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$

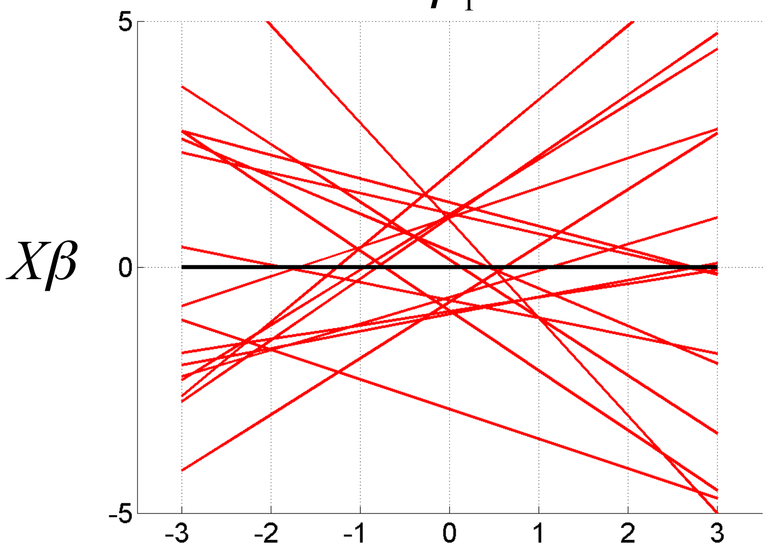
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

Bayesian linear regression: *priors and likelihood*



Model: $$y = X\beta + e$$

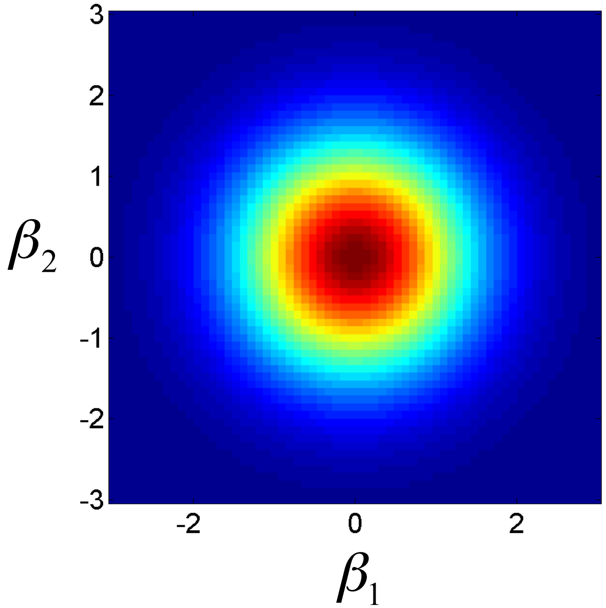
Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$



Sample curves from prior (before observing any data)

Mean curve

Bayesian linear regression:
priors and likelihood

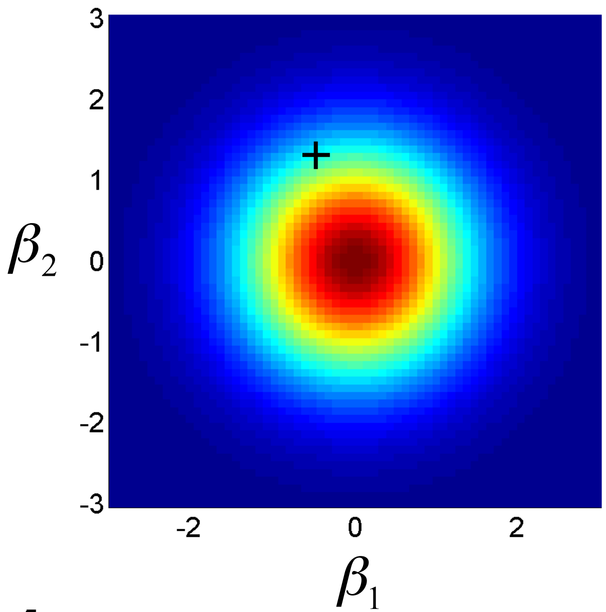


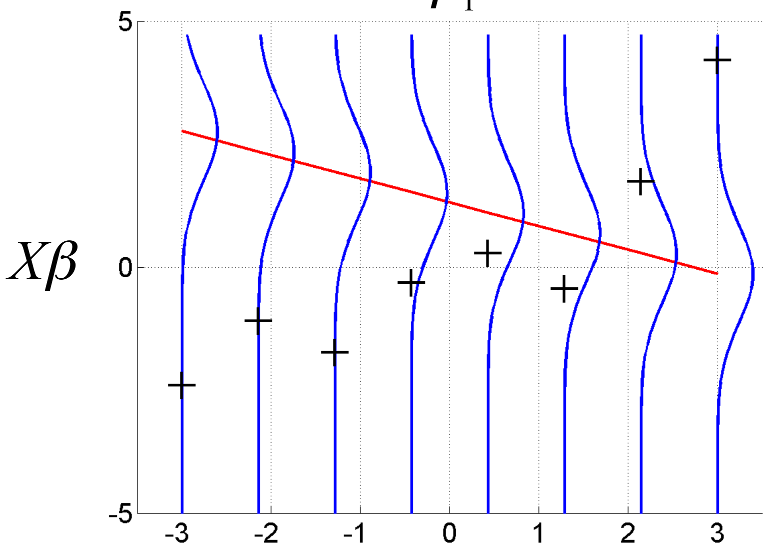
Model: $$y = X\beta + e$$

Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

Likelihood:
$$p(y|\beta, \alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1^{-1})$$
$$p(y_i|\beta, \alpha_1) = N(X_i\beta, \alpha_1^{-1})$$
$$\propto \exp(-\alpha_1 (y_i - X_i\beta)^2 / 2)$$

