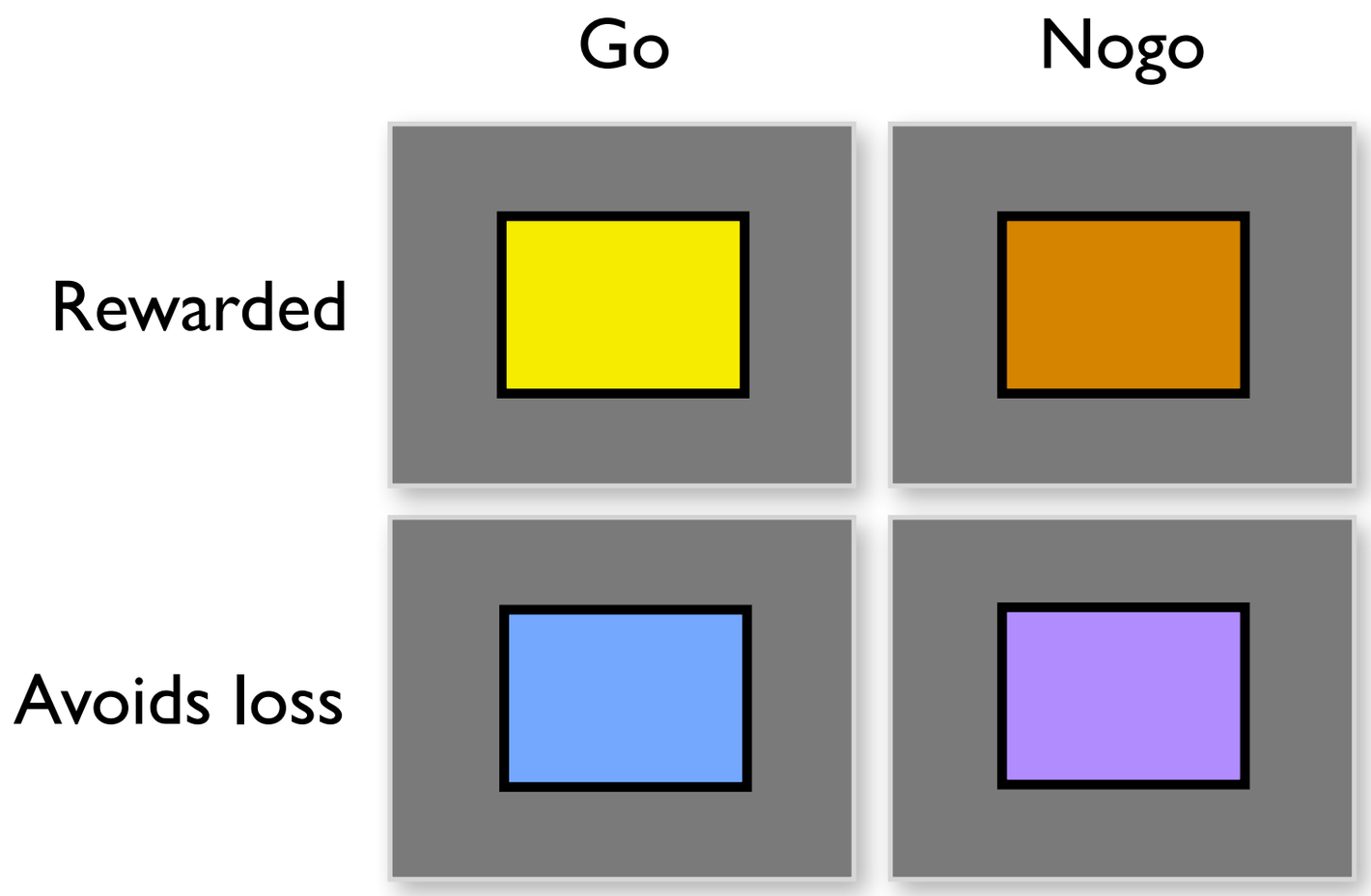
Fitting behavioural models

Quentin Huys
Dr. med. MBBS PhD

Max Planck UCL Centre for Computational Psychiatry and Ageing
and Camden and Islington NHS Foundation Trust

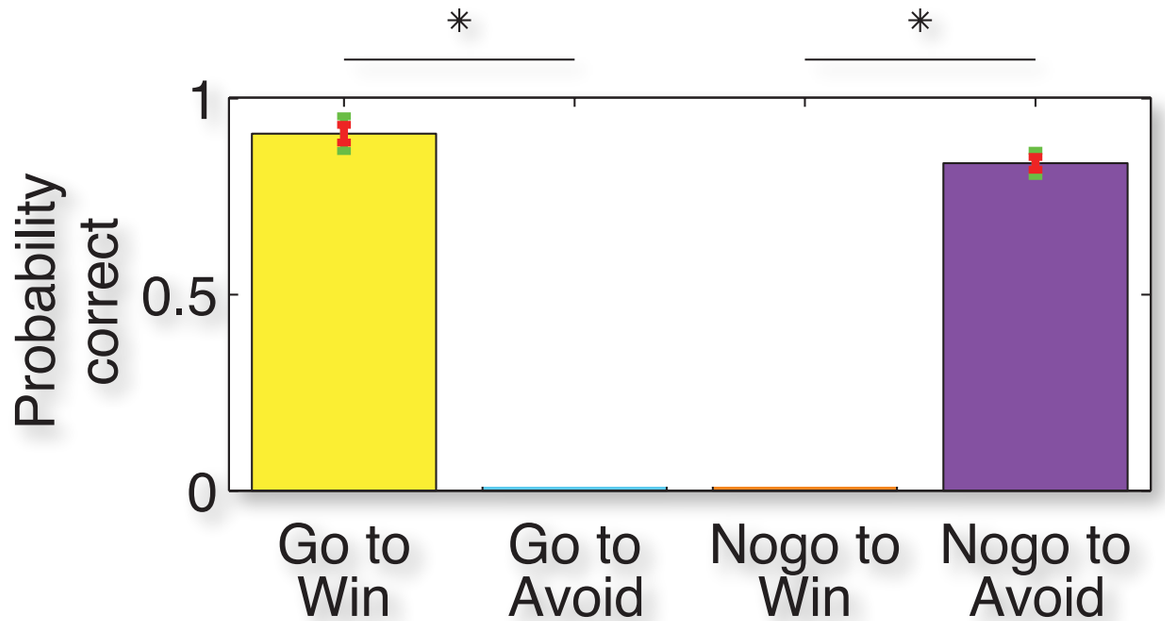
Ringberg
26/9/18

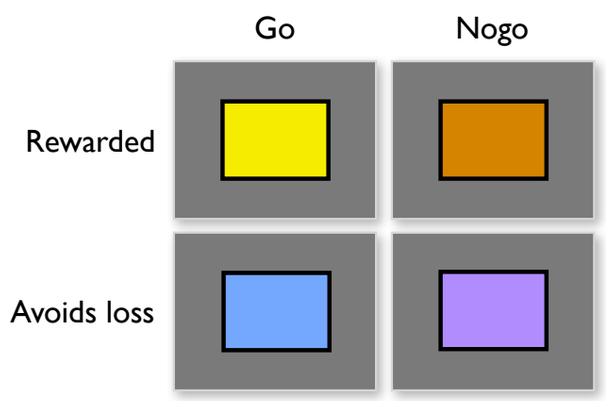
Example task

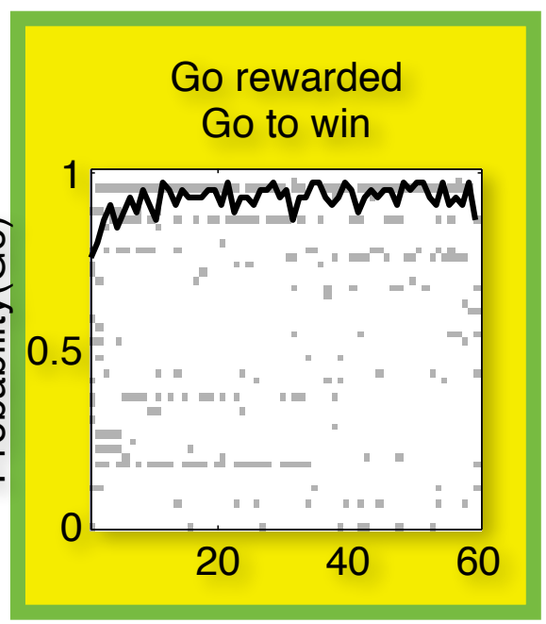


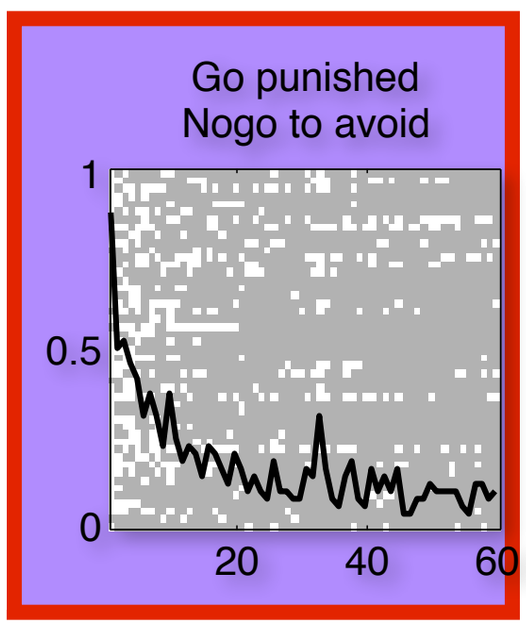
Guitart-Masip, Huys et al. *2012*

Example task



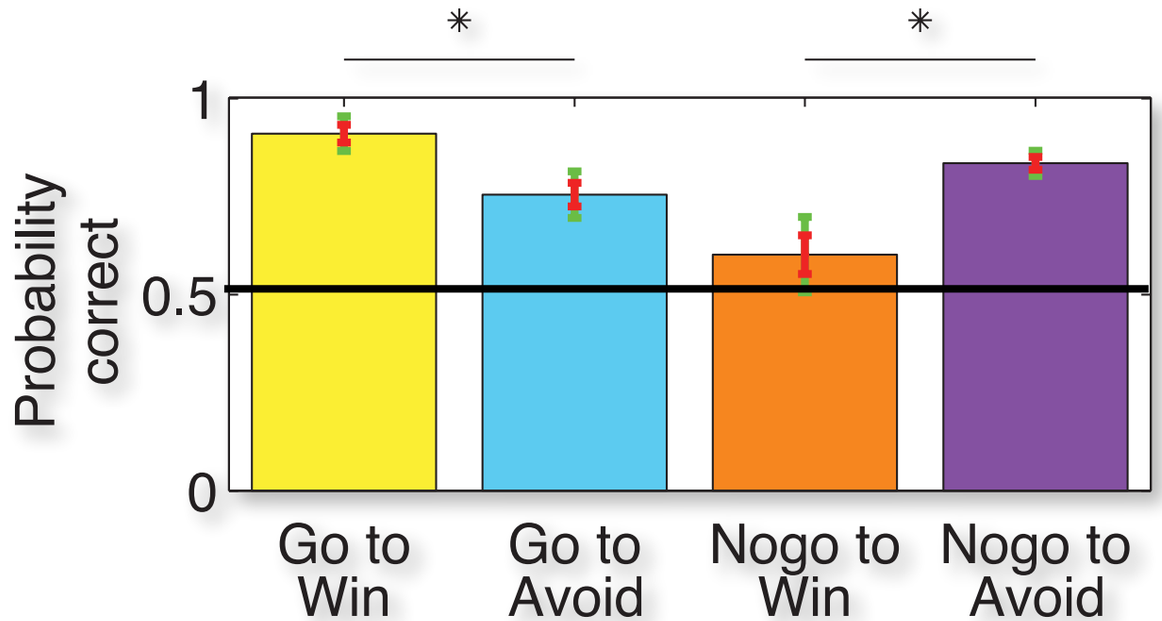


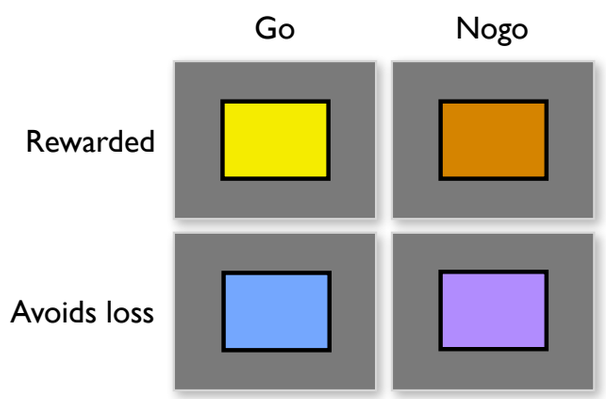


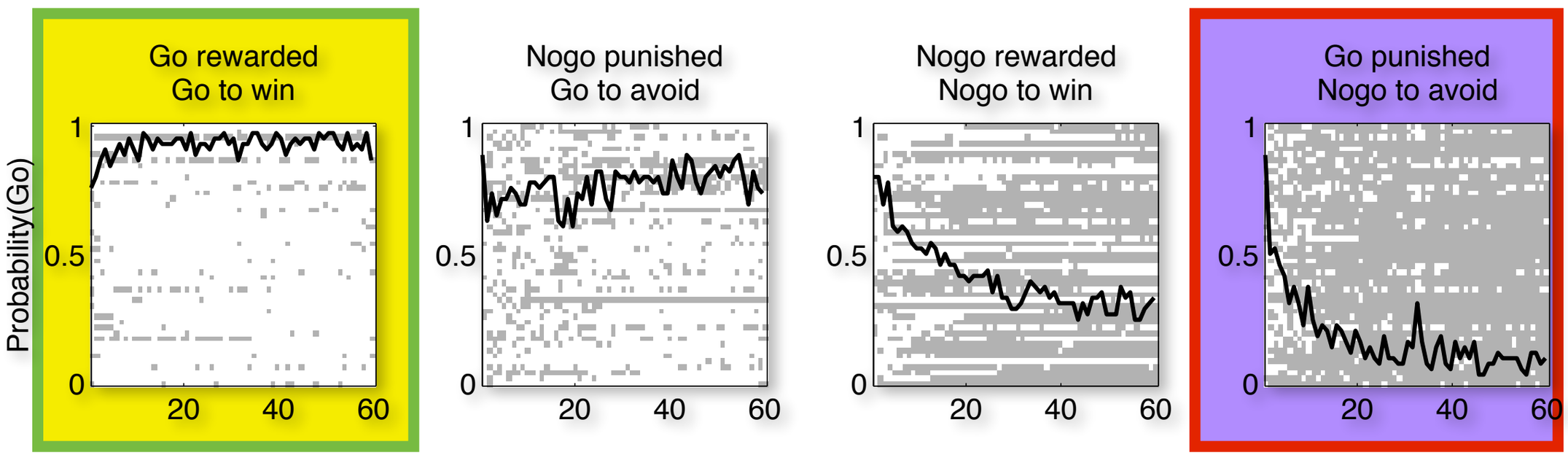


Guitart-Masip, Huys et al. *2012*

Example task

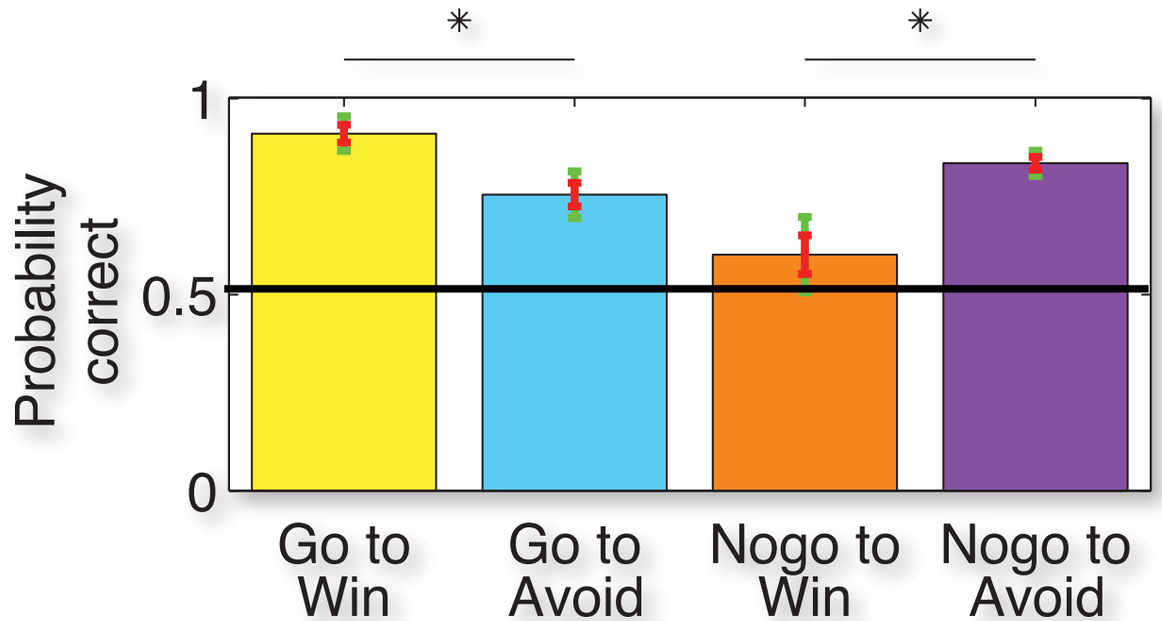


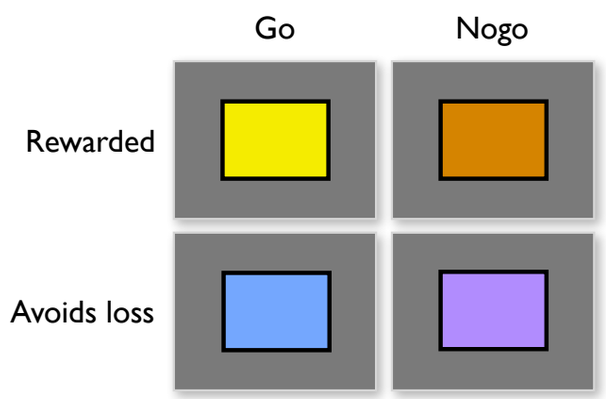


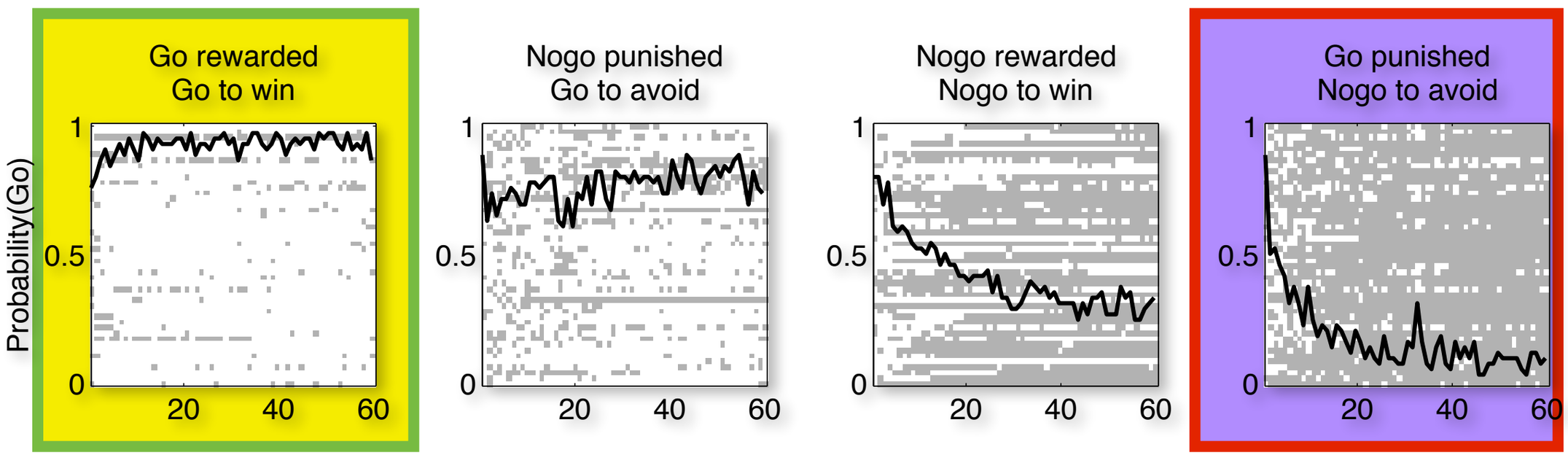


Guitart-Masip, Huys et al. *2012*

Example task







Think of it as four separate two-armed bandit tasks

Guitart-Masip, Huys et al. *2012*

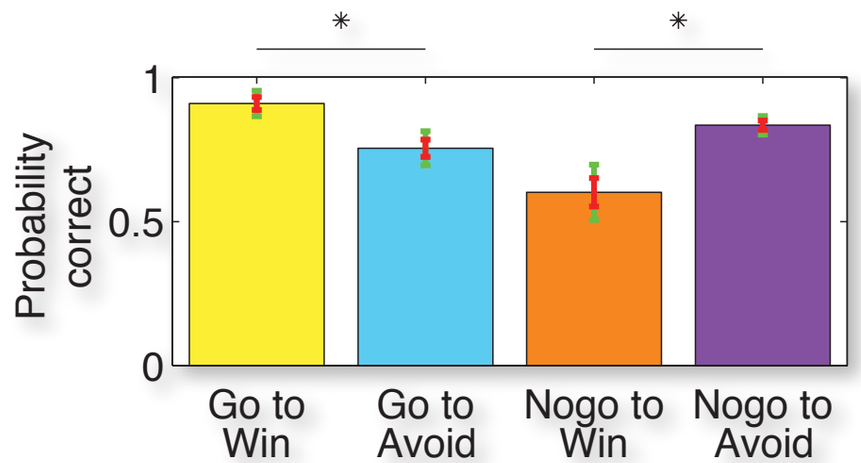
Analysing behaviour

- Standard approach:
  - Decide which feature of the data you care about
  - Run descriptive statistical tests, e.g. ANOVA

- Many strengths
- Weakness
  - Piecemeal, not holistic / global
  - Descriptive, not generative
  - No internal variables

Analysing behaviour

- Standard approach:
  - Decide which feature of the data you care about
  - Run descriptive statistical tests, e.g. ANOVA



- Many strengths
- Weakness
  - Piecemeal, not holistic / global
  - Descriptive, not generative
  - No internal variables

Models

- Holistic
  - Aim to model the process by which the data came about in its “entirety”
- Generative
  - They can be run on the task to generate data as if a subject had done the task
- Inference process
  - Capture the inference process subjects have to make to perform the task.
  - Do this in sufficient detail to replicate the data.
- Parameters
  - replace test statistics
  - their meaning is explicit in the model
  - their contribution to the data is assessed in a holistic manner

How to fit a model and believe the

Model building The first step is to build a series of models. Each contains an internal process by which different choice options are valued, and a link function which describes how preferences turn into observed decisions. At least two models should be built: a model M0 of 'no interest' that performs the task, but without involving the process of interest, and a model M1 that does contain the process of interest.