Bayesian linear regression: *priors and likelihood*



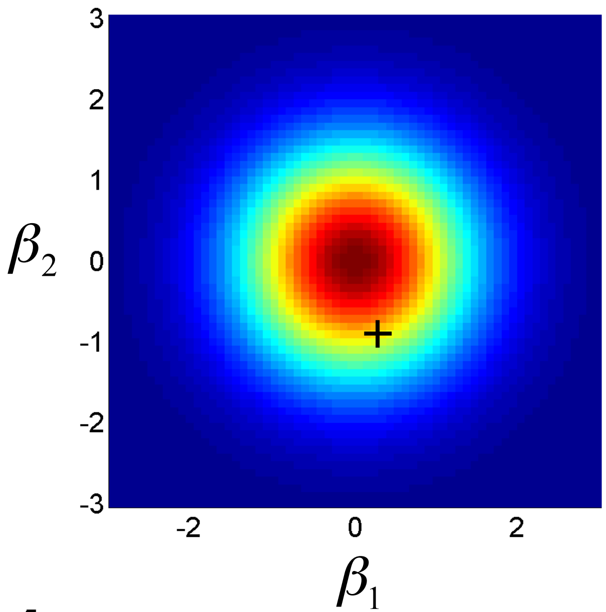


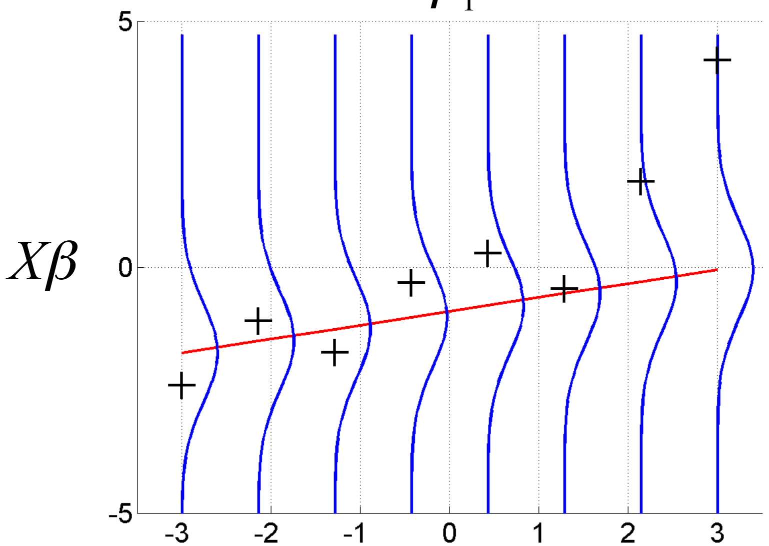
Model: $$y = X\beta + e$$

Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

Likelihood:
$$p(y|\beta, \alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1^{-1})$$
$$p(y_i|\beta, \alpha_1) = N(X_i \beta, \alpha_1^{-1})$$
$$\propto \exp(-\alpha_1 (y_i - X_i \beta)^2 / 2)$$

Bayesian linear regression: *priors and likelihood*



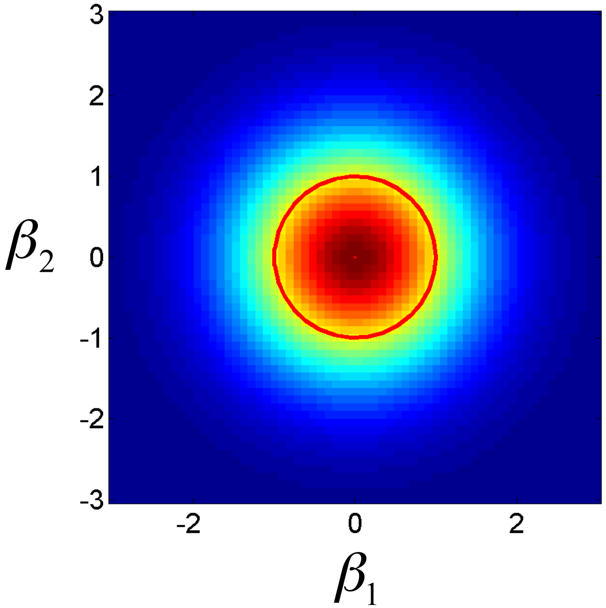


Model: $y = X\beta + e$

Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

Likelihood:
$$p(y|\beta, \alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1^{-1})$$
$$p(y_i|\beta, \alpha_1) = N(X_i\beta, \alpha_1^{-1})$$
$$\propto \exp(-\alpha_1 (y_i - X_i\beta)^2 / 2)$$

Bayesian linear regression: *posterior*



Model: $y = X\beta + e$

Prior: $p(\beta | \alpha_2) = N_k(0, \alpha_2^{-1} I_k)$

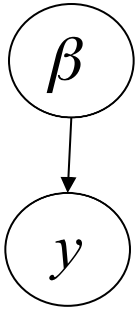
$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$

Likelihood:

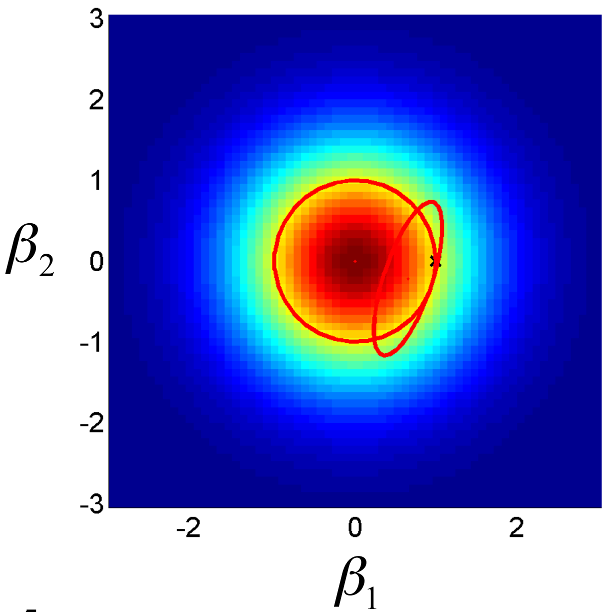
$$p(y | \beta, \alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1)$$

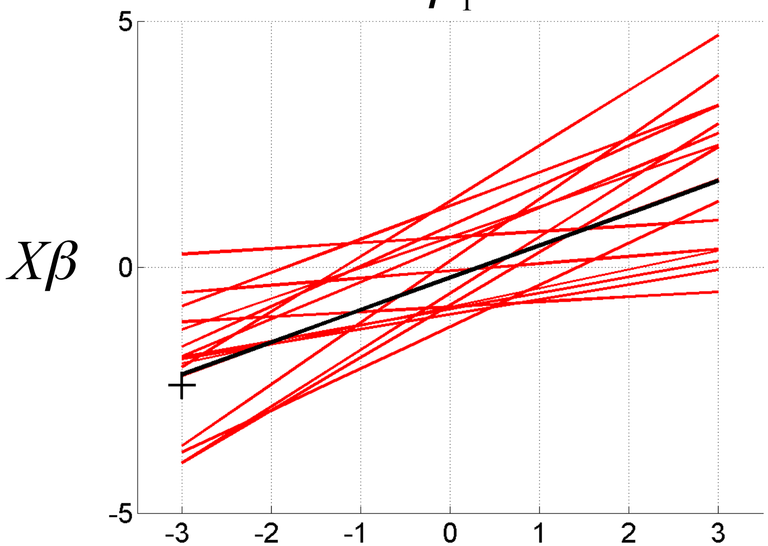
Bayes Rule:

$$p(\beta | y, \alpha) \propto p(y \mid \beta, \alpha) p(\beta \mid \alpha)$$



Bayesian linear regression: *posterior*





Model:

$$y = X\beta + e$$

Prior:

$$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$

$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

Likelihood:

$$p(y|\beta, \alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1)$$

Bayes Rule:

$$p(\beta|y, \alpha) \propto p(y \mid \beta, \alpha) p(\beta \mid \alpha)$$

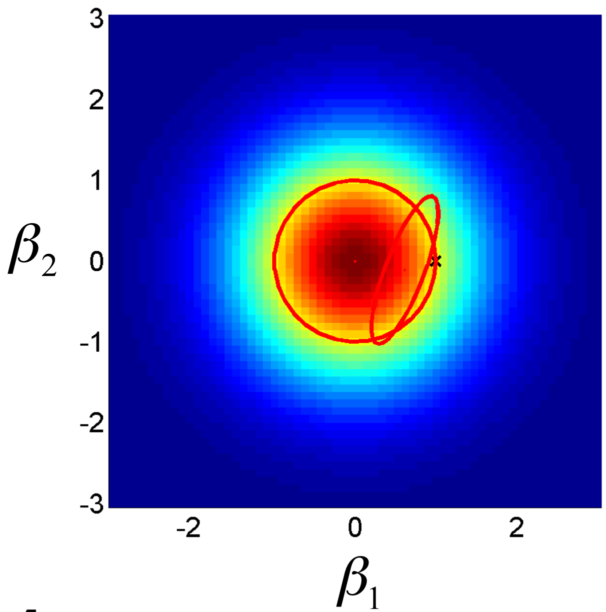
Posterior:

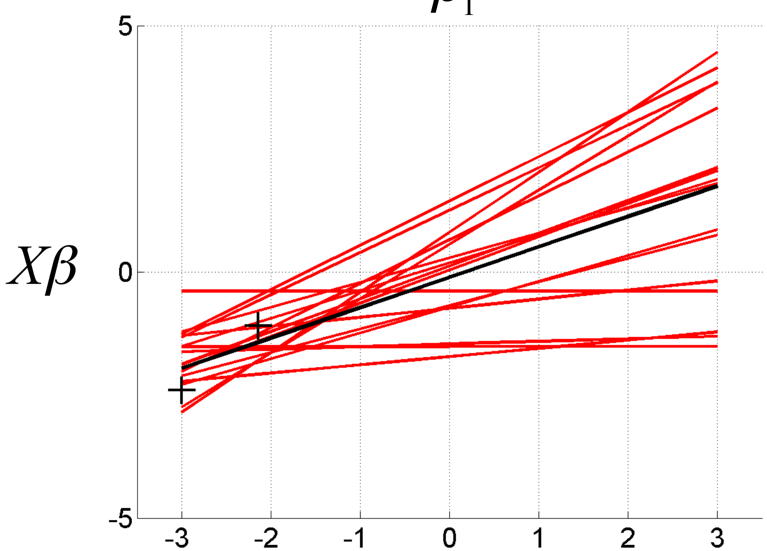
$$p(\beta \mid y, \alpha) = N(\mu, C)$$

$$C = \left(\alpha_1 X^T X + \alpha_2 I_k\right)^{-1}$$

$$\mu = \alpha_1 C X^T y$$

Bayesian linear regression: *posterior*





Model: $$y = X\beta + e$$

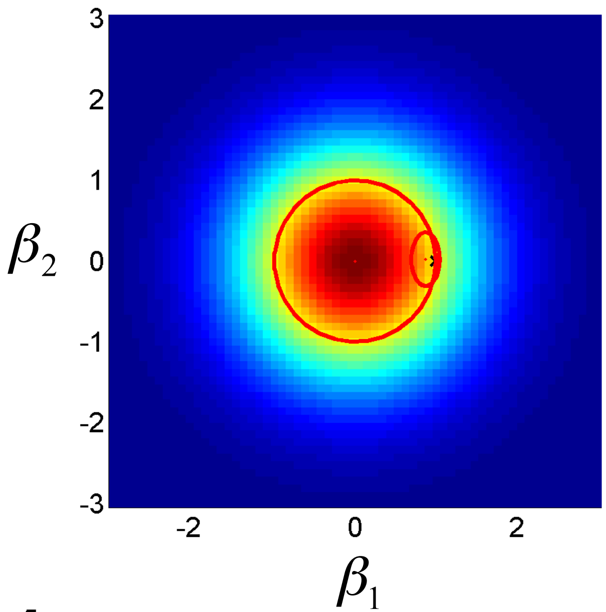
Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

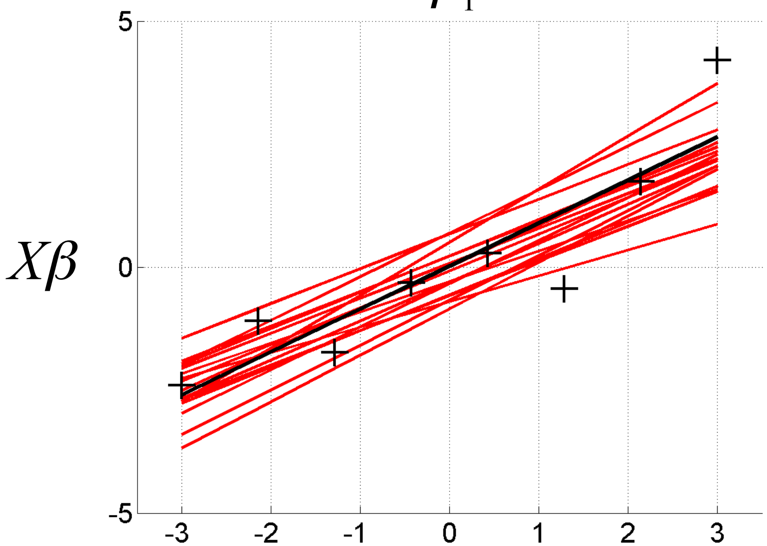
Likelihood:
$$p(y|\beta,\alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1)$$

Bayes Rule:
$$p(\beta|y,\alpha) \propto p(y \mid \beta, \alpha) p(\beta \mid \alpha)$$

Posterior: $$p(\beta \mid y, \alpha) = N(\mu, C)$$
$$C = \left(\alpha_1 X^T X + \alpha_2 I_k\right)^{-1}$$
$$\mu = \alpha_1 C X^T y$$

Bayesian linear regression: *posterior*





Model: $$y = X\beta + e$$

Prior: $$p(\beta|\alpha_2) = N_k(0, \alpha_2^{-1} I_k)$$
$$\propto \exp(-\alpha_2 \|\beta\|^2 / 2)$$

Likelihood:
$$p(y|\beta, \alpha_1) = \prod_{i=1}^{N} p(y_i \mid \beta, \alpha_1)$$

Bayes Rule:
$$p(\beta|y, \alpha) \propto p(y \mid \beta, \alpha) p(\beta \mid \alpha)$$

Posterior: $$p(\beta \mid y, \alpha) = N(\mu, C)$$
$$C = \left(\alpha_1 X^T X + \alpha_2 I_k\right)^{-1}$$
$$\mu = \alpha_1 C X^T y$$

Posterior Probability Maps (PPMs)

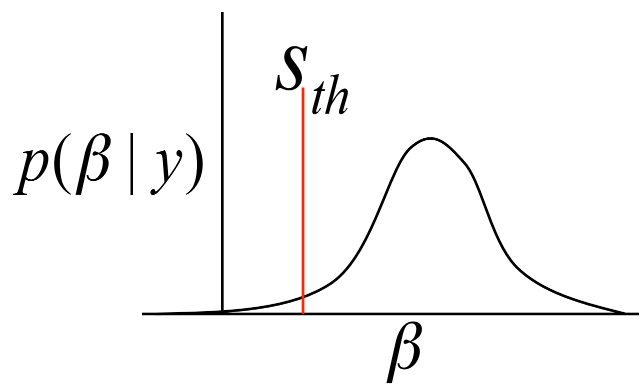
Posterior distribution: probability of the effect given the data

$$p(\beta \mid y)$$

mean: size of effect
precision: variability

Posterior probability map: images of the probability (confidence) that an activation exceeds some specified threshold s_{th}, given the data y