Validation on surrogate data

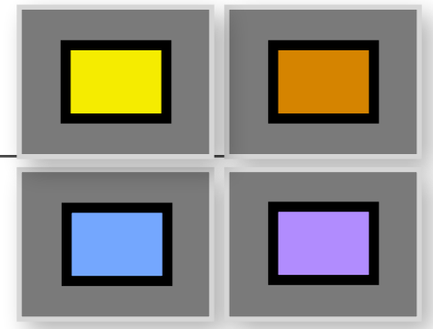
1. **Data generation**: Run each model on the experiment from which data will be examined. Do the generated data look reasonable?
2. **Surrogate model fitting**: Fit each model to the data generated from it. Are the true parameters readily recovered? Are some parameters not identifiable?
3. **Surrogate model comparison**: Does the model comparison procedure correctly identify the data generated by each model?

Real data analysis

1. **Real model fitting**: Fit each model to the real data.
2. **Real model validation**: Run each model with the fitted parameters on the exact experimental instance presented to that particular subject. Are the key features of the real data captured reasonably?
3. **Real model comparison**: choose the least complex model that best accounts for the data.
4. **Parameter examination**: only at this point should the parameters of the model be examined, and only the parameters of the most parsimonious model should be ascribed meaning.

Huys 2017

Actions



- Q values “the process”

$$\mathcal{Q}_t(a_t, s_t) = \mathcal{Q}_{t-1}(a_t, s_t) + \epsilon(r_t - \mathcal{Q}_{t-1}(a_t, s_t))$$

- Probabilities “link function”

$$p(a_t|s_t, h_t, \beta) = p(a_t|\mathcal{Q}(a_t, s_t), \beta)$$

$$= \frac{e^{\beta \mathcal{Q}(a_t, s_t)}}{\sum_{a'} e^{\beta \mathcal{Q}(a', s_t)}}$$

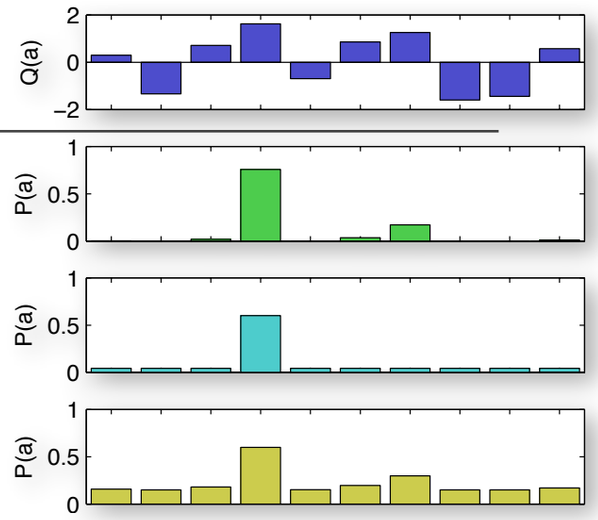
- Features:

$$p(a_t|s_t) \propto \mathcal{Q}(a_t, s_t)$$

$$0 \leq p(a) \leq 1$$

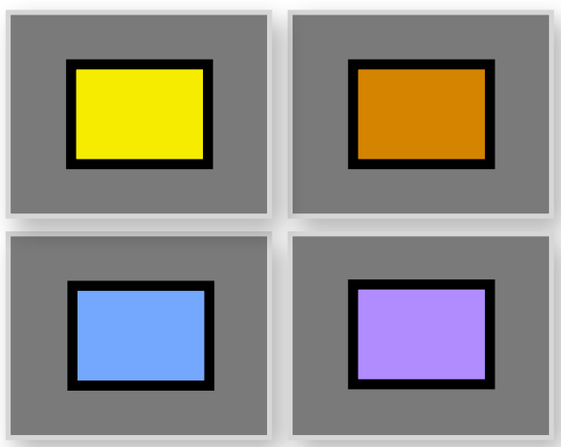
- links learning process and observations
  - choices, RTs, or any other data
  - link function in GLMs
  - many other forms

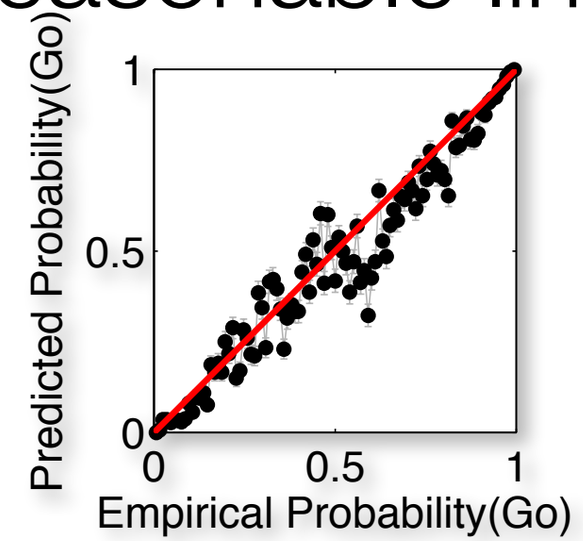
Link functions



- Sigmoid $p(a|s) = \frac{e^{\beta \mathcal{Q}(a,s)}}{\sum_{a'} e^{\beta \mathcal{Q}(a',s)}}$
- ϵ - greedy $p(a|s) = \begin{cases} c & \text{if } a = \text{argmax}_a \, \mathcal{Q}(a,s) \\ \frac{1-c}{|a|-1} & \text{else} \end{cases}$
- irreducible noise $p(a|s) = \frac{1-g}{2} + g \frac{e^{\beta \mathcal{Q}(a,s)}}{\sum_{a'} e^{\beta \mathcal{Q}(a',s)}}$

- critical sanity check 1: reasonable link function?





- other link functions for other observations

Fitting models I

- Maximum likelihood (ML) parameters

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta)$$

- where the likelihood of all choices is:

$$\begin{aligned}
\mathcal{L}(\theta) &= \log p(\{a_t\}_{t=1}^T | \{s_t\}_{t=1}^T, \{r_t\}_{t=1}^T, \underbrace{\theta}_{\beta,\epsilon}) \\
&= \log p(\{a_t\}_{t=1}^T | \{\mathcal{Q}(s_t, a_t; \epsilon)\}_{t=1}^T, \beta) \\
&= \log \prod_{t=1}^{T} p(a_t | \mathcal{Q}(s_t, a_t; \epsilon), \beta) \\
&= \sum_{t=1}^{T} \log p(a_t | \mathcal{Q}(s_t, a_t; \epsilon), \beta)
\end{aligned}$$

Fitting models II

- No closed form
- Use your favourite method
  - gradients
  - fminunc / fmincon...
- Gradients for RW model

$$
\begin{aligned}
\frac{d\mathcal{L}(\theta)}{d\theta} &= \frac{d}{d\theta}\sum_t \log p(a_t|\mathcal{Q}_t(a_t,s_t;\epsilon),\beta) \\
&= \sum_t \frac{d}{d\theta}\beta\mathcal{Q}_t(a_t,s_t;\epsilon) - \sum_{a'} p(a'|\mathcal{Q}_t(a',s_t;\epsilon),\beta)\frac{d}{d\theta}\beta\mathcal{Q}_t(a',s_t;\epsilon) \\
\frac{d\mathcal{Q}_t(a_t,s_t;\epsilon)}{d\epsilon} &= (1-\epsilon)\frac{d\mathcal{Q}_{t-1}(a_t,s_t;\epsilon)}{d\epsilon} + (r_t - \mathcal{Q}_{t-1}(a_t,s_t;\epsilon))
\end{aligned}
$$

Little tricks

- Transform your variables

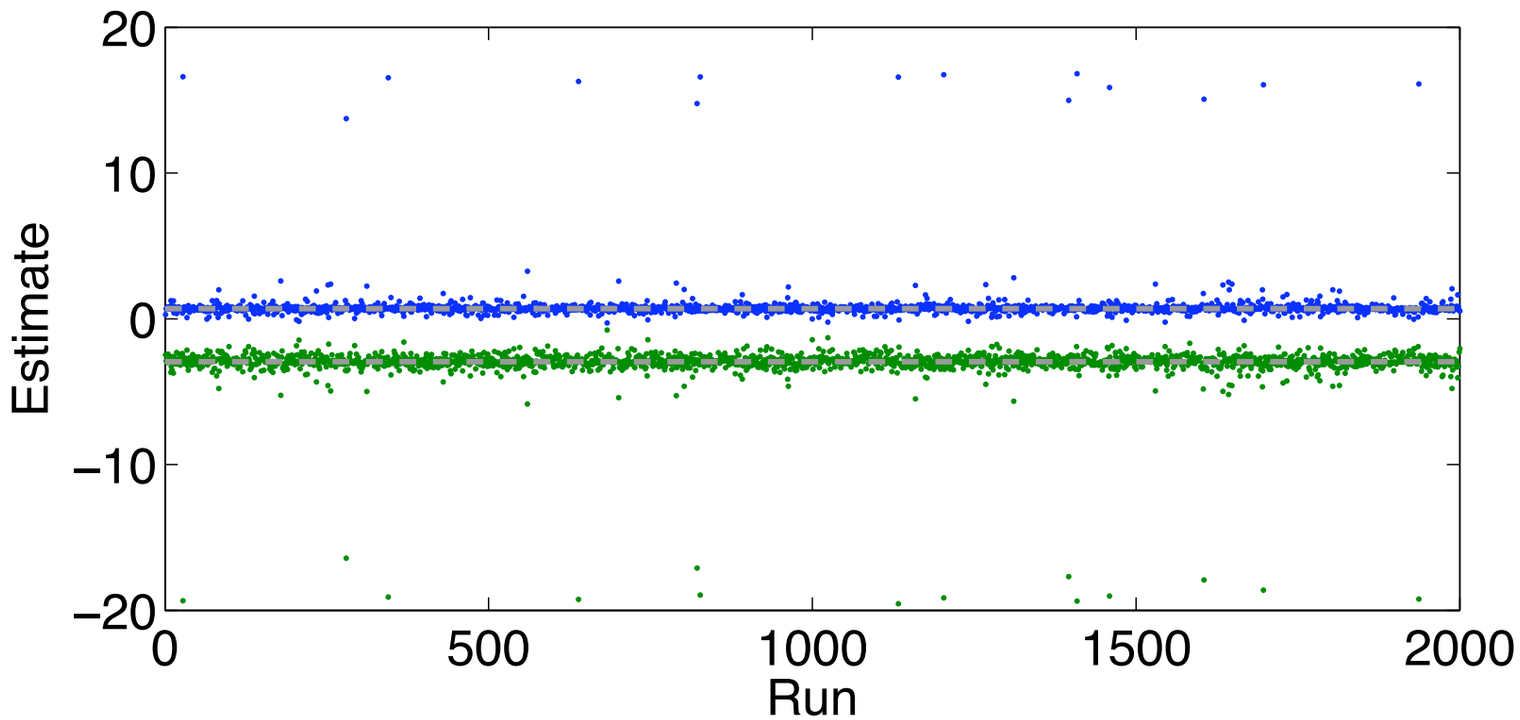
$$\begin{aligned}\beta &= e^{\beta'} \\ &\Rightarrow \quad \beta' = \log(\beta) \\ \epsilon &= \frac{1}{1+e^{-\epsilon'}} \\ &\Rightarrow \quad \epsilon' = \log\left(\frac{\epsilon}{1-\epsilon}\right)\end{aligned}$$

$$\frac{d\log\mathcal{L}(\theta')}{d\theta'}$$

- Avoid over/underflow

$$\begin{aligned}y(a) &= \beta\mathcal{Q}(a) \\ y_m &= \max_a y(a) \\ p &= \frac{e^{y(a)}}{\sum_b e^{y(b)}} = \frac{e^{y(a)-y_m}}{\sum_b e^{y(b)-y_m}}\end{aligned}$$

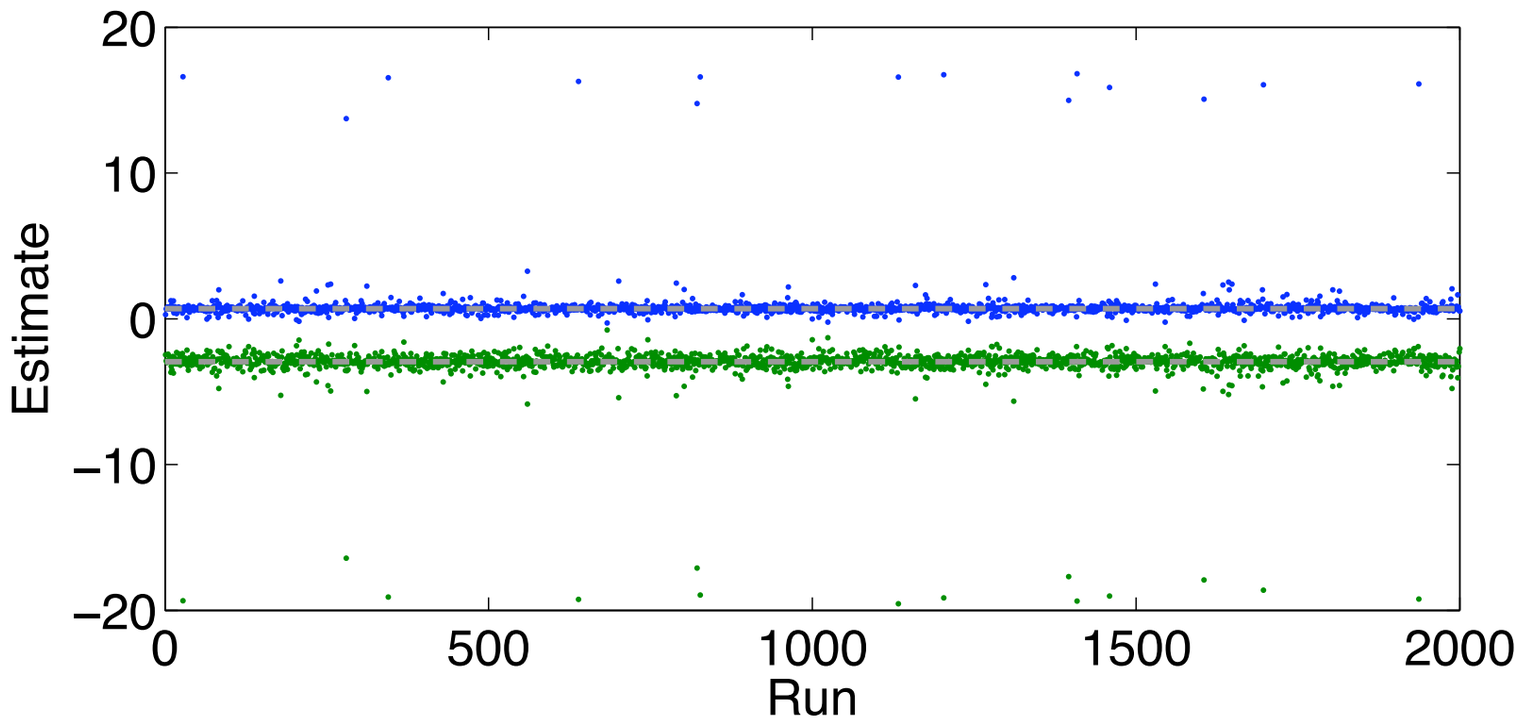
ML characteristics



200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2

ML characteristics

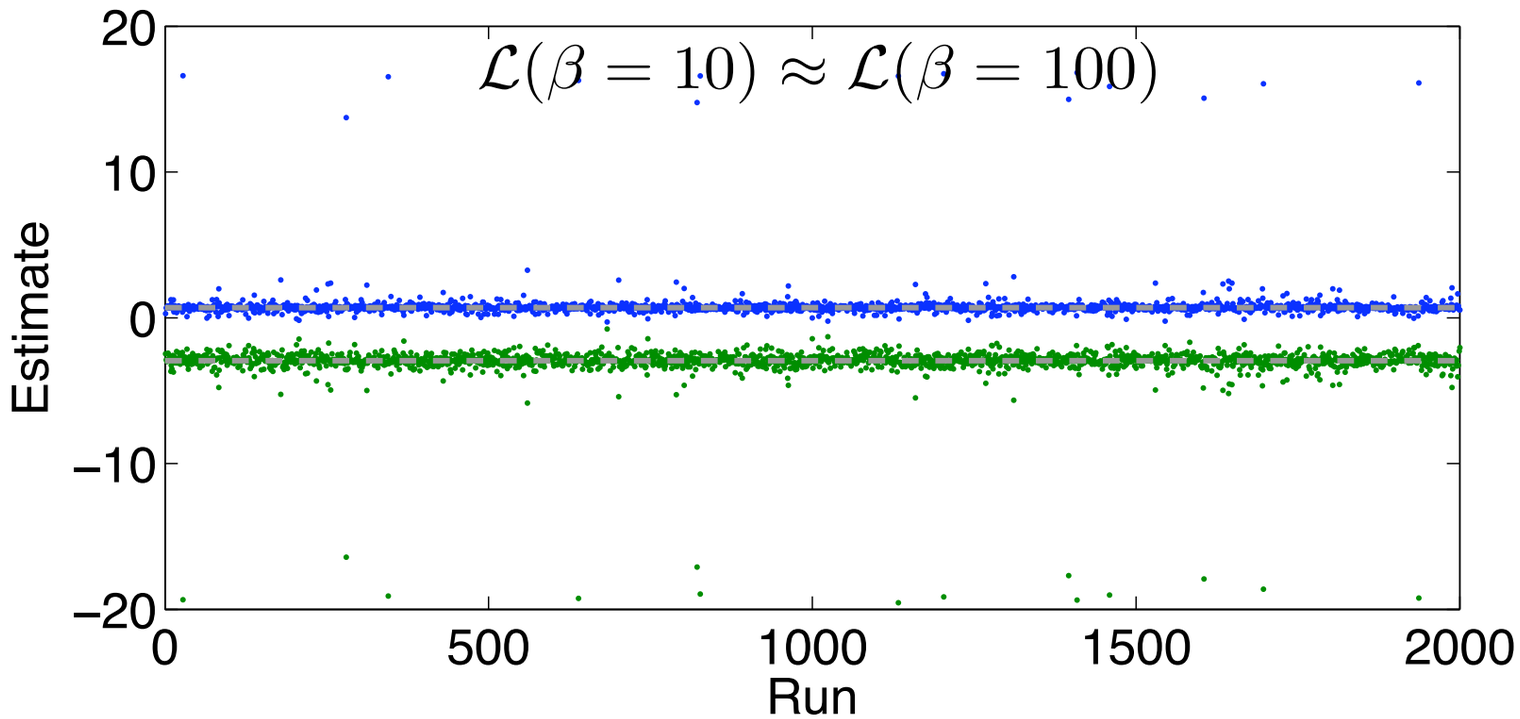
- ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon



200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2

ML characteristics

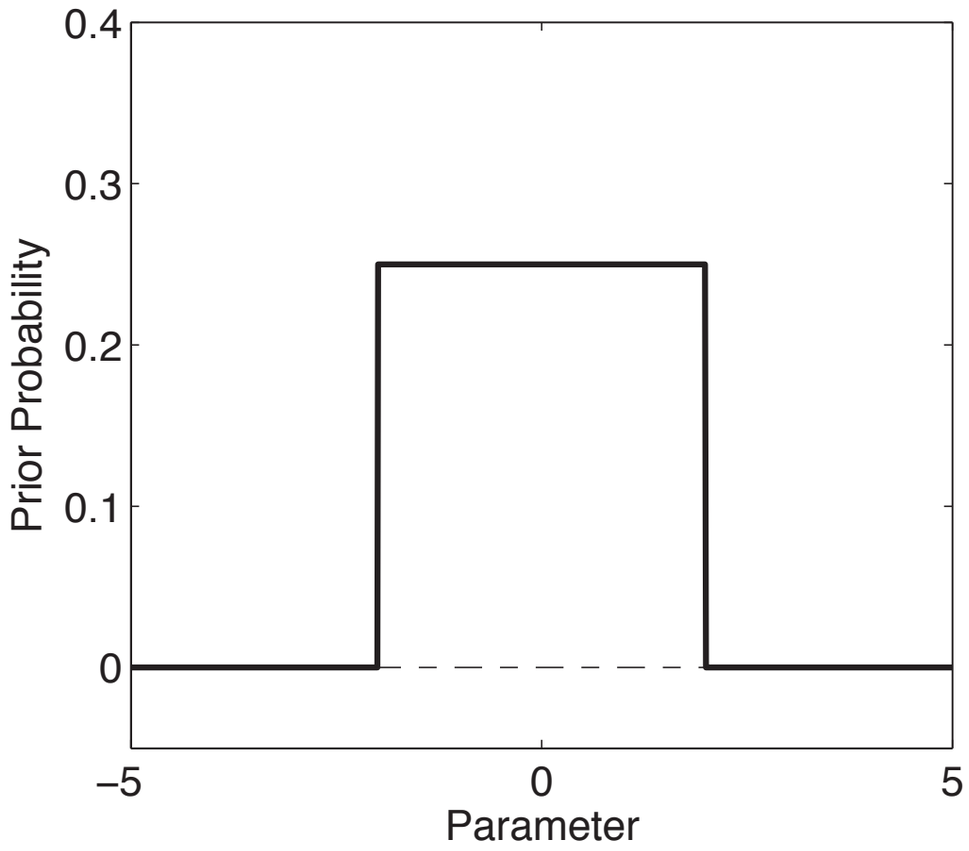
- ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon



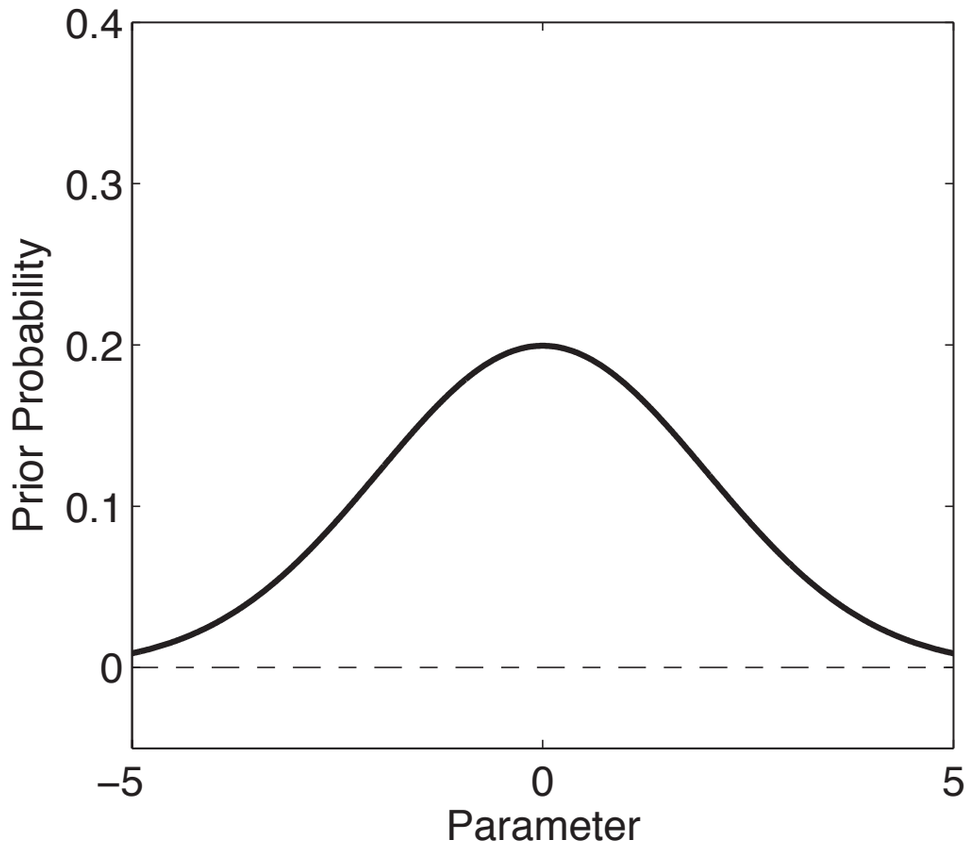
200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2

Priors

Not so smooth

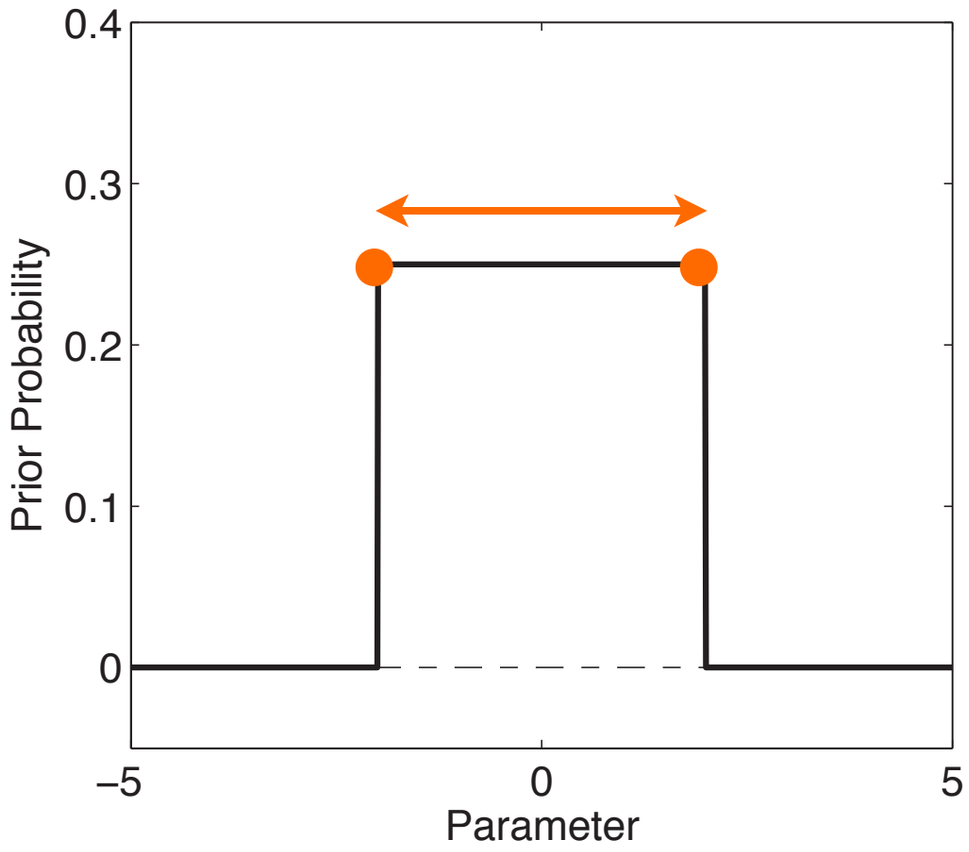


Smooth

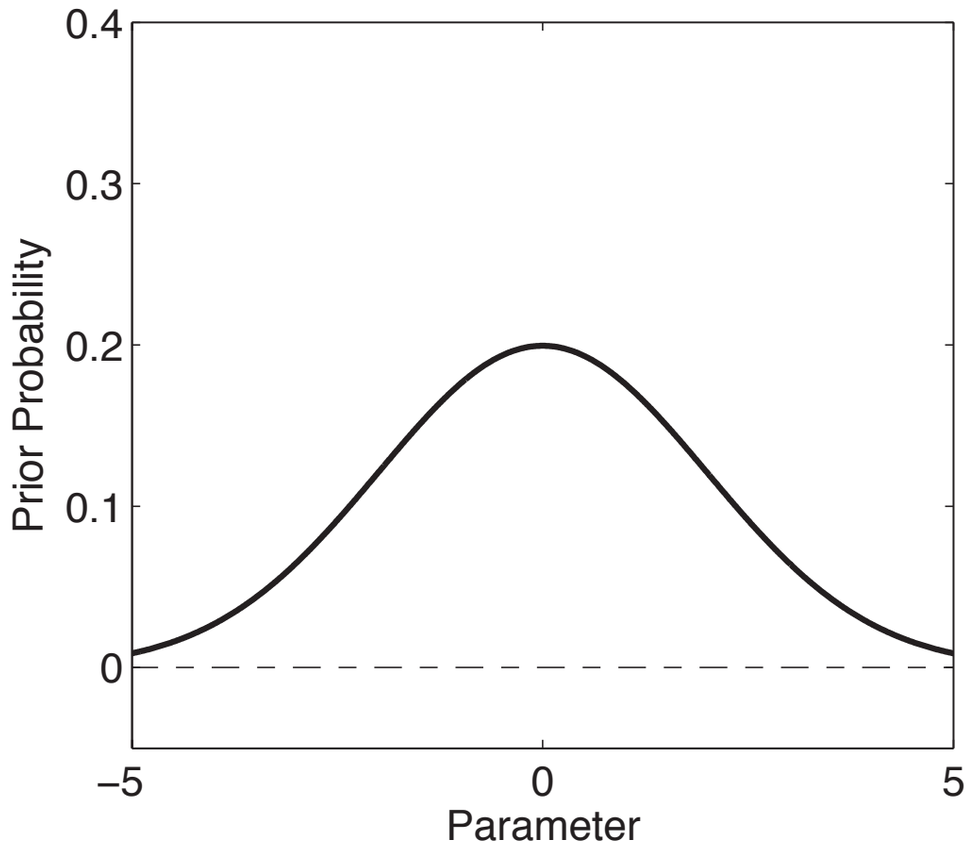


Priors

Not so smooth

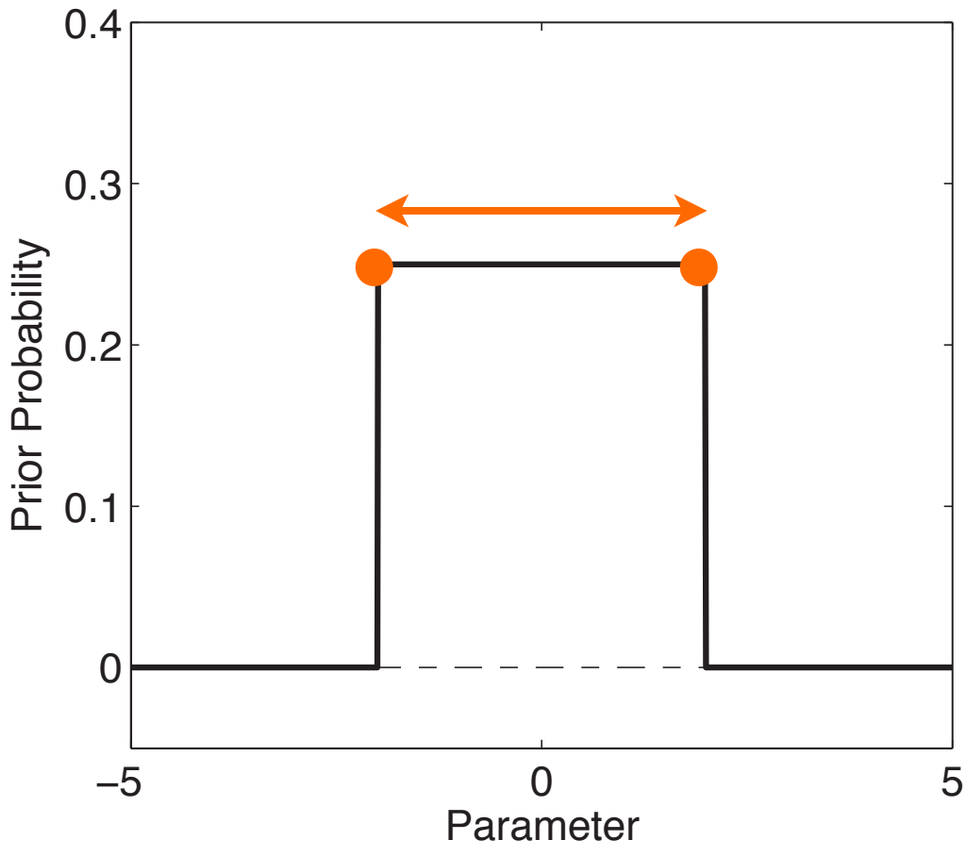


Smooth

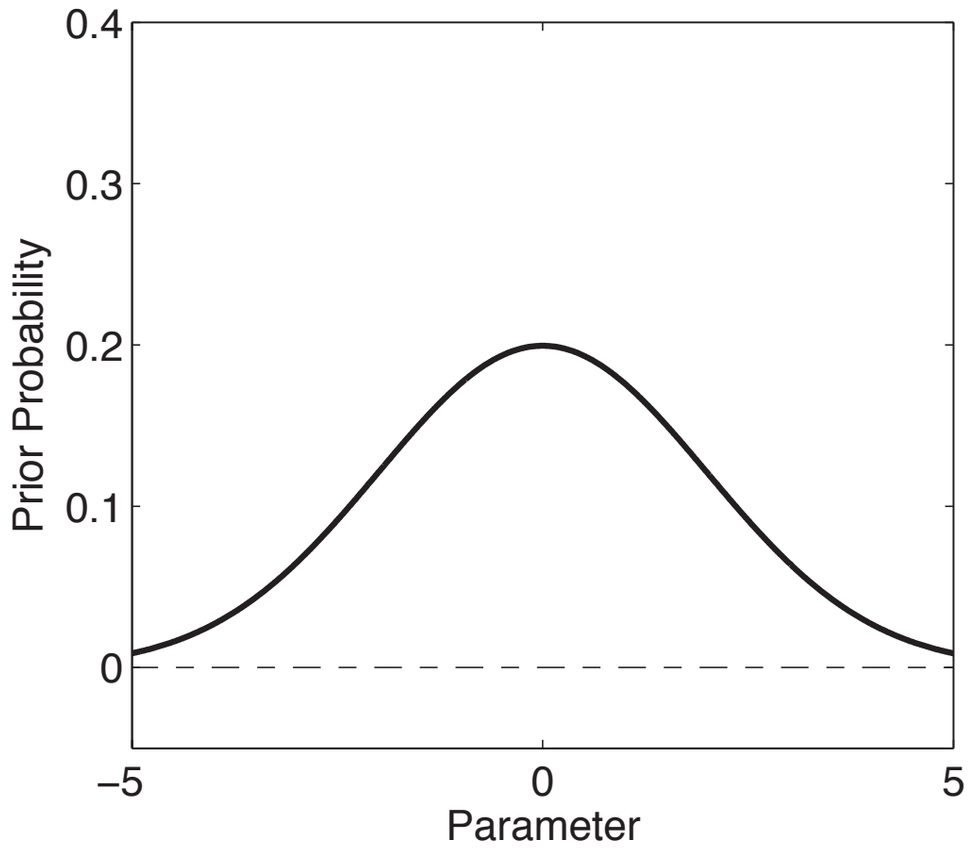


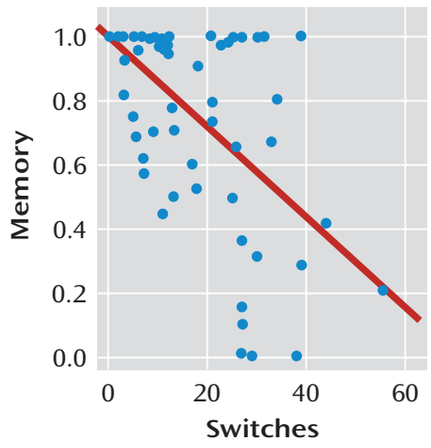
Priors

Not so smooth



Smooth





Maximum a posteriori estimate

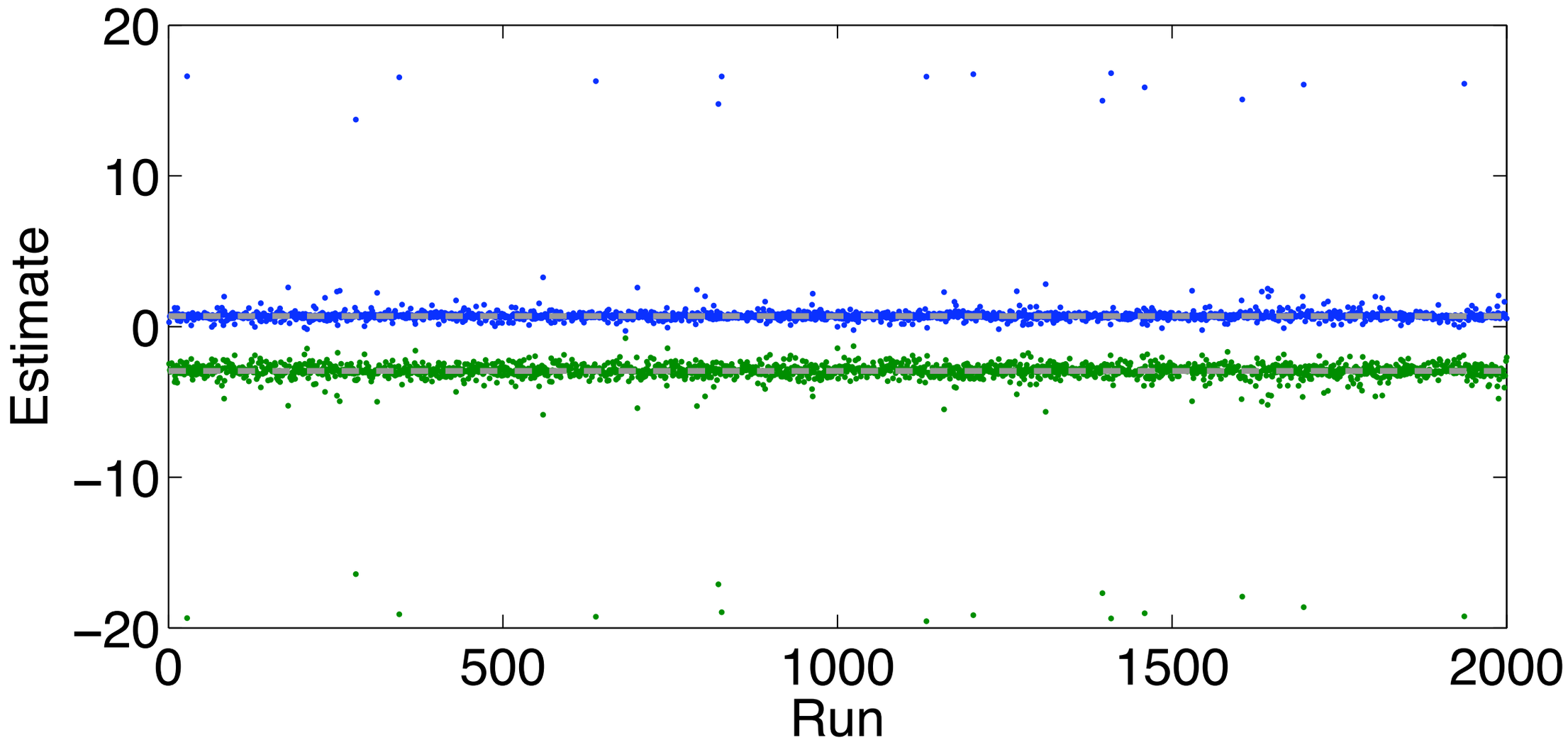
$$\mathcal{P}(\theta) = p(\theta|a_{1...T}) = \frac{p(a_{1...T}|\theta)p(\theta)}{\int d\theta p(\theta|a_{1...T})p(\theta)}$$

$$\log \mathcal{P}(\theta) = \sum_{t=1}^{T} \log p(a_t|\theta) + \log p(\theta) + const.$$

$$\frac{\log \mathcal{P}(\theta)}{d\alpha} = \frac{\log \mathcal{L}(\theta)}{d\alpha} + \frac{d\, p(\theta)}{d\theta}$$