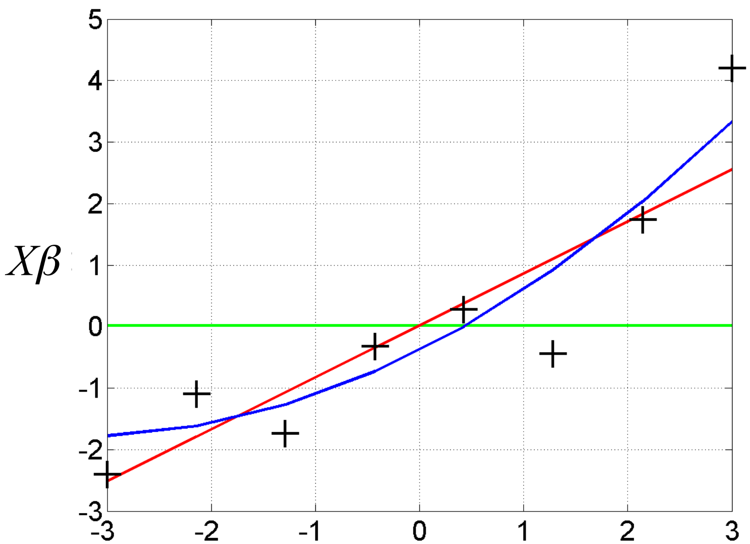
$$p(\beta > s_{th} \mid y) > p_{th}$$



Two thresholds:

- activation threshold s_{th} : percentage of whole brain mean signal (physiologically relevant size of effect)
- probability p_{th} that voxels must exceed to be displayed (e.g. 95%)

Bayesian linear regression: *model selection*

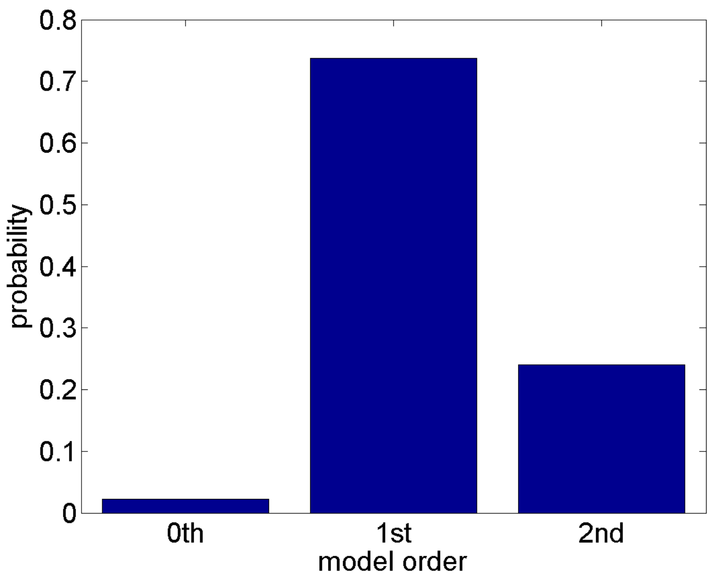


Bayes Rule:

$$p(\beta|y,\alpha,m) = \frac{p(y\,|\,\beta,\alpha,m)\,p(\beta\,|\,\alpha,m)}{\underset{\text{normalizing constant}}{p(y|\alpha,m)}}$$

Model evidence:

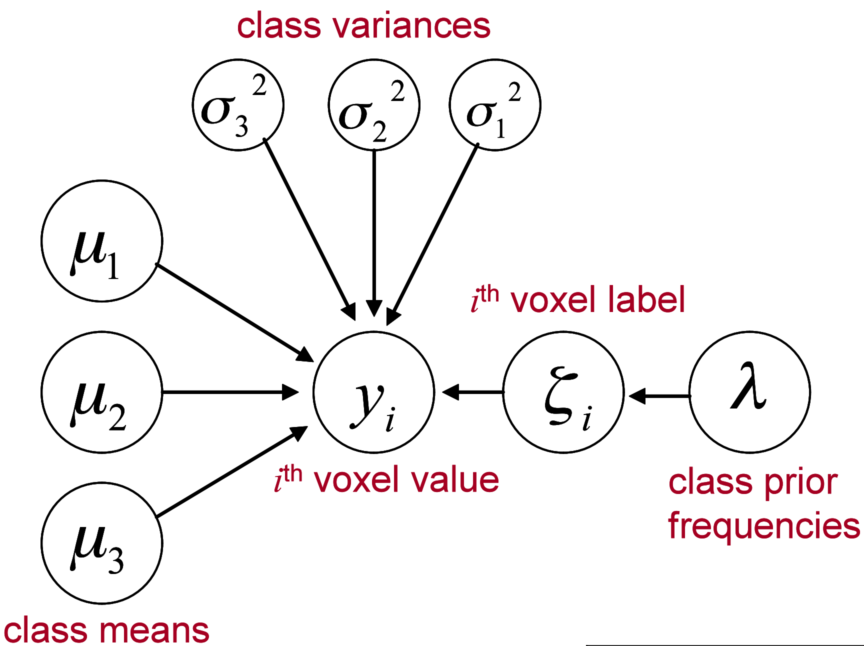
$$p(y|\alpha,m) = \int p(y\,|\,\beta,\alpha,m)\,p(\beta\,|\,\alpha,m)\,d\beta$$

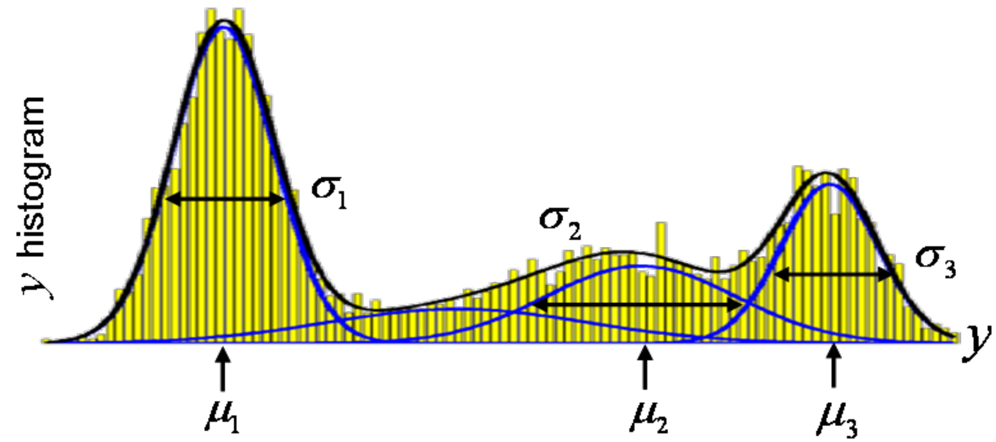


$$\log p(y\,|\,\alpha,m) \;=$$
$$accuracy(m) - complexity(m)$$

$$accuracy(m) \propto \|y - X\mu\|^2$$
$$complexity(m) \propto k\log\alpha_2^{-1} + \alpha_2\|\mu\|^2$$

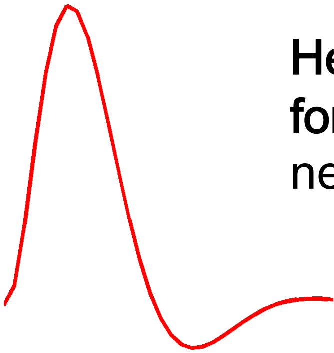
aMRI segmentation





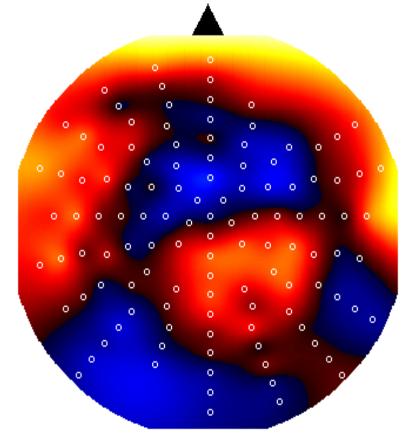
PPM of belonging to… grey matter white matter CSF

Dynamic Causal Modelling:
generative model for fMRI and ERPs



Hemodynamic forward model:
neural activity→BOLD

Electric/magnetic forward model:
neural activity→EEG
MEG
LFP



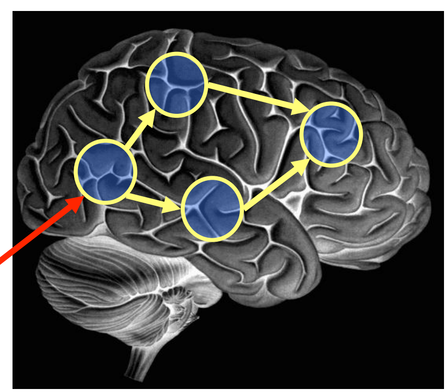
Neural state equation:

$$\dot{x} = F(x, u, \theta)$$

fMRI

ERPs

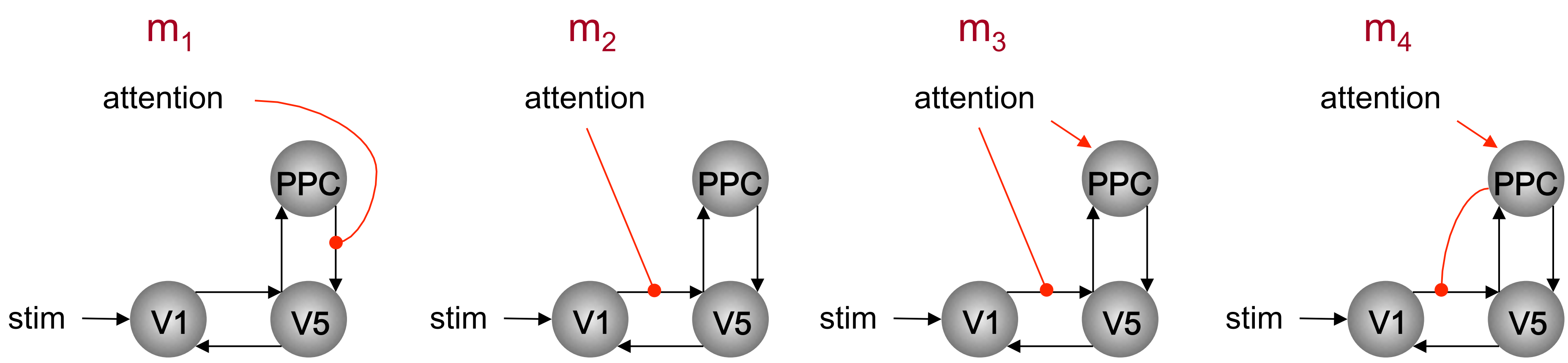
Neural model:
1 state variable per region
bilinear state equation
no propagation delays



inputs

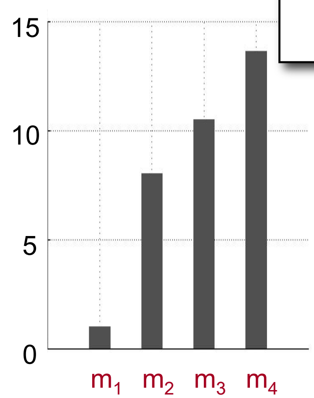
Neural model:
8 state variables per region
nonlinear state equation
propagation delays

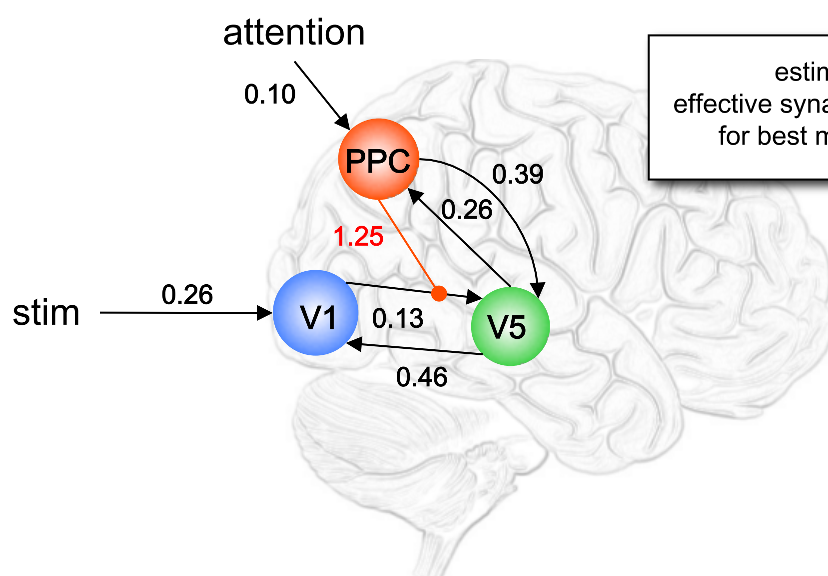
Bayesian Model Selection for fMRI



models marginal likelihood

$$\ln p\left(y|m\right)$$





estimated
effective synaptic strengths
for best model (m_4)

[Stephan et al., Neuroimage, 2008]

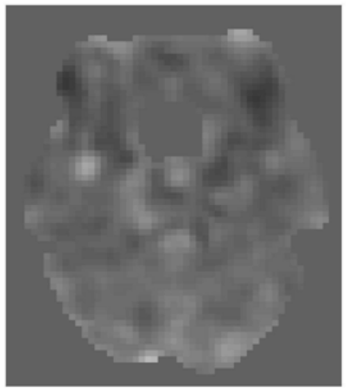
fMRI time series analysis with spatial priors

degree of smoothness Spatial precision matrix

$$Y = X\beta + \varepsilon$$

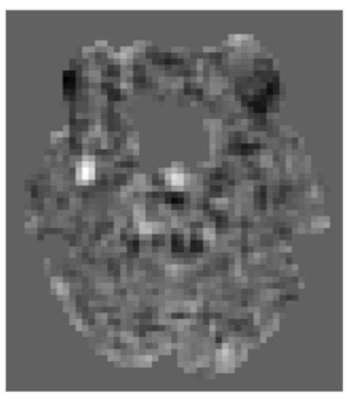
$$p(\beta) = N\left(0, \underline{\alpha}^{-1}\underline{L}^{-1}\right)$$

aMRI

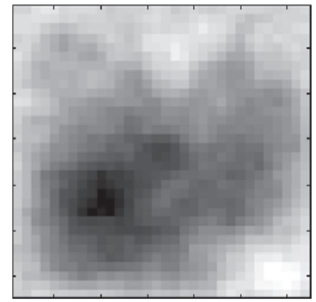


Smooth Y (RFT)