- If likelihood is strong, prior will have little effect
  - mainly has influence on poorly constrained parameters
  - if a parameter is strongly constrained to be outside the typical range of the prior, then it will win over the prior

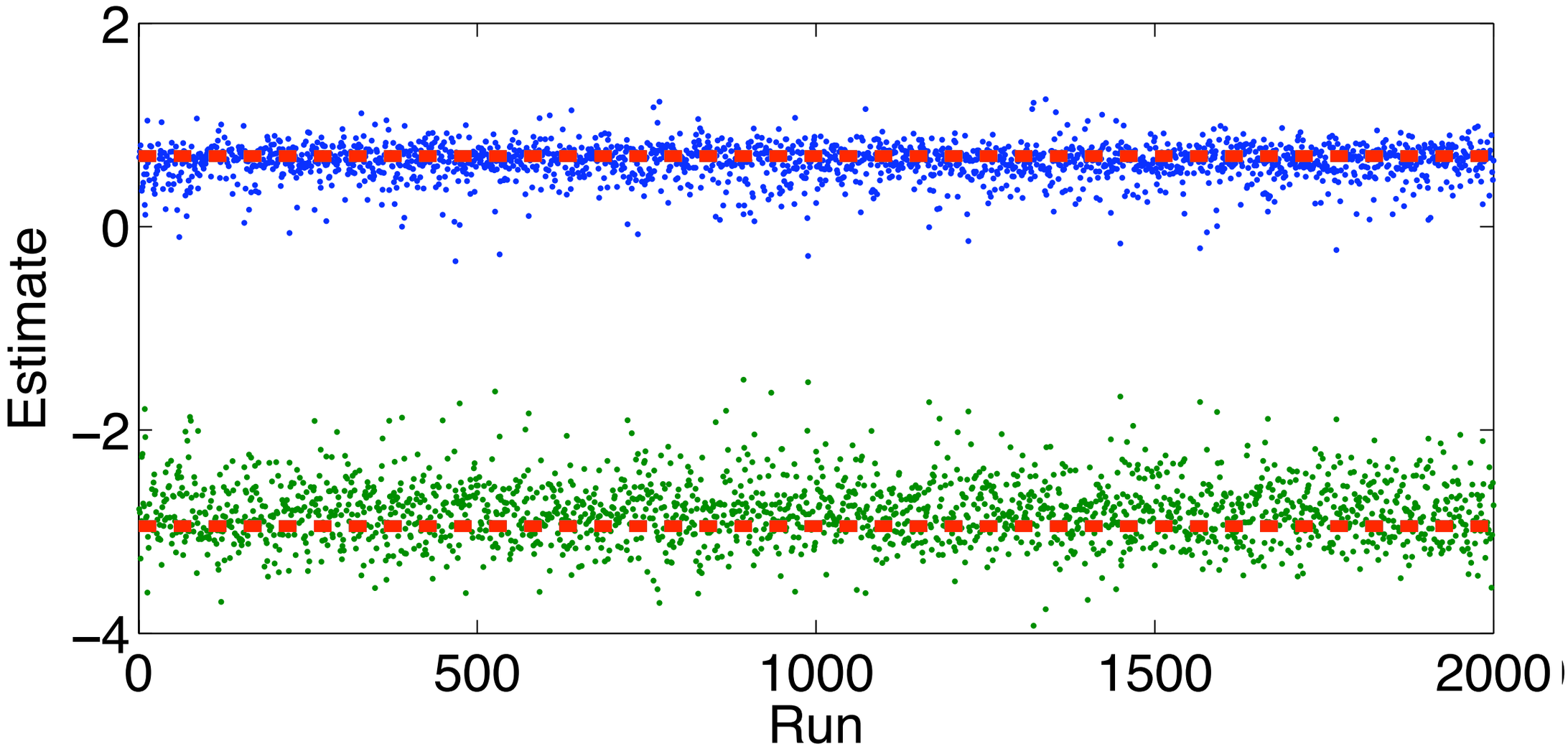
Maximum a posteriori estimate



200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2

m_{beta}=0, m_{eps}=-3, n=1

Maximum a posteriori estimate



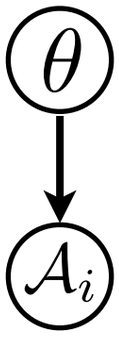
200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2

m_{beta}=0, m_{eps}=-3, n=1

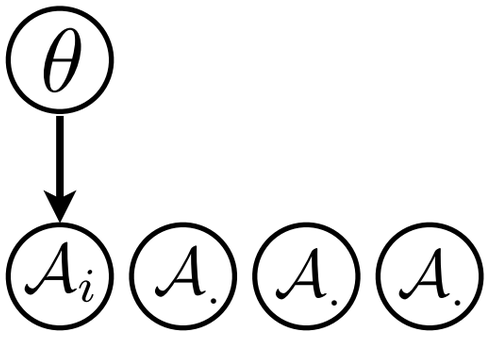
But

What prior parameters should I use?

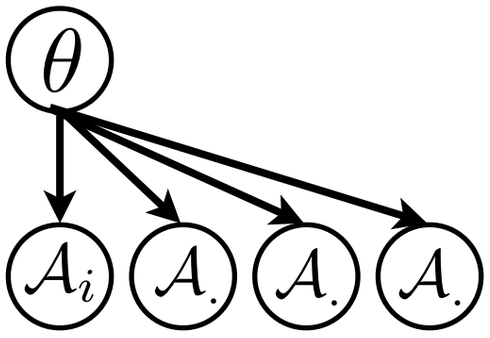
Hierarchical estimation - “random” effects



Hierarchical estimation - “random” effects

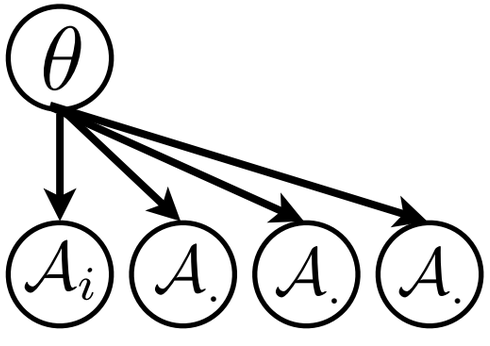


Hierarchical estimation - “random” effects



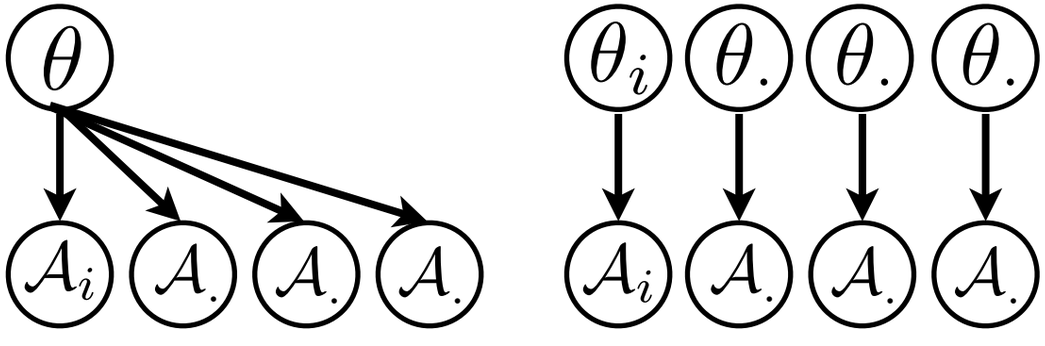
- Fixed effect
  - conflates within- and between- subject variability

Hierarchical estimation - “random” effects



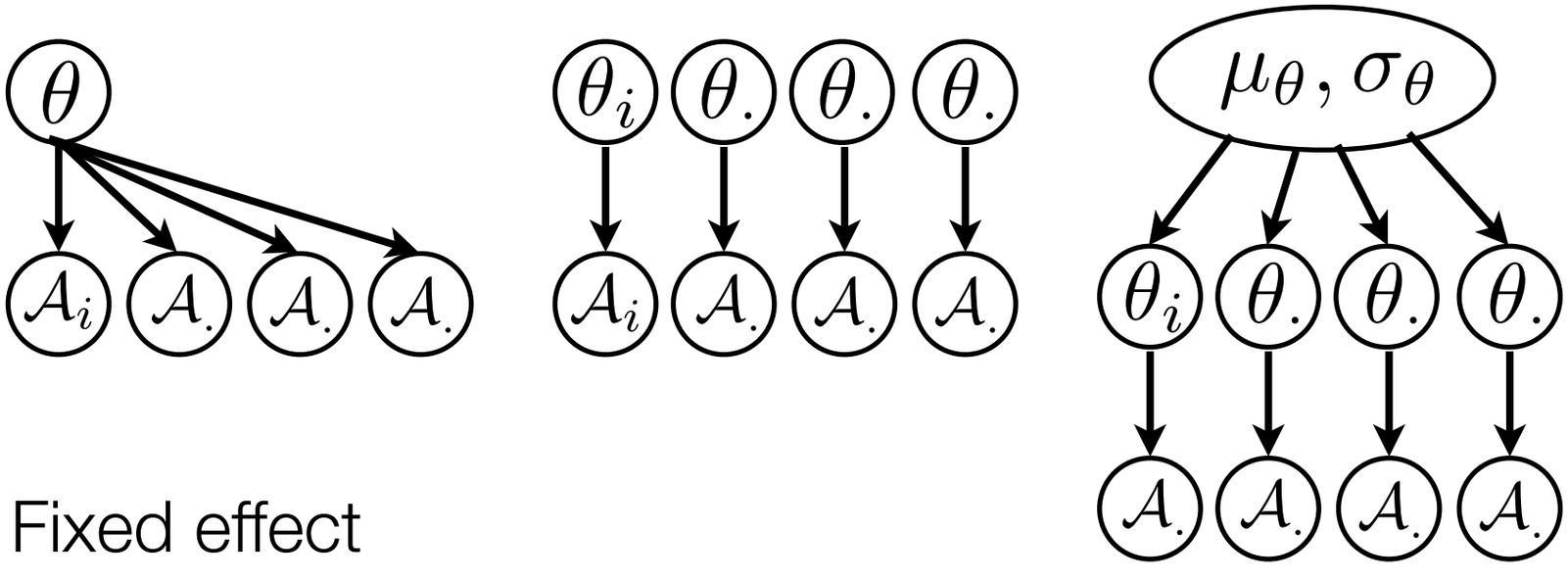
- Fixed effect
  - conflates within- and between- subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model

Hierarchical estimation - “random” effects



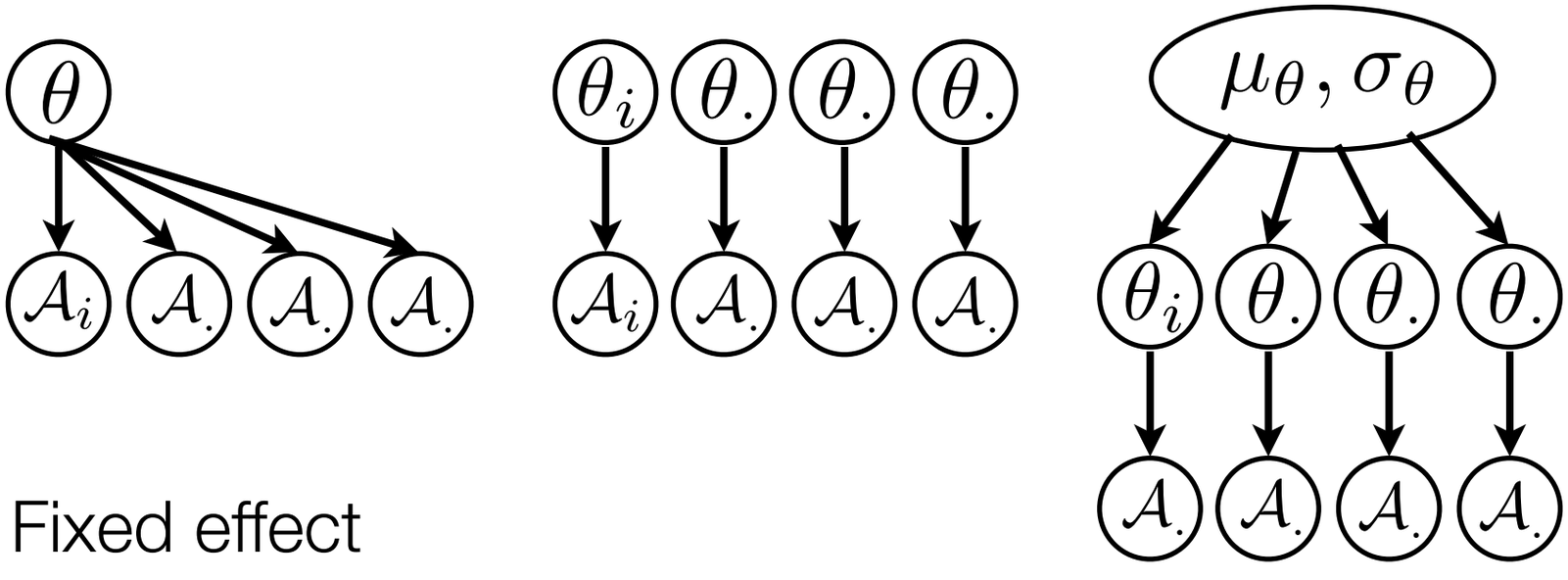
- Fixed effect
  - conflates within- and between- subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model
- Summary statistic
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy

Hierarchical estimation - “random” effects



- Fixed effect
  - conflates within- and between- subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model
- Summary statistic
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
- Random effects
  - prior mean = group mean

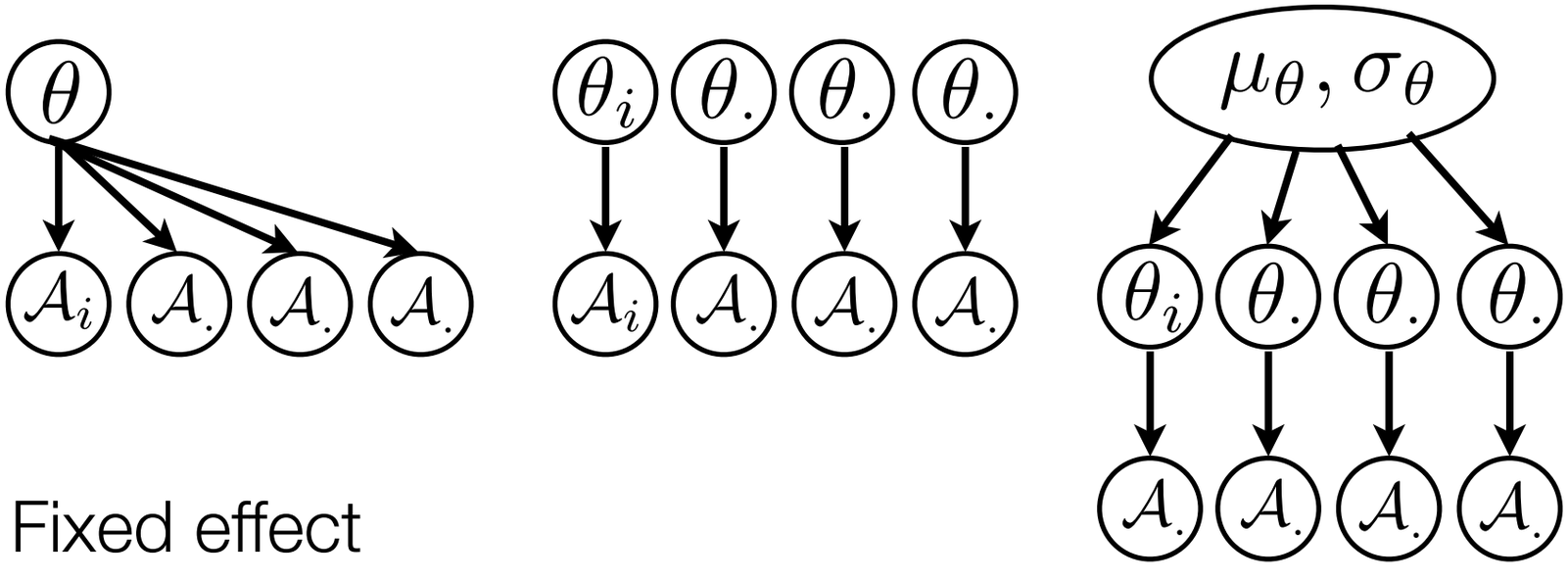
Hierarchical estimation - “random” effects



- Fixed effect
  - conflates within- and between- subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model
- Summary statistic
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
- Random effects
  - prior mean = group mean

$$p(\mathcal{A}_i|\mu_\theta, \sigma_\theta) = \int d\theta_i \, p(\mathcal{A}_i|\theta_i) \, p(\theta_i|\mu_\theta, \sigma_\theta)$$

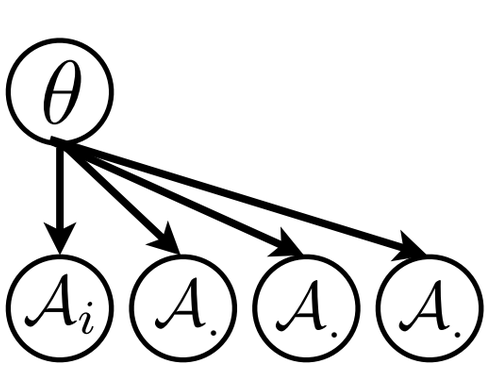
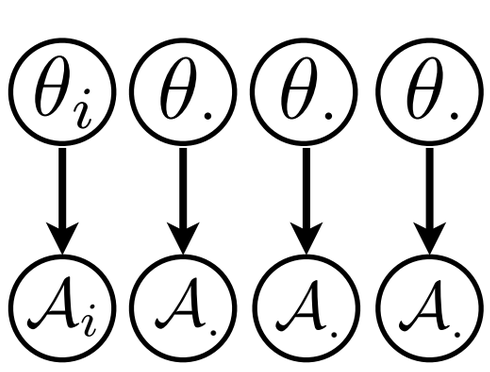
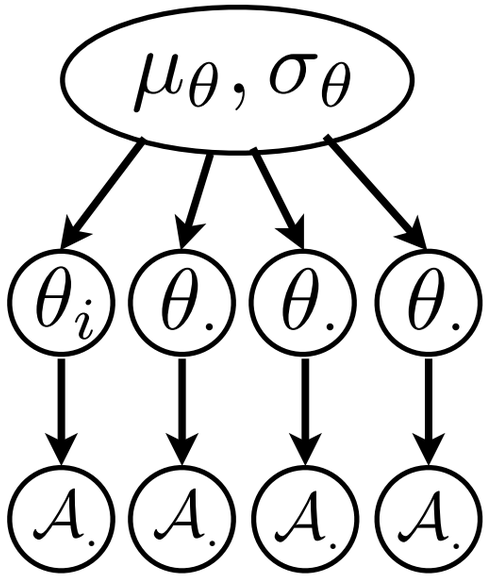
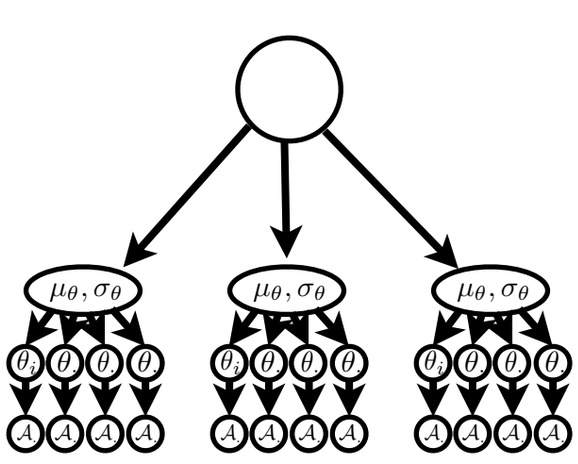
Hierarchical estimation - “random” effects



- Fixed effect
  - conflates within- and between- subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model
- Summary statistic
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
- Random effects
  - prior mean = group mean

$$p(\mathcal{A}_i|\mu_\theta, \sigma_\theta) = \int d\theta_i \, p(\mathcal{A}_i|\theta_i) \, p(\theta_i| \underbrace{\mu_\theta, \sigma_\theta}_{\zeta})$$

Hierarchical estimation - “random” effects

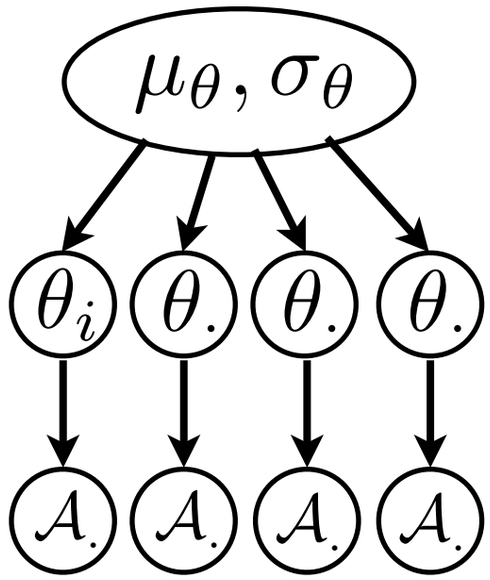
- Fixed effect
  - conflates within- and between- subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model
- Summary statistic
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
- Random effects
  - prior mean = group mean

$$p(\mathcal{A}_i|\mu_\theta, \sigma_\theta) = \int d\theta_i \, p(\mathcal{A}_i|\theta_i) \, p(\theta_i| \underbrace{\mu_\theta, \sigma_\theta}_{\zeta})$$

Random effects

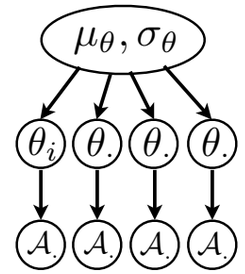
Random effects

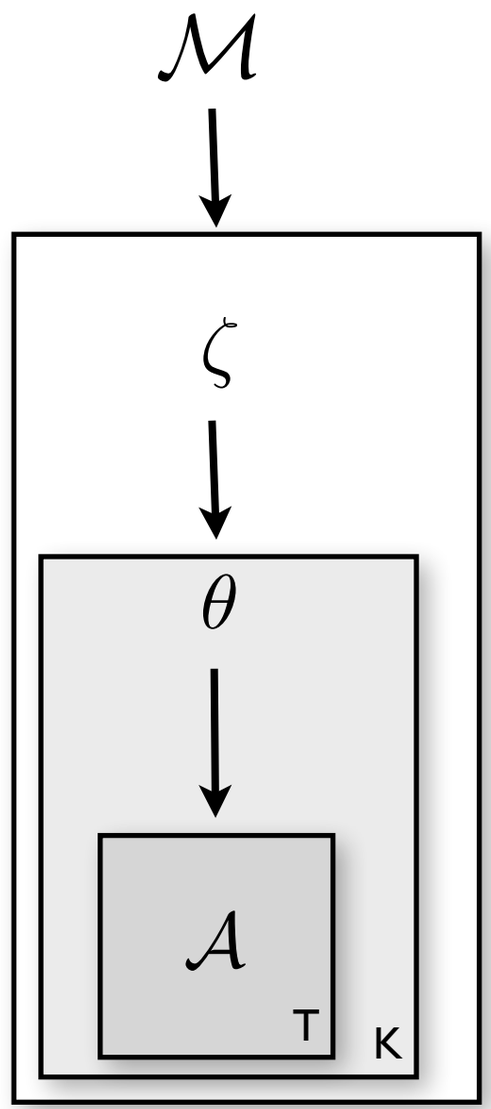
- See subjects as drawn from group
- Fixed models



Random effects

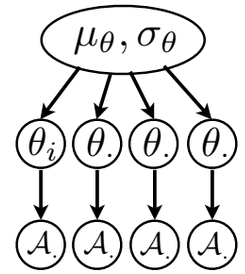
- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model

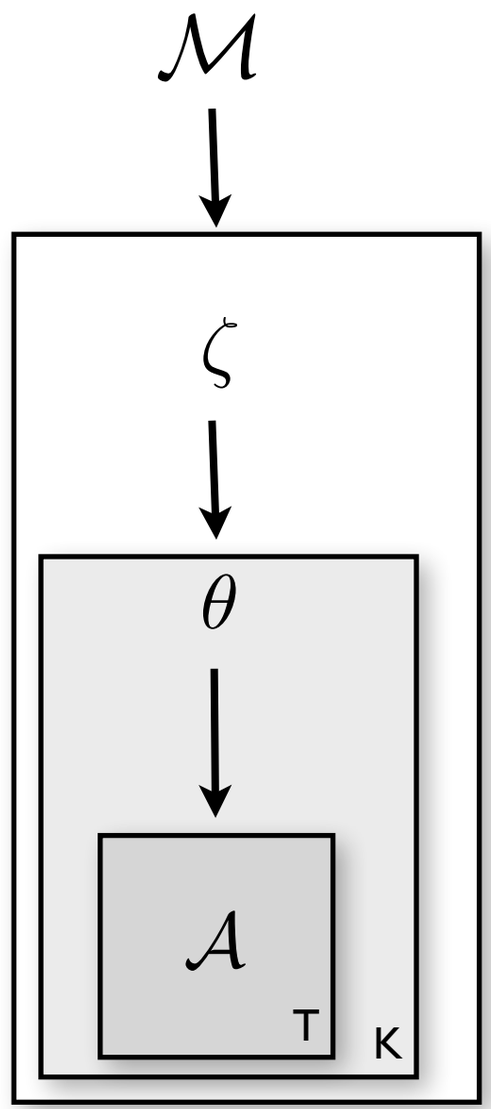




Random effects

- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested



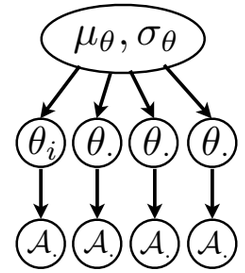


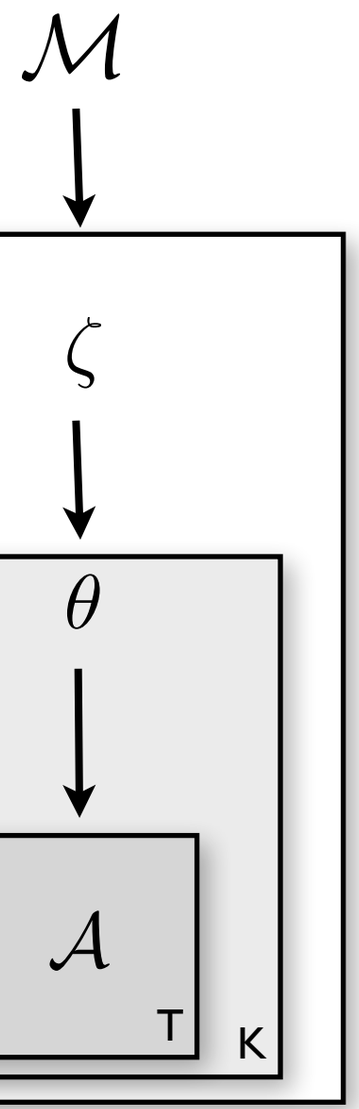
Random effects

- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$

  - assumes within-subject mixture, rather than a group mixture of perfect types

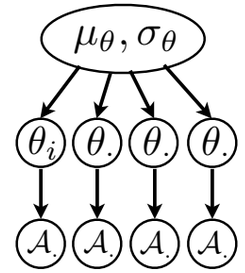




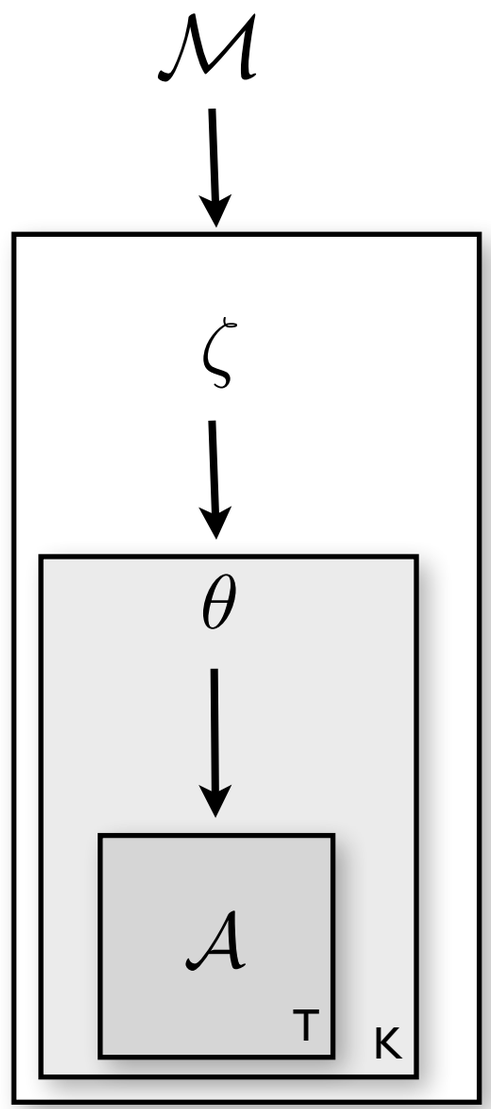
Random effects

- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$



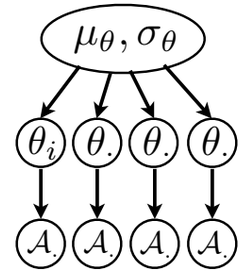
- assumes within-subject mixture, rather than a group mixture of perfect types
- w/in subject model comparison?

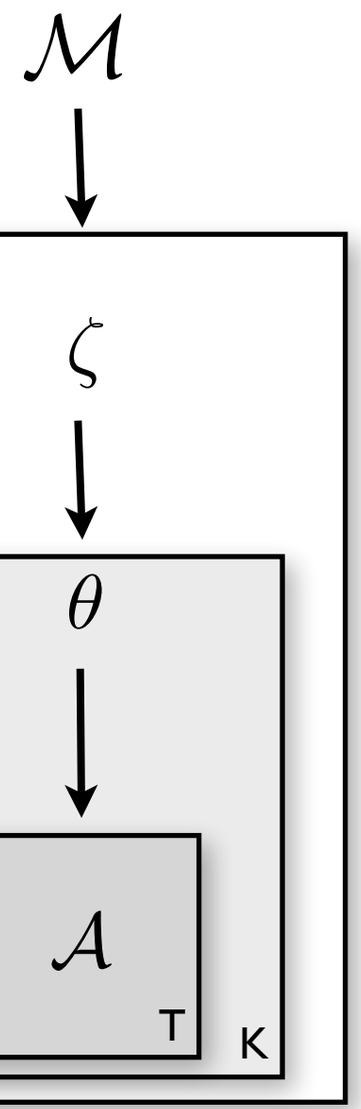


Random effects

- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$



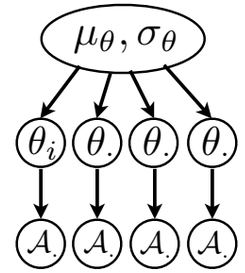


  - assumes within-subject mixture, rather than a group mixture of perfect types
  - w/in subject model comparison?
- HMM

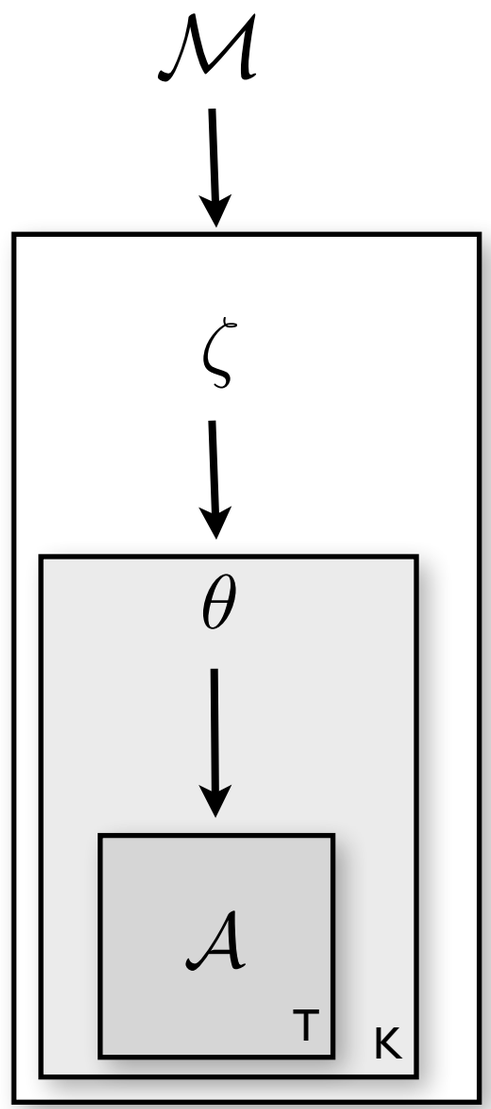
Random effects

- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

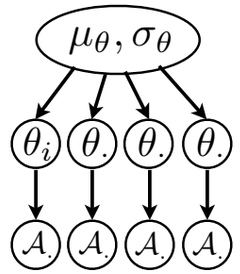
$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$



  - assumes within-subject mixture, rather than a group mixture of perfect types
  - w/in subject model comparison?
- HMM
  - switch between models over trials



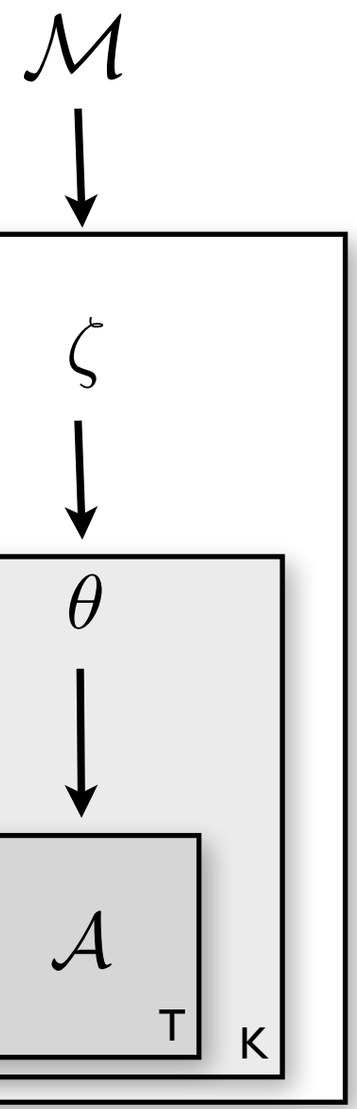
Random effects



- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$

    - assumes within-subject mixture, rather than a group mixture of perfect types
    - w/in subject model comparison?
  - HMM
    - switch between models over trials
- Random effects in models

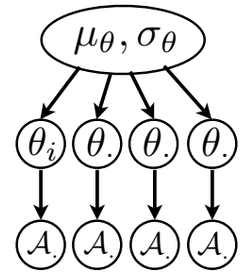


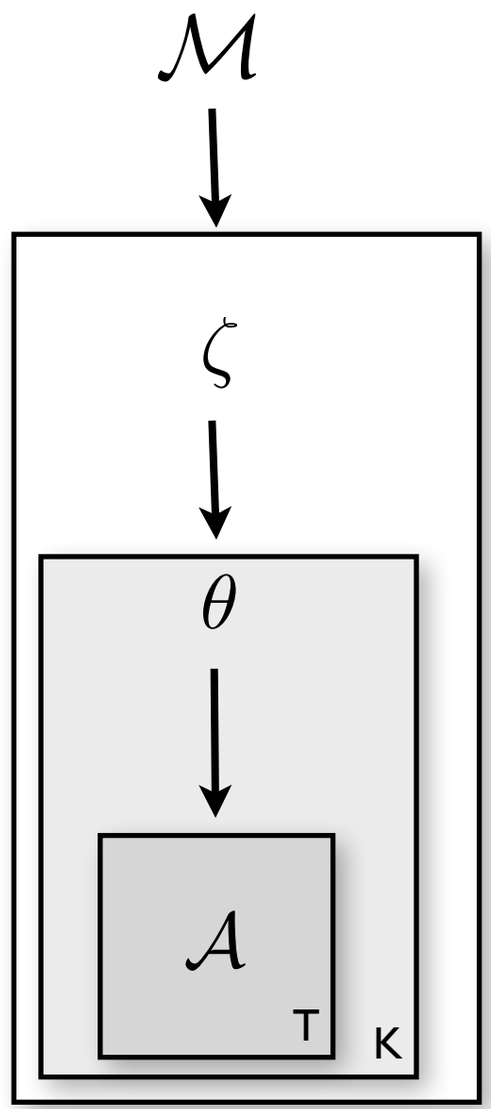
Random effects

- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$

    - assumes within-subject mixture, rather than a group mixture of perfect types
    - w/in subject model comparison?
  - HMM
    - switch between models over trials
- Random effects in models
  - Bayesian model averaging



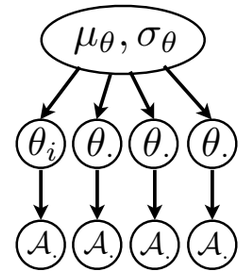


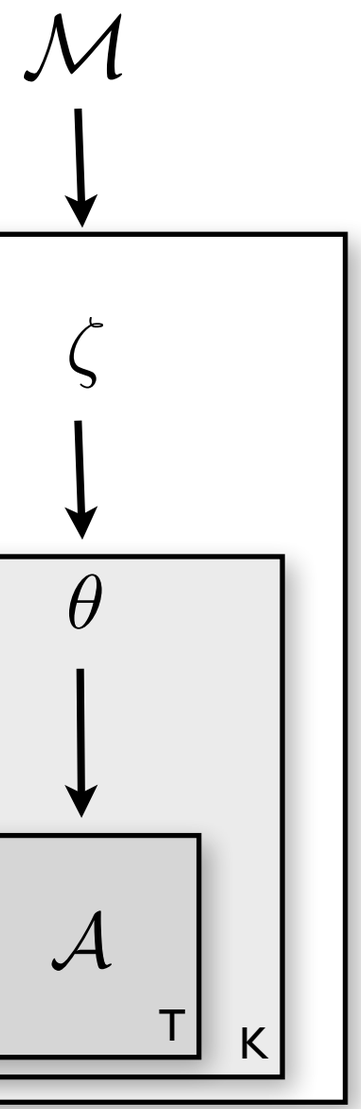
Random effects

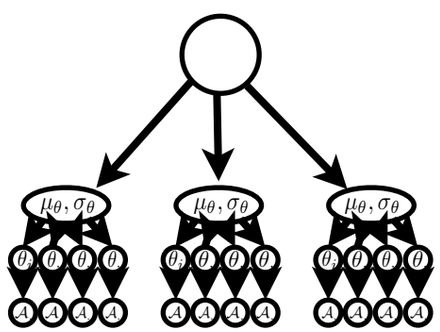
- See subjects as drawn from group
- Fixed models
  - all the same: fixed effect wrt model
  - parametrically nested

$$\mathcal{Q}(a,s) = \omega_1 \mathcal{Q}^1(a,s) + \omega_2 \mathcal{Q}^2(a,s)$$

    - assumes within-subject mixture, rather than a group mixture of perfect types
    - w/in subject model comparison?
  - HMM
    - switch between models over trials
- Random effects in models
  - Bayesian model averaging
  - parameter interpretation?