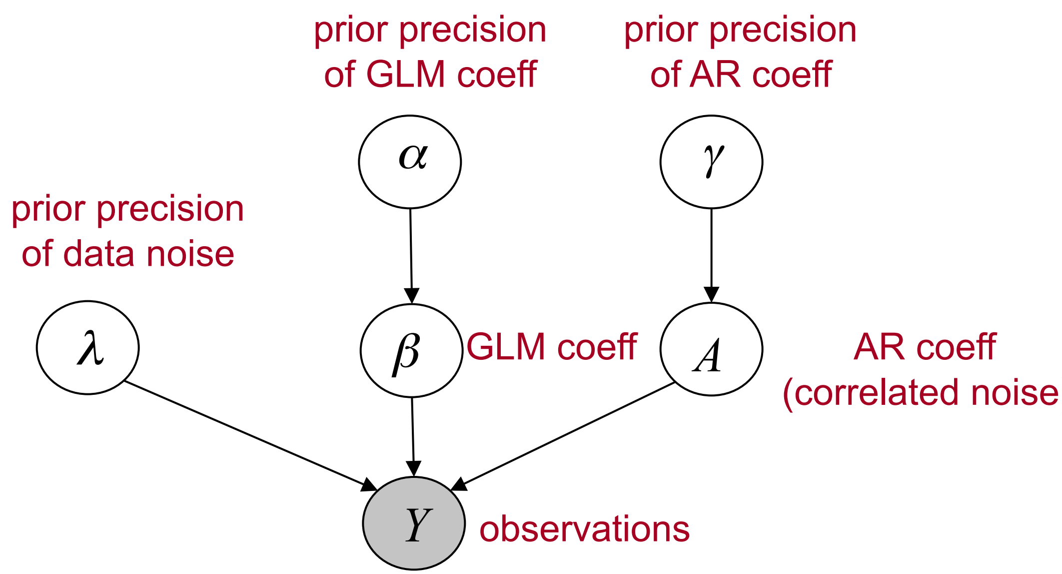
ML estimate of β



VB estimate of β

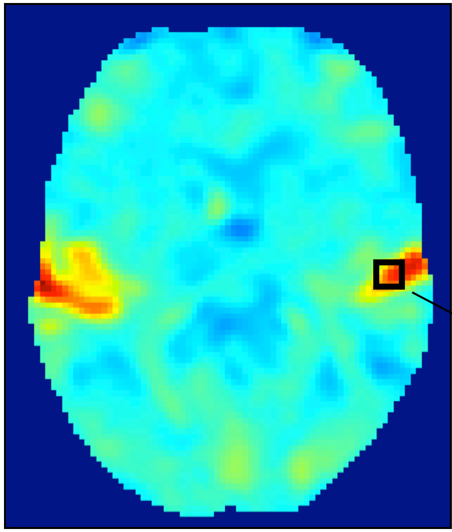


		1		
	2	-8	2	
1	-8	20	-8	1
	2	-8	2	
		1		

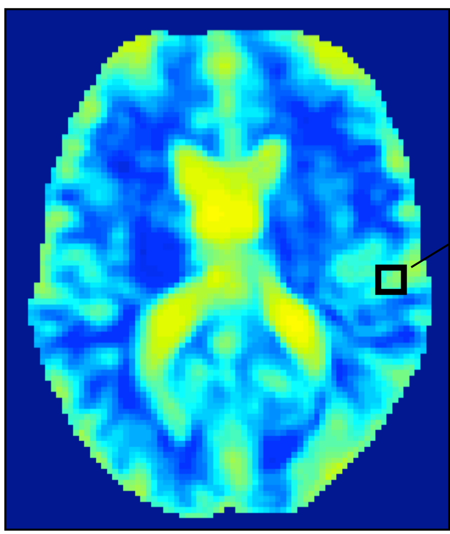


Penny et al 2005

fMRI time series analysis with spatial priors: *posterior probability maps*

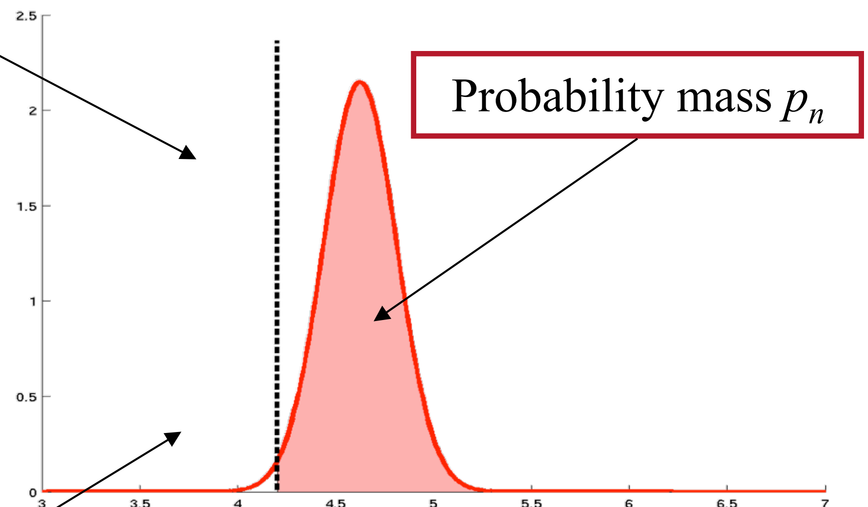


Mean *(Cbeta_*.img)*



Std dev *(SDbeta_*.img)*

activation threshold s_{th}

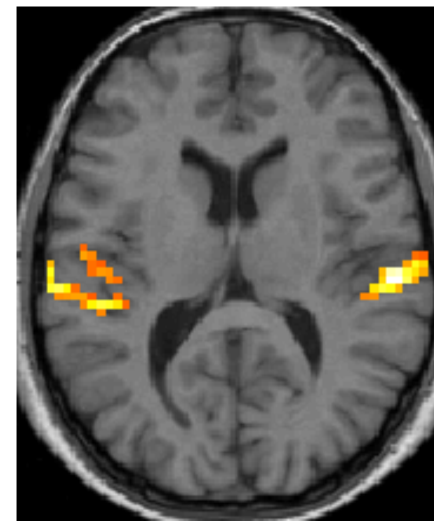


Posterior density $q(\beta_n)$

Display only voxels that exceed e.g. 95%

$$p > p_{th}$$

$$p = q(\beta > s_{th})$$



PPM *(spmP_*.img)*

probability of getting an effect, given the data

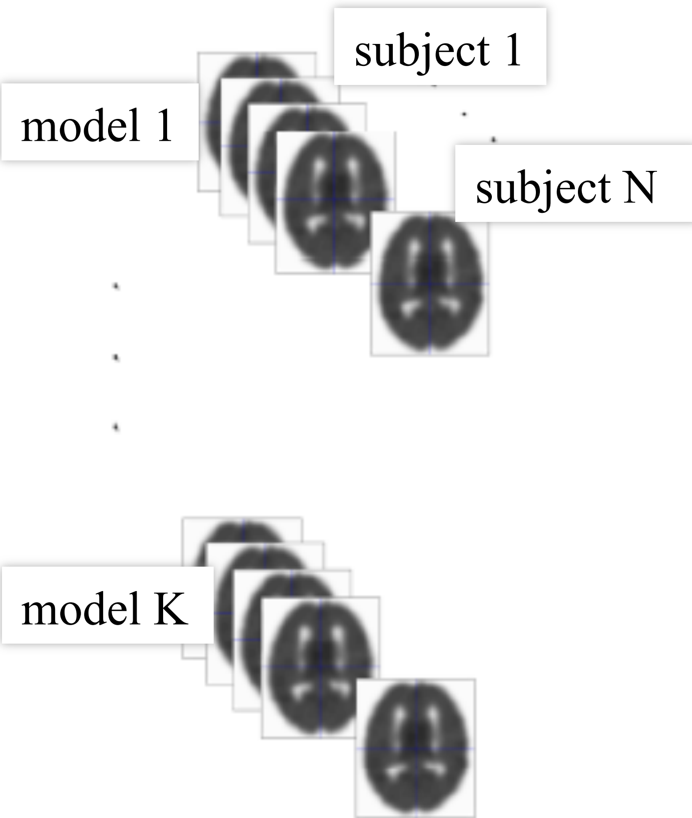
$$q(\beta_n) = N(\mu_n, \Sigma_n)$$

mean: *size of effect*
covariance: *uncertainty*

fMRI time series analysis with spatial priors: *Bayesian model selection*

$$\log p(y|m) \approx F(q)$$

Log-evidence maps



Compute log-evidence for each model/subject

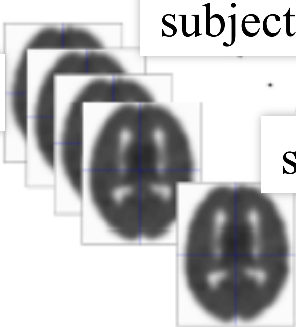
fMRI time series analysis with spatial priors: *Bayesian model selection*