Estimating the hyperparameters

- Effectively we now want to do gradient ascent on:

$$\frac{d}{d\zeta} p(\mathcal{A}|\zeta)$$

- But this contains an integral over individual parameters:

$$p(\mathcal{A}|\zeta) = \int d\theta p(\mathcal{A}|\theta)\, p(\theta|\zeta)$$

- So we need to:

$$\begin{aligned} \hat{\zeta} &= \underset{\zeta}{\operatorname{argmax}}\, p(\mathcal{A}|\zeta) \\ &= \underset{\zeta}{\operatorname{argmax}} \int d\,\theta p(\mathcal{A}|\theta)\, p(\theta|\zeta) \end{aligned}$$

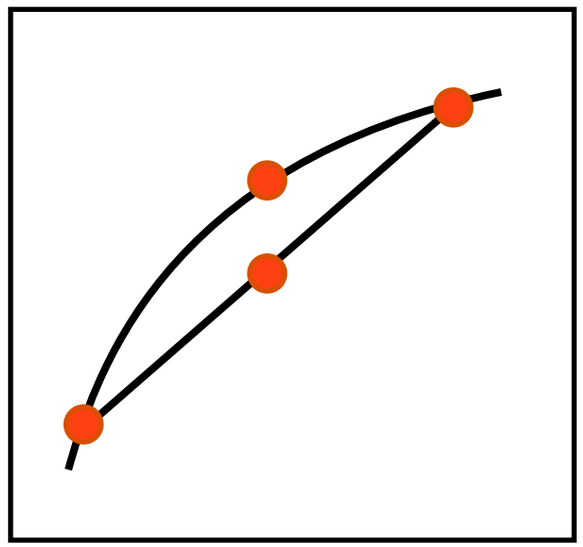
Inference

$$\begin{aligned}\hat{\zeta} &= \underset{\zeta}{\operatorname{argmax}}\, p(\mathcal{A}|\zeta) \\ &= \underset{\zeta}{\operatorname{argmax}} \int d\,\theta p(\mathcal{A}|\theta)\, p(\theta|\zeta)\end{aligned}$$

- analytical - rare
- brute force - for simple problems
- Expectation Maximisation - approximate, easy
- Variational Bayes - approximate, often hard
- Sampling / MCMC - slow, easy

Expectation Maximisation

$$\begin{aligned}
\log p(\mathcal{A}|\zeta) &= \log \int d\theta\, p(\mathcal{A}, \theta|\zeta) \\
&= \log \int d\theta\, q(\theta) \frac{p(\mathcal{A}, \theta|\zeta)}{q(\theta)} \\
&\geq \int d\theta\, q(\theta) \log \frac{p(\mathcal{A}, \theta|\zeta)}{q(\theta)}
\end{aligned}$$



Jensen's inequality

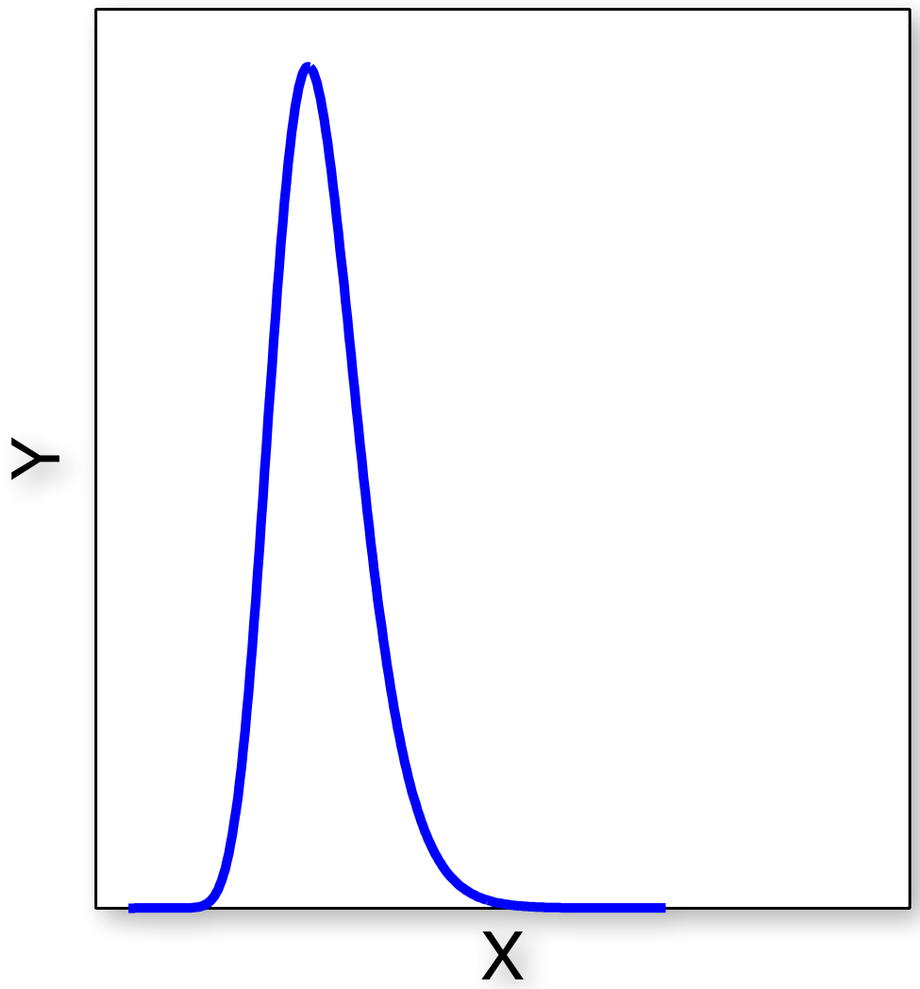
$$k^{\text{th}} \text{ E step: } q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$$

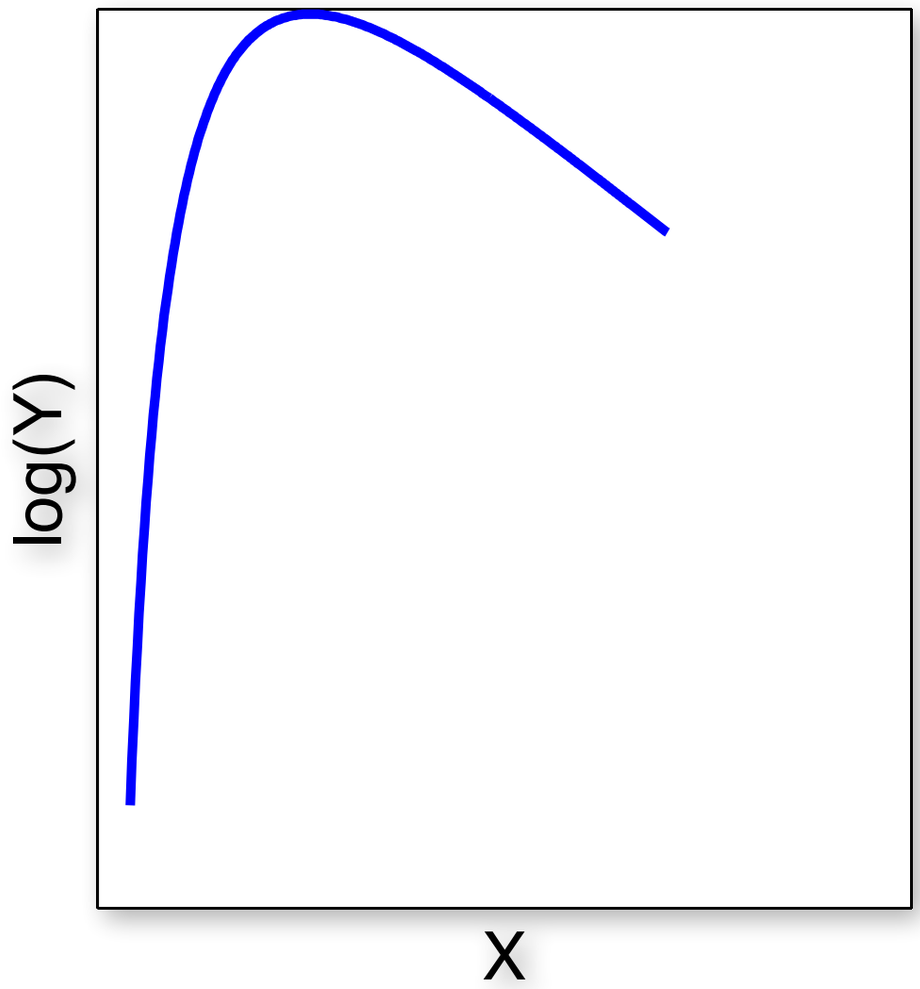
$$k^{\text{th}} \text{ M step: } \zeta^{(k+1)} \leftarrow \underset{\zeta}{\text{argmax}} \int d\theta\, q(\theta) \log p(\mathcal{A}, \theta|\zeta)$$

- Iterate between
  - Estimating MAP parameters given prior parameters
  - Estimating prior parameters from MAP parameters

Laplace's approximation

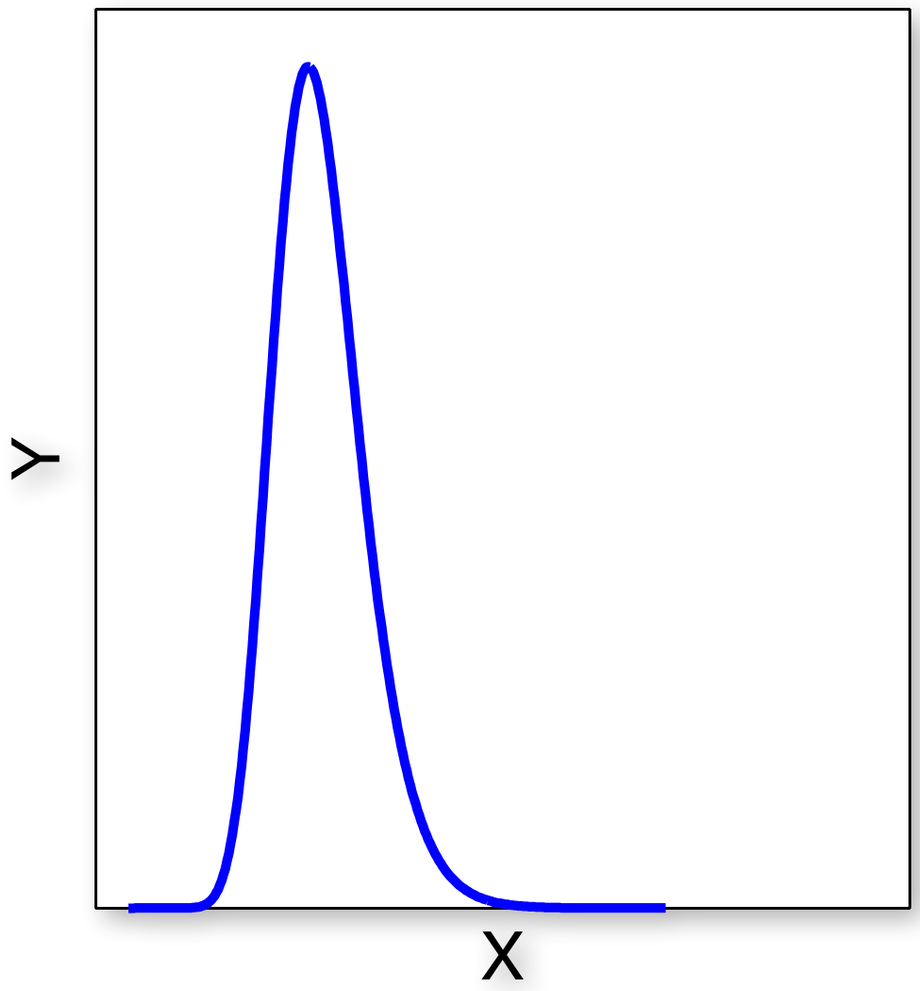
- Problem: $\text{E step:} \quad q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$

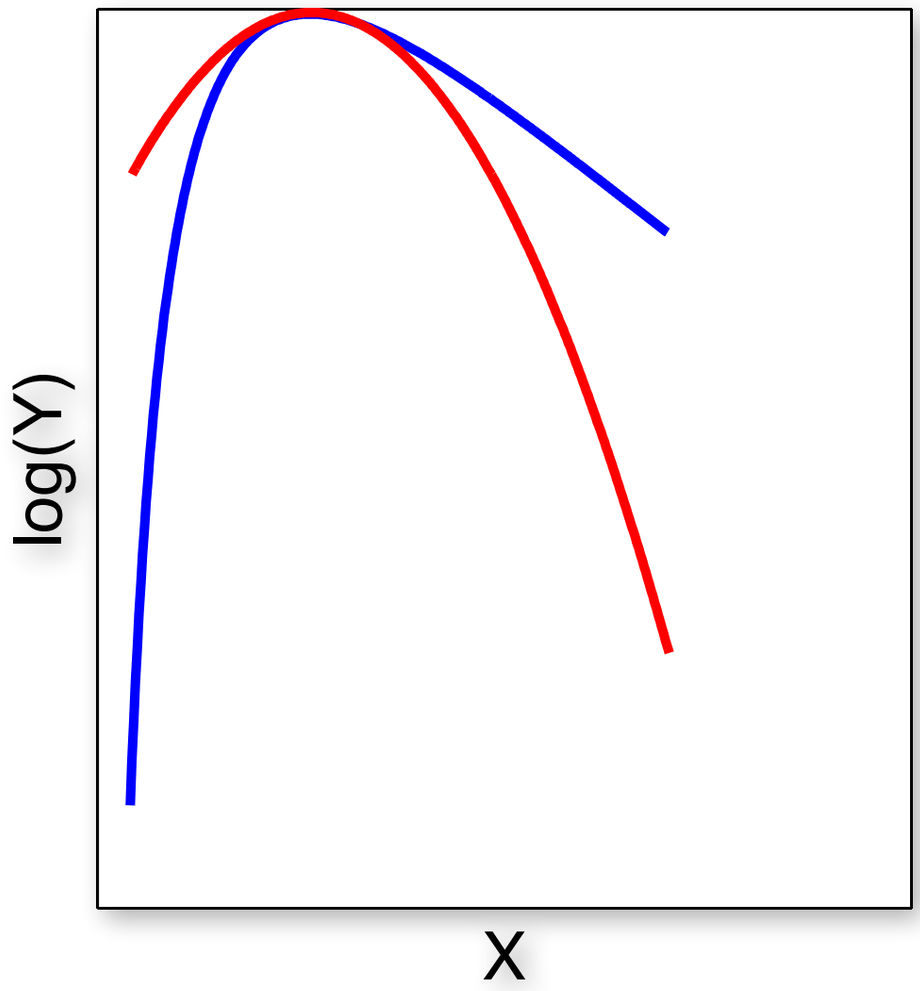




Laplace's approximation

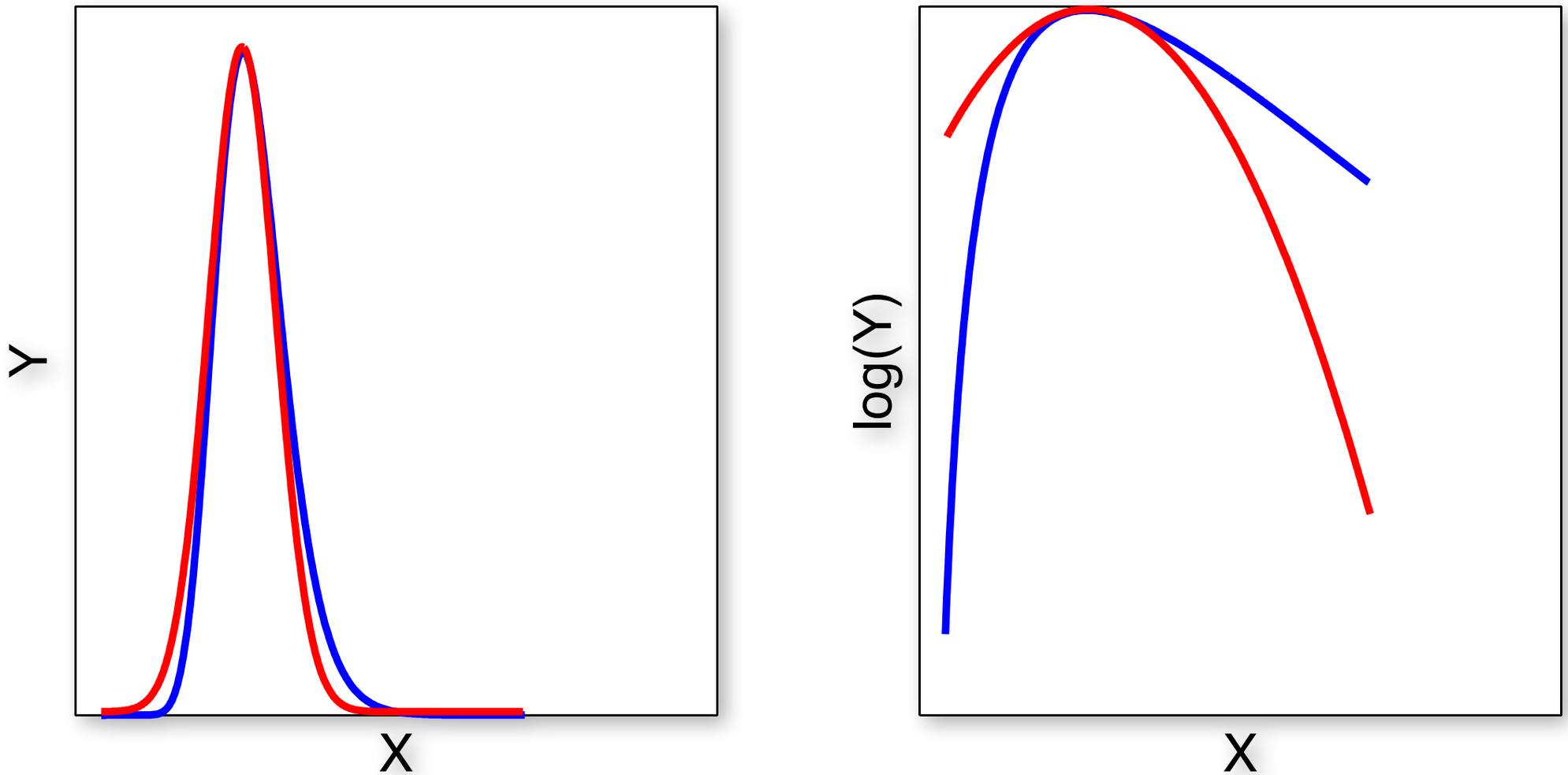
- Problem: E step: $q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$





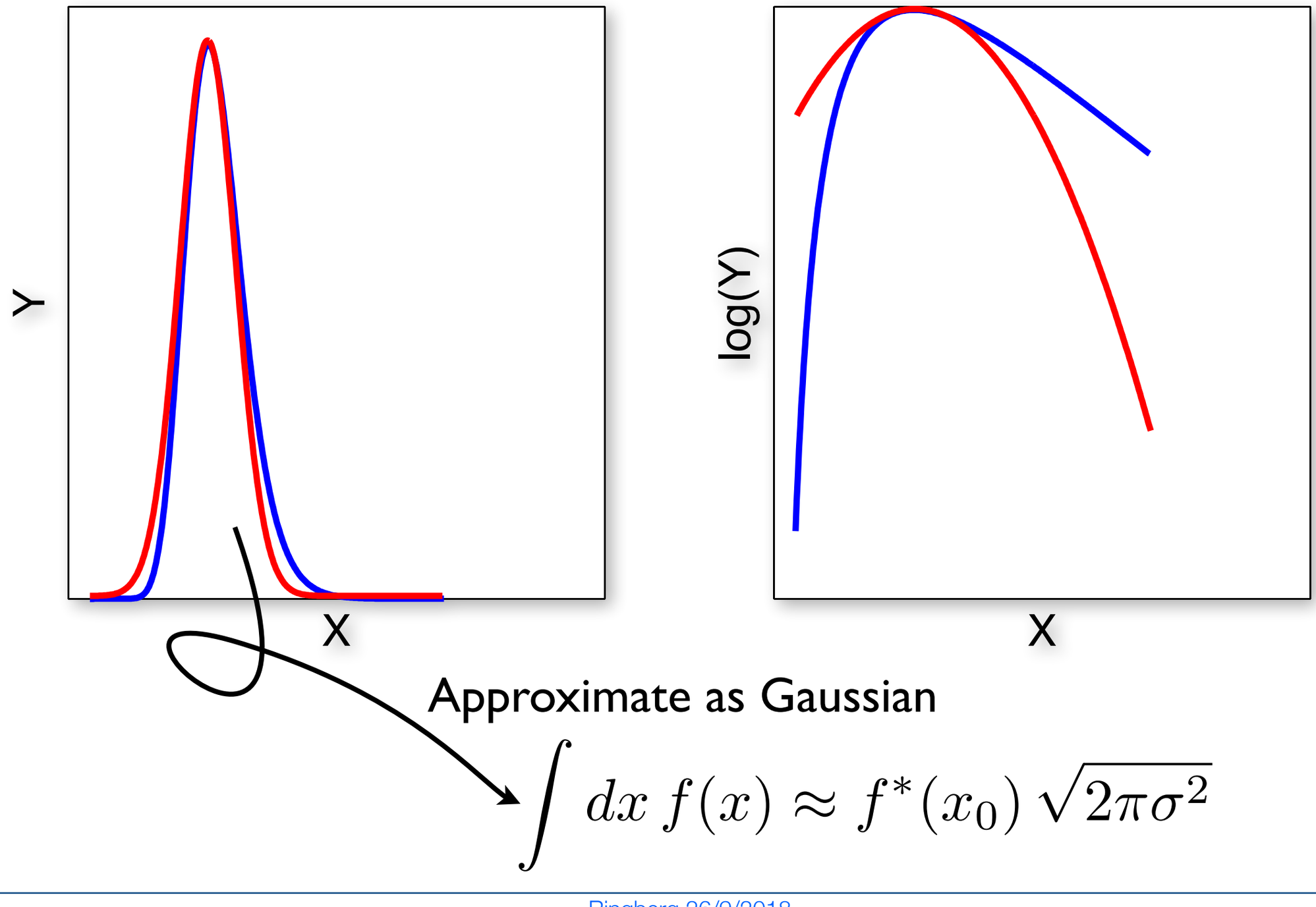
Laplace's approximation

- Problem: $\mathrm{E\ step:} \quad q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$



Laplace's approximation

- Problem: $\mathrm{E\ step:} \quad q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$



EM with Laplace approximation

- E step: $q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$
  - only need sufficient statistics to perform M step
  - Approximate $p(\theta|\mathcal{A}, \zeta^{(k)}) \sim \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k)$
  - and hence:

$$\text{E step:} \quad q_k(\theta) \quad = \mathcal{N}(\mathbf{m}_k, \mathbf{S}_k)$$

$$\mathbf{m}_k \quad \leftarrow \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{a}_k|\theta)p(\theta|\zeta^{(i)})$$

$$\mathbf{S}_k^{-1} \quad \leftarrow \left.\frac{\partial^2 p(\mathbf{a}^k|\theta)p(\theta|\zeta^{(i)})}{\partial\theta^2}\right|_{\theta=\mathbf{m}_k}$$

matlab: `[m,L,,,S]=fminunc(…)`

Just what we had before: MAP inference given some prior parameters

EM with Laplace approximation

- Updates

Prior mean = mean of MAP estimates

$$\text{M step:} \qquad \zeta_{\mu}^{(i+1)} = \frac{1}{K}\sum_{k}\mathbf{m}_k$$

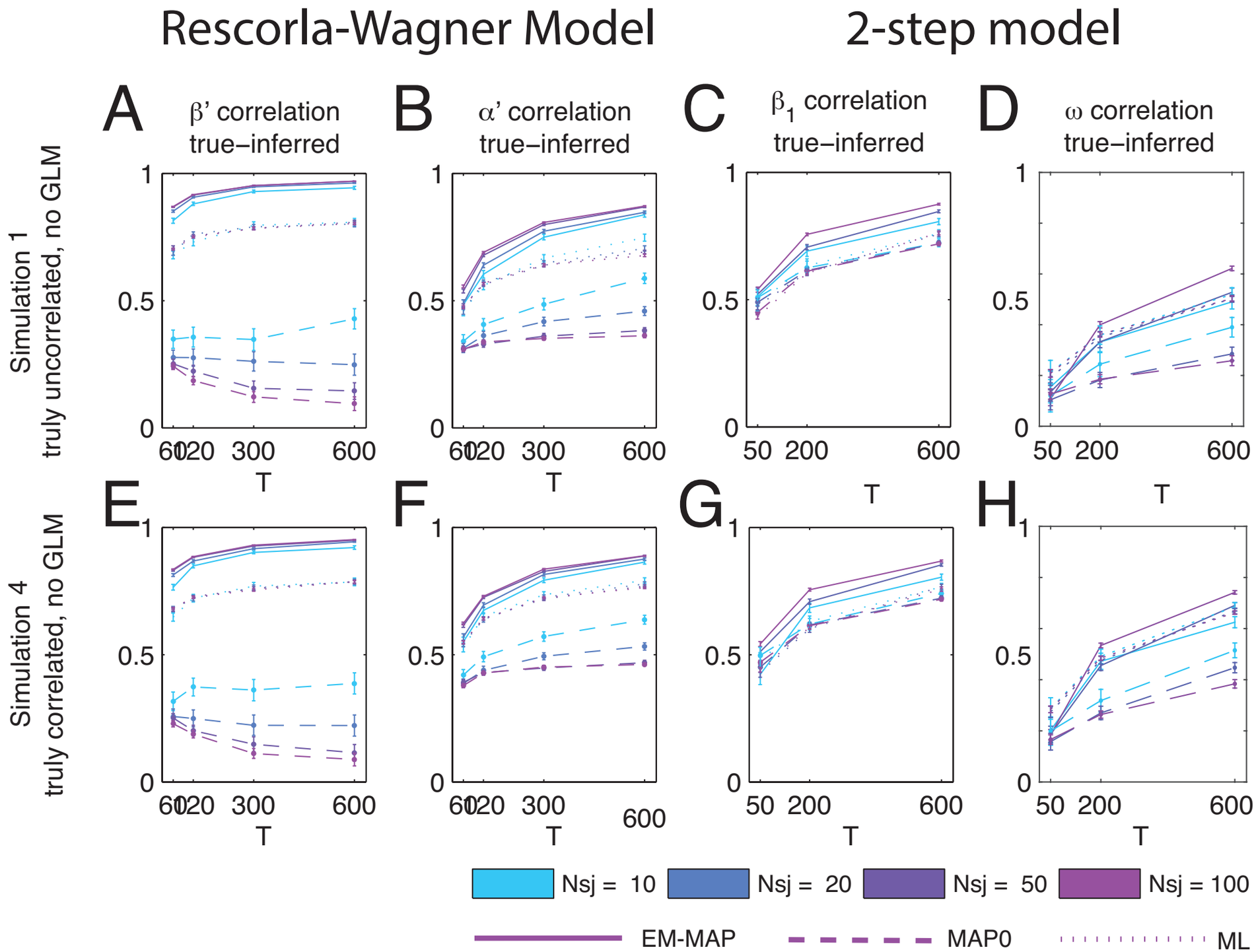
$$\zeta_{\nu^2}^{(i+1)} = \frac{1}{N}\sum_{i}\left[(\mathbf{m}_k)^2 + \mathbf{S}_k\right] - (\zeta_{\mu}^{(i+1)})^2$$

Prior variance depends on inverse Hessian S and variance of MAP estimates

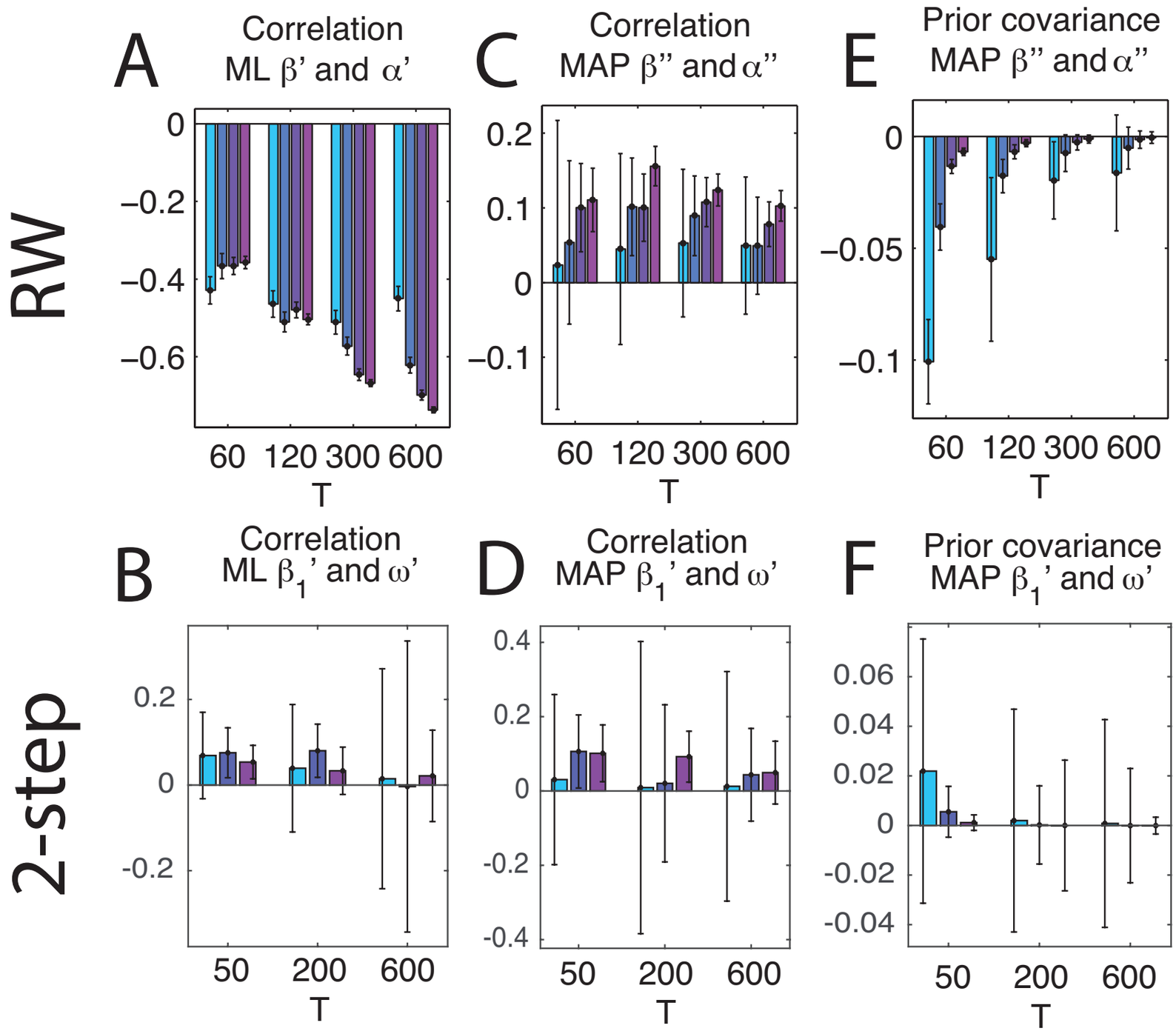
Take uncertainty of estimates into account

- And now iterate until convergence

Parameter recovery

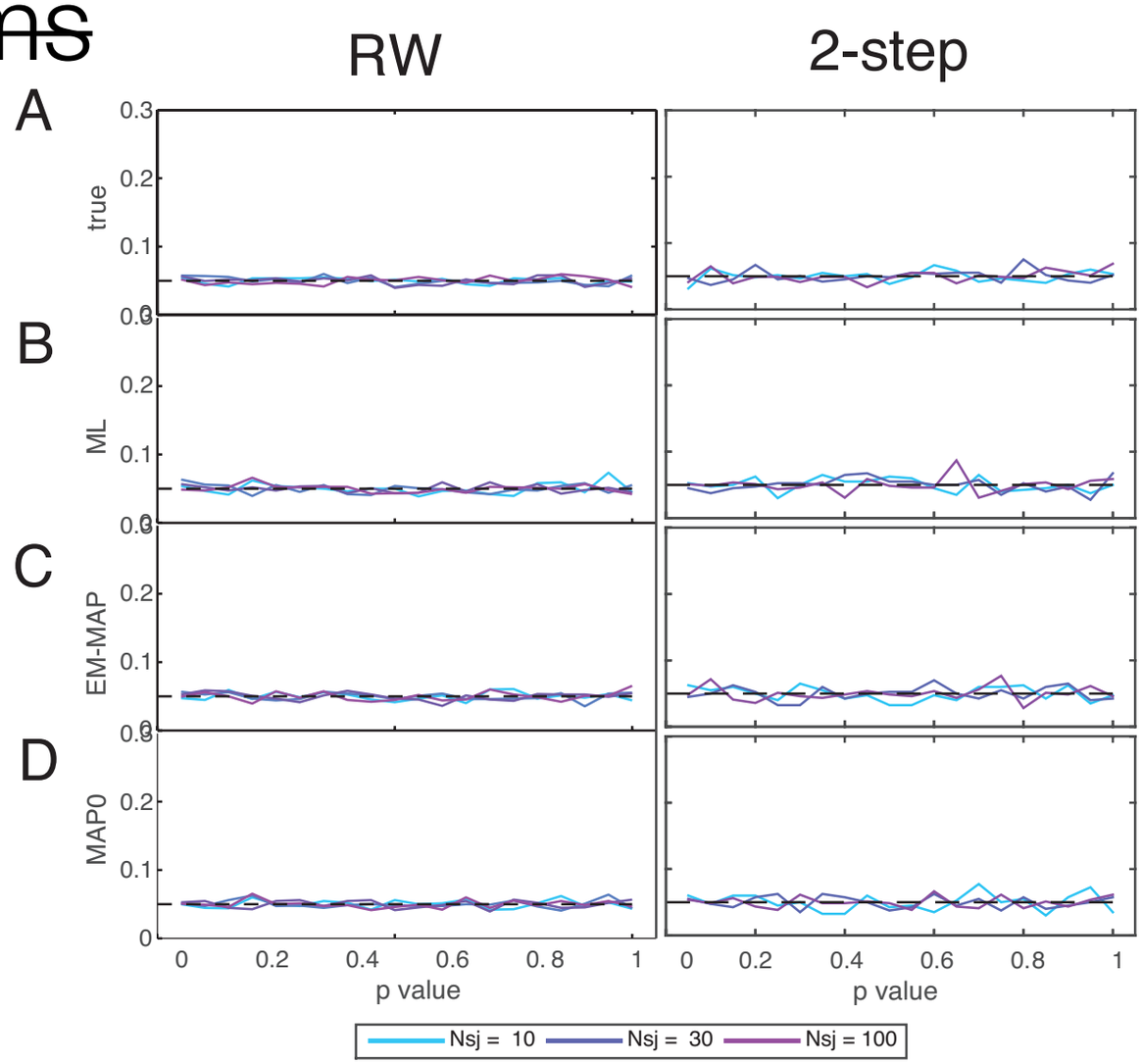


Correlations



Are parameters ok for correlations?

- Individual subject parameter estimates NO LONGER INDEPENDENT!
  - Change group -> change parameter estimates
- ~~compare different params~~
  - if different priors
- correlations, t-tests
  - within same prior ok
  - more power than ML



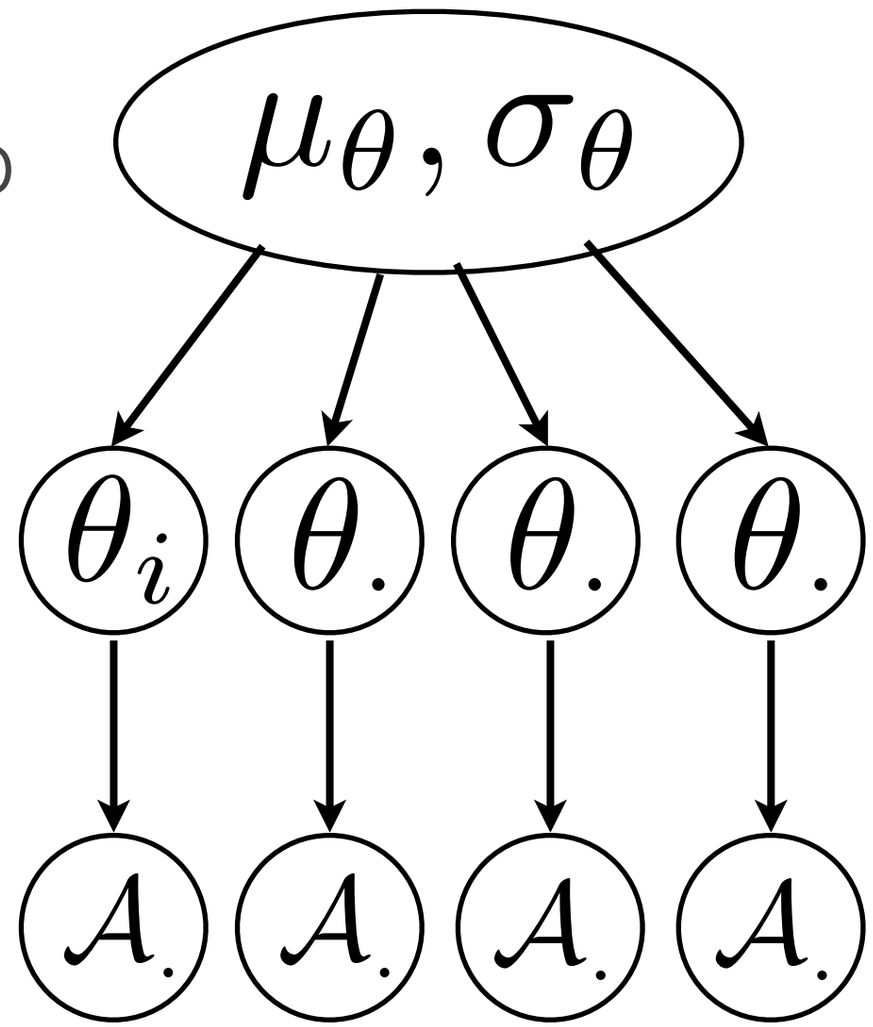
GLM

GLM

- So far
  - infer individual parameters
  - apply standard tests

GLM

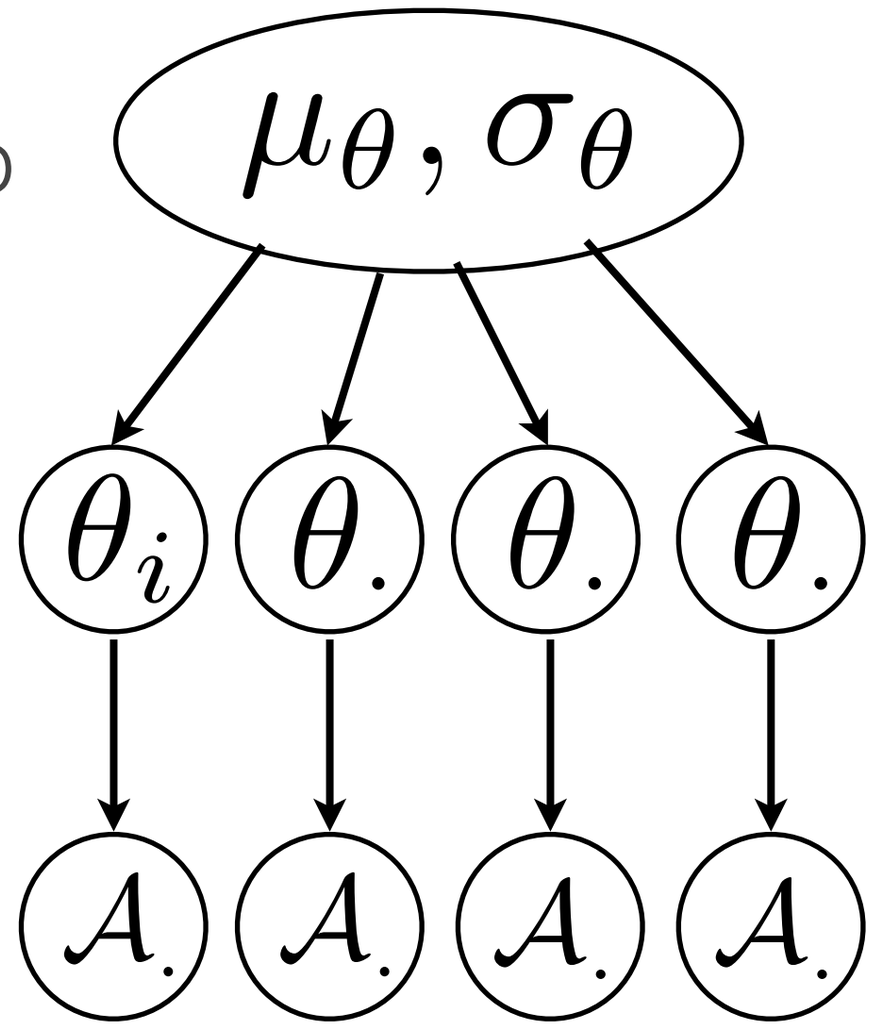
- So far
  - infer individual parameters
  - apply standard tests
- Alternative
  - View as variation across group
  - Specific - more powerful?