$$\log p(y|m) \approx F(q)$$

Log-evidence maps

subject 1

model 1



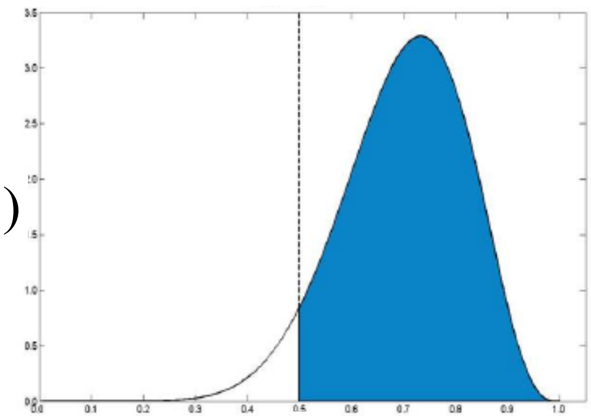
subject N

model K

Compute log-evidence for each model/subject

$q(r_k > 0.5) = 0.941$

$q(r_k)$

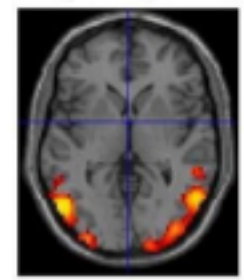


r_k

Probability that model k generated data

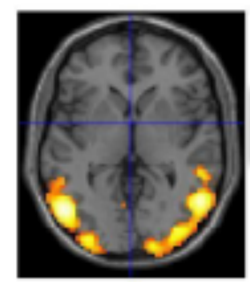
BMS maps

$\langle r_k \rangle > \gamma$



PPM

$\varphi_k > \gamma$

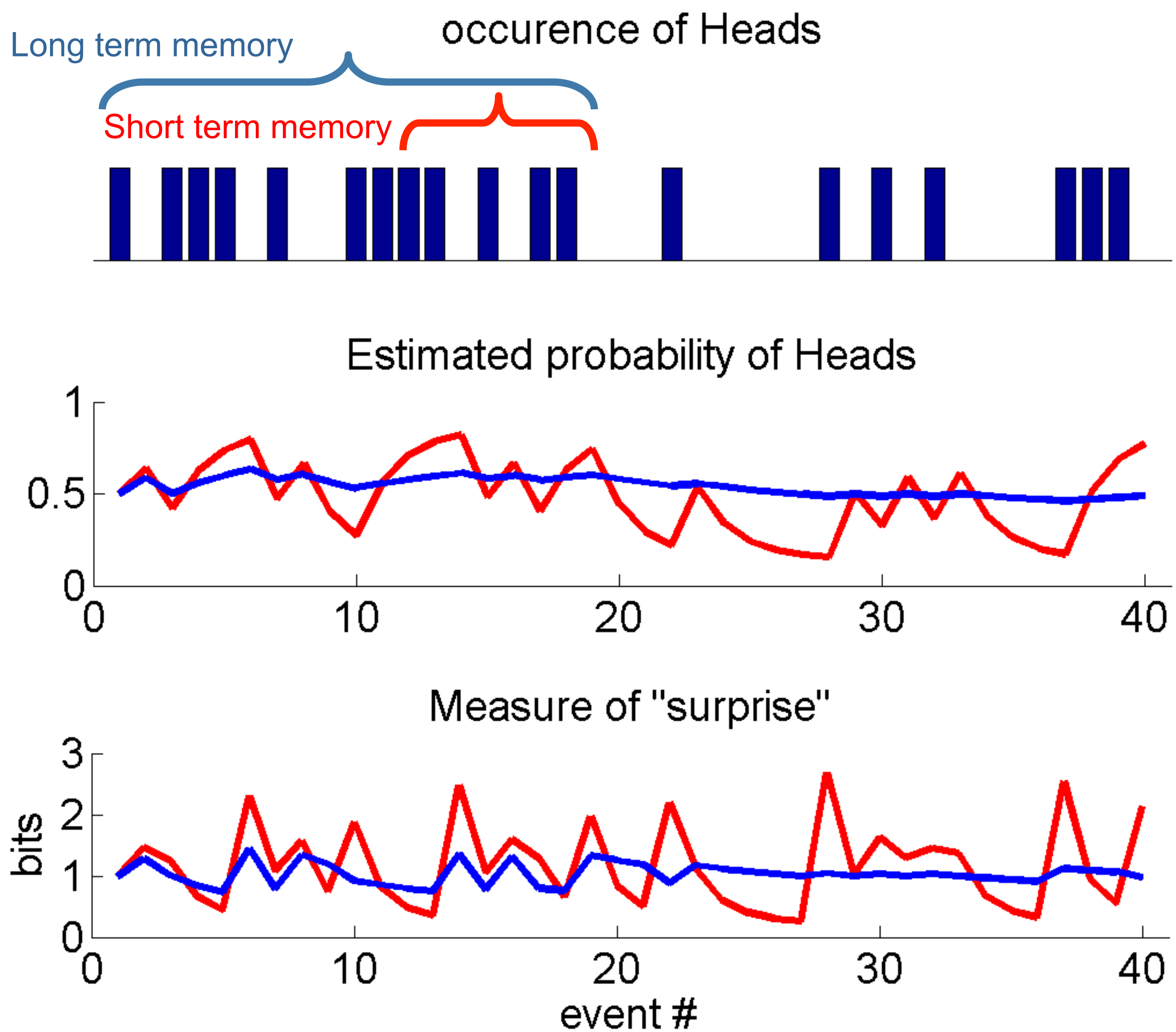


EPM

model k

Joao et al, 2009

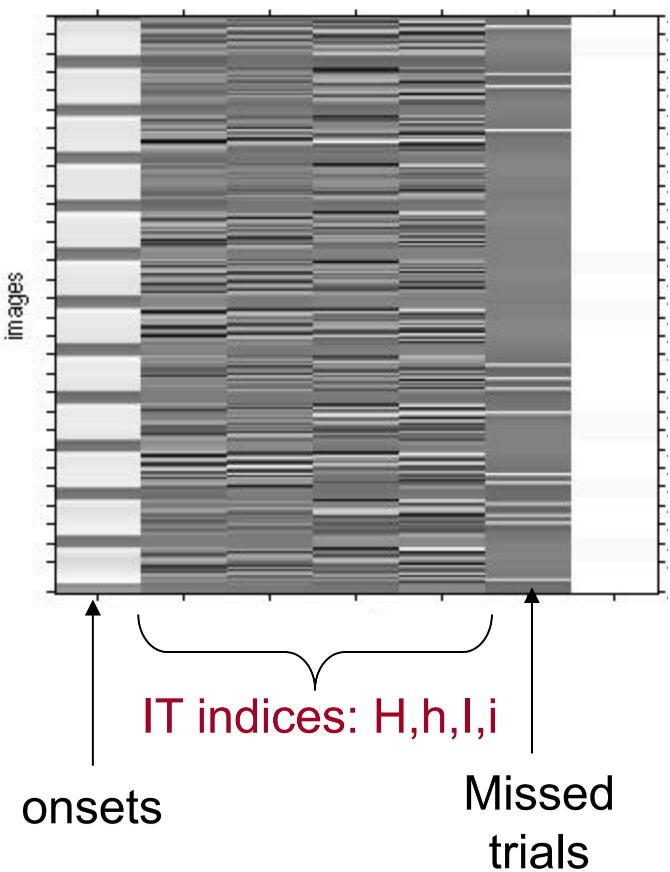
Reminder…



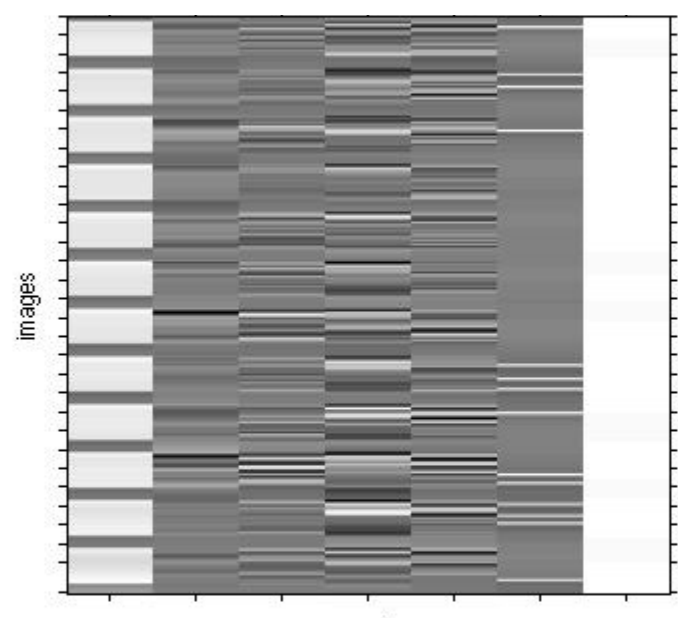
occurence of Heads
Long term memory
Short term memory
Estimated probability of Heads
1
0.5
0
0
10
20
30
40
Measure of "surprise"
bits
3
2
1
0
0
10
20
30
40
event #

Compare two models

Short-term memory model



long-term memory model

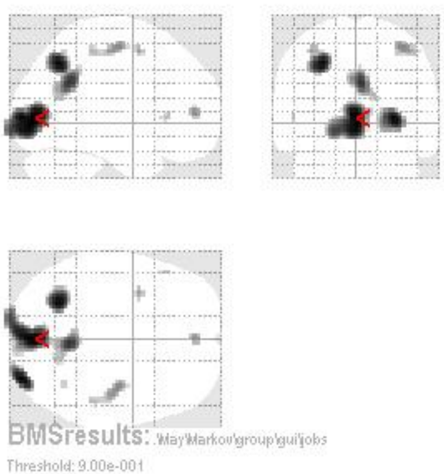


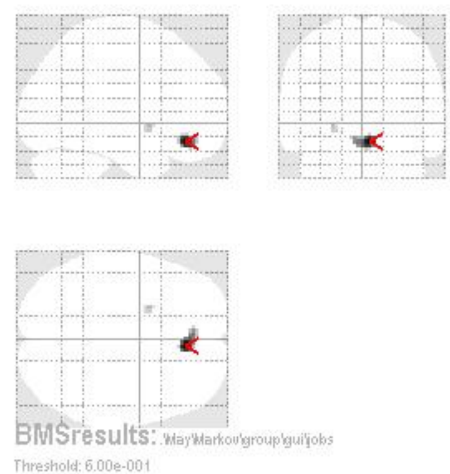
IT indices are smoother

H=entropy; h=surprise; I=mutual information; i=mutual surprise

Group data: Bayesian Model Selection maps

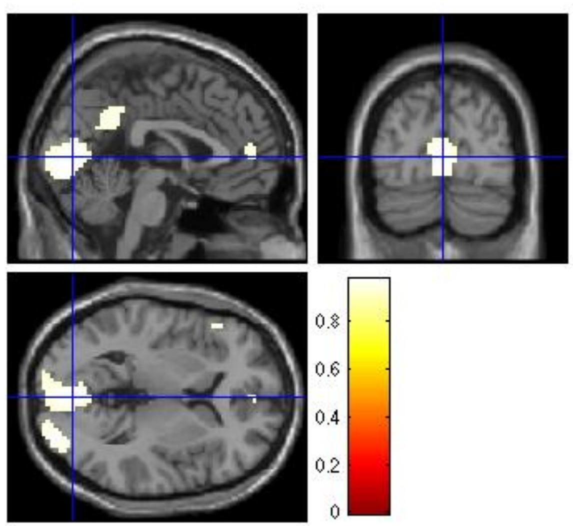
Regions best explained by short-term memory model

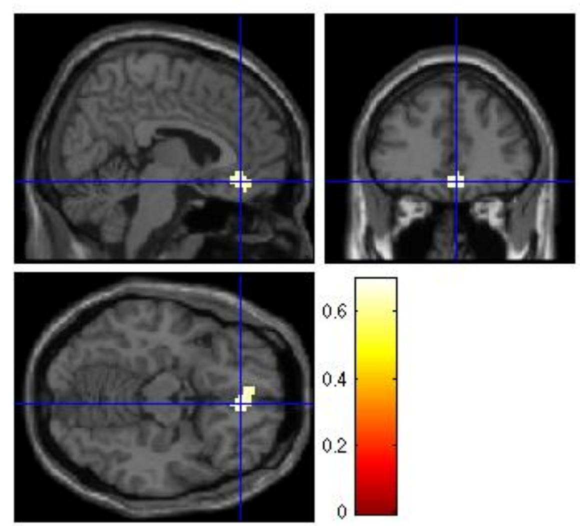




Regions best explained by long-term memory model

primary visual cortex





frontal cortex (executive control)

Thank-you