GLM

- So far
  - infer individual parameters
  - apply standard tests
- Alternative
  - View as variation across group
  - Specific - more powerful?

$$\mu_\theta^i = \mu_\theta^{\text{Group}} + \beta\psi_i$$

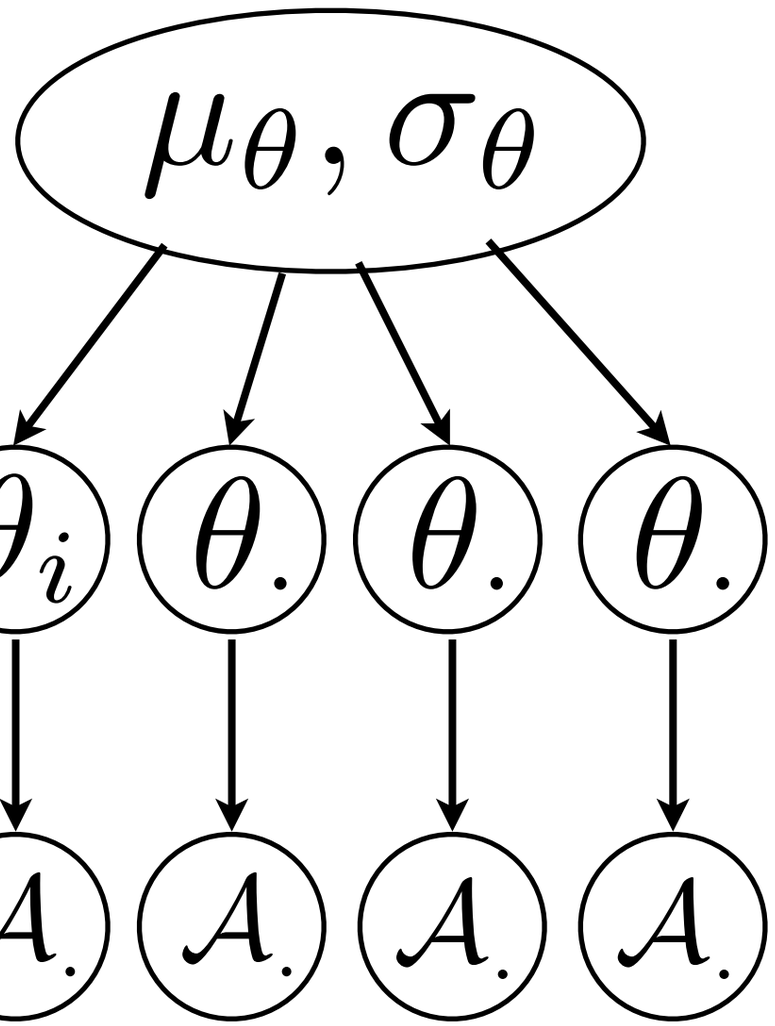


GLM

- So far
  - infer individual parameters
  - apply standard tests
- Alternative
  - View as variation across group
  - Specific - more powerful?

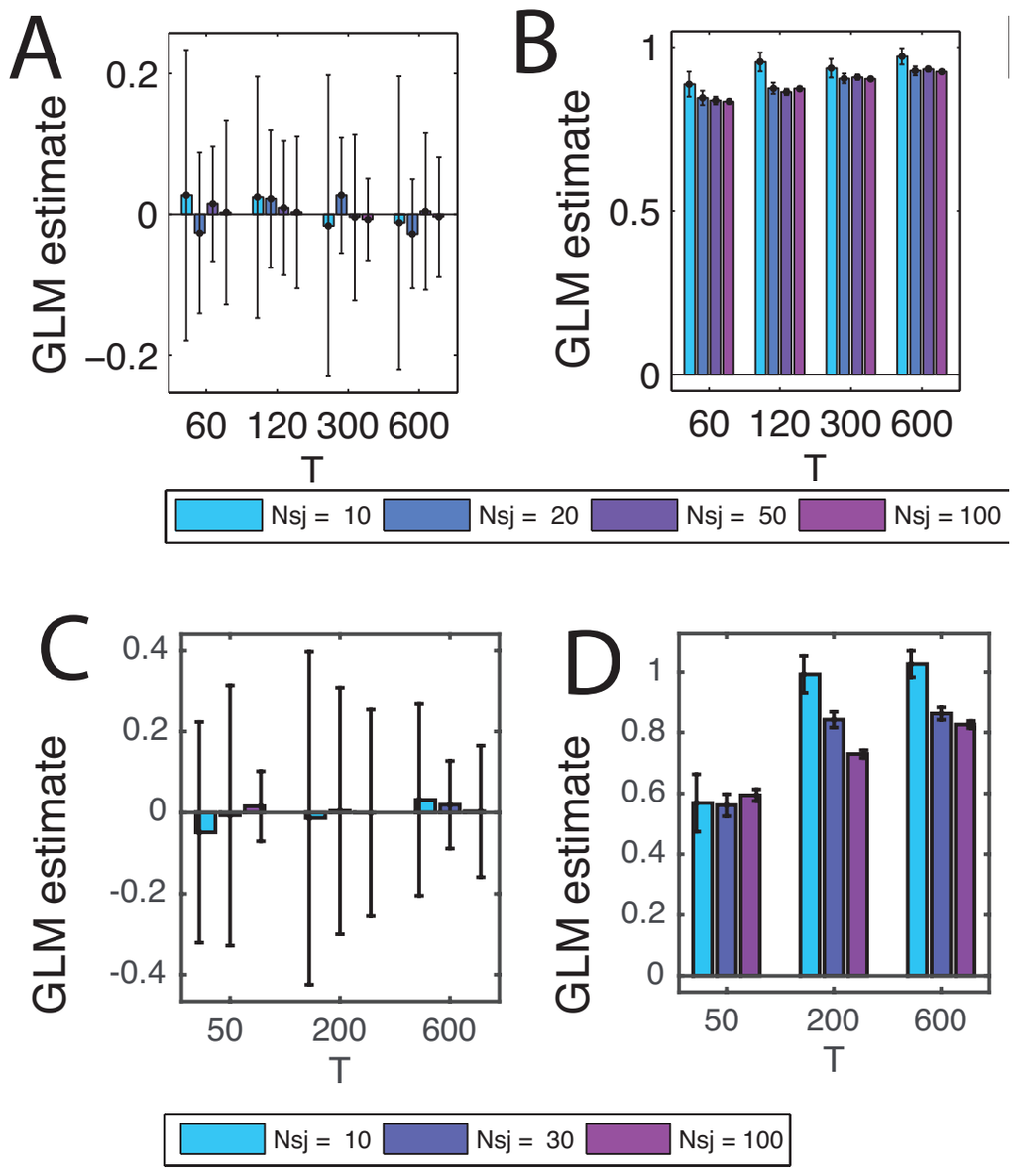
$$\mu_\theta^i = \mu_\theta^{\text{Group}} + \beta\psi_i$$

Infer



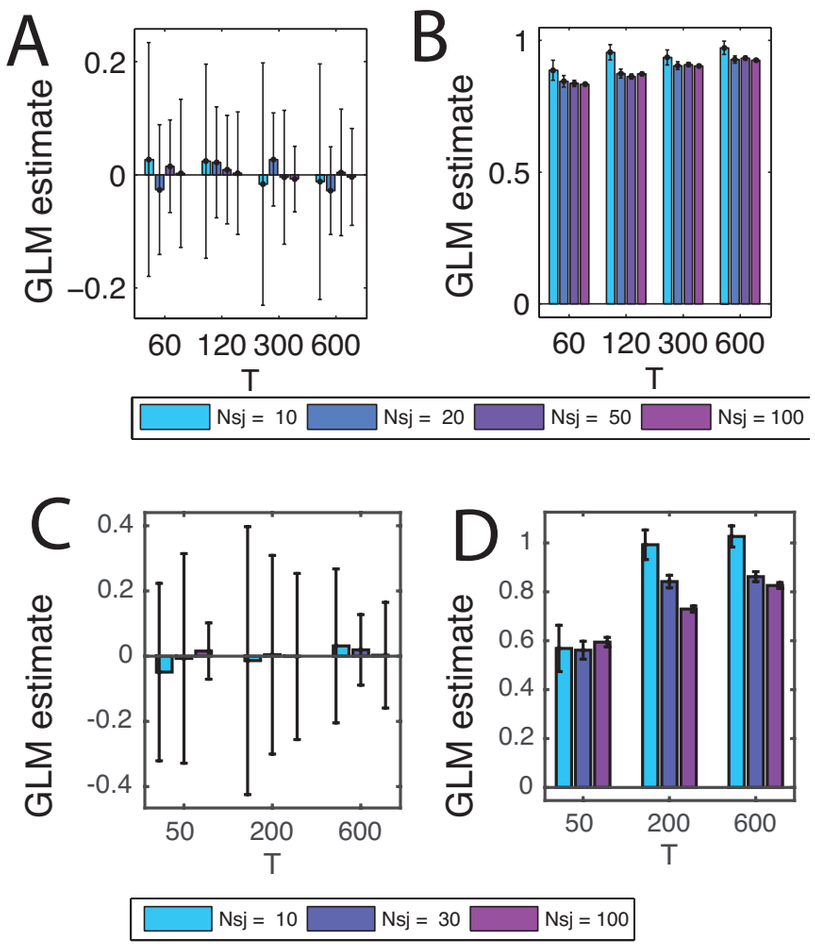
GLM

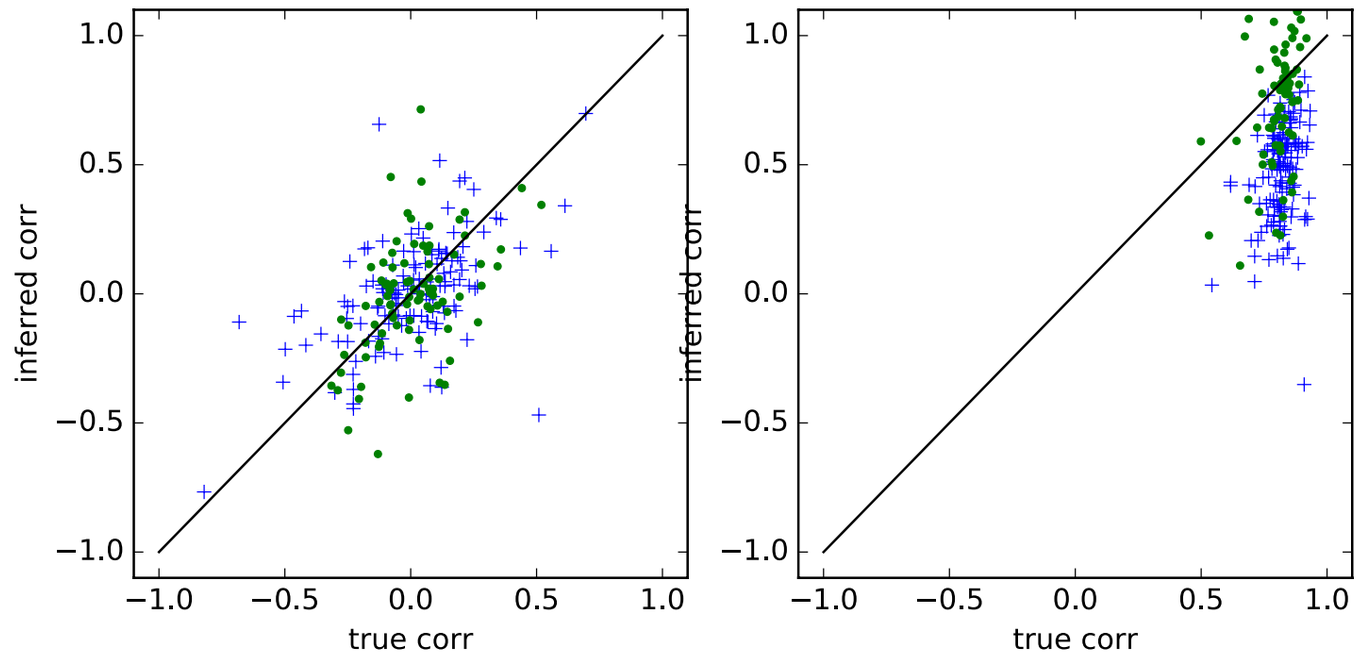
- Group-level regressor



GLM

- Group-level regressor





Group error bars

- EM standardly does not provide error bars
- Shifted samples

$$p(\mathcal{A}|\zeta) \approx \sum_i p(\mathcal{A}|\theta_i); \qquad \theta_i \sim p(\theta|\zeta) \tag{1}$$

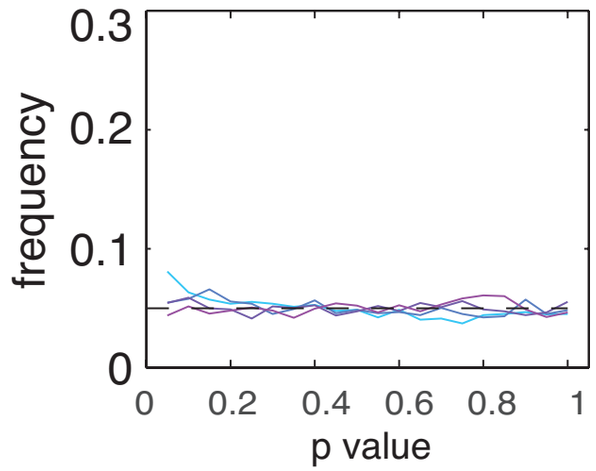
$$\frac{\partial}{\partial \mu} p(\mathcal{A}|\zeta) \approx \frac{1}{\delta}\left[\sum_i p(\mathcal{A}|\theta_i + \delta) - \sum_i p(\mathcal{A}|\theta_i)\right] \tag{2}$$

- Oakes 1999
  - analytical description of gradients
  - tricky, but combined with forward differentiation it is automatic (julialang.org Pkg ForwardDiff)

GLM error bars

Shifted samples

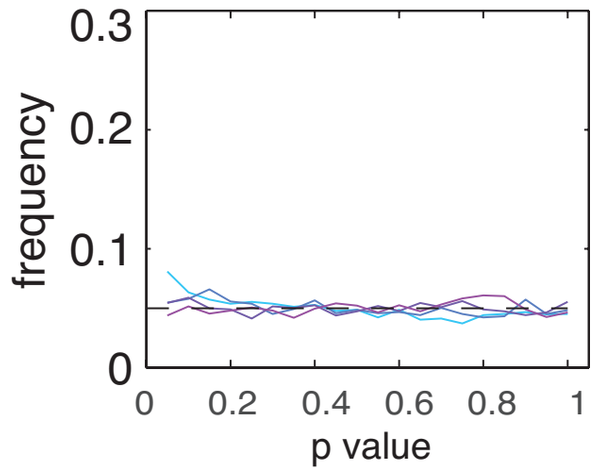
RW



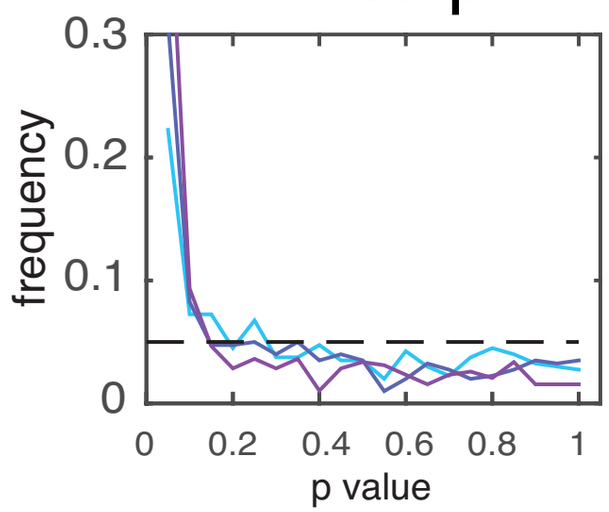
GLM error bars

Shifted samples

RW



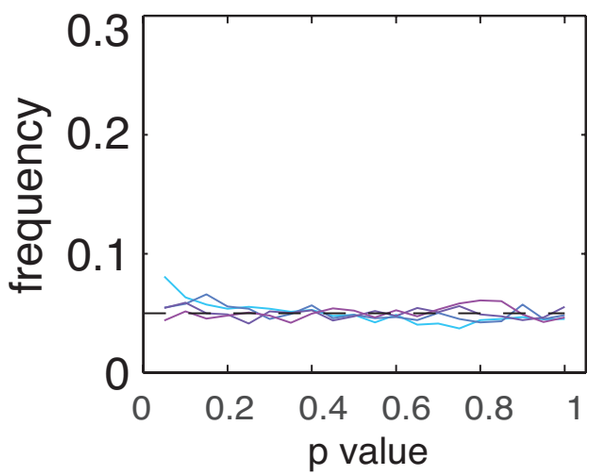
Twostep



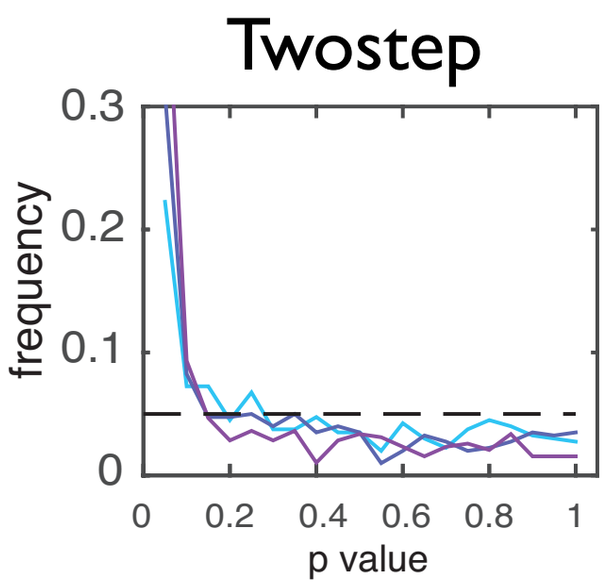
GLM error bars

Shifted samples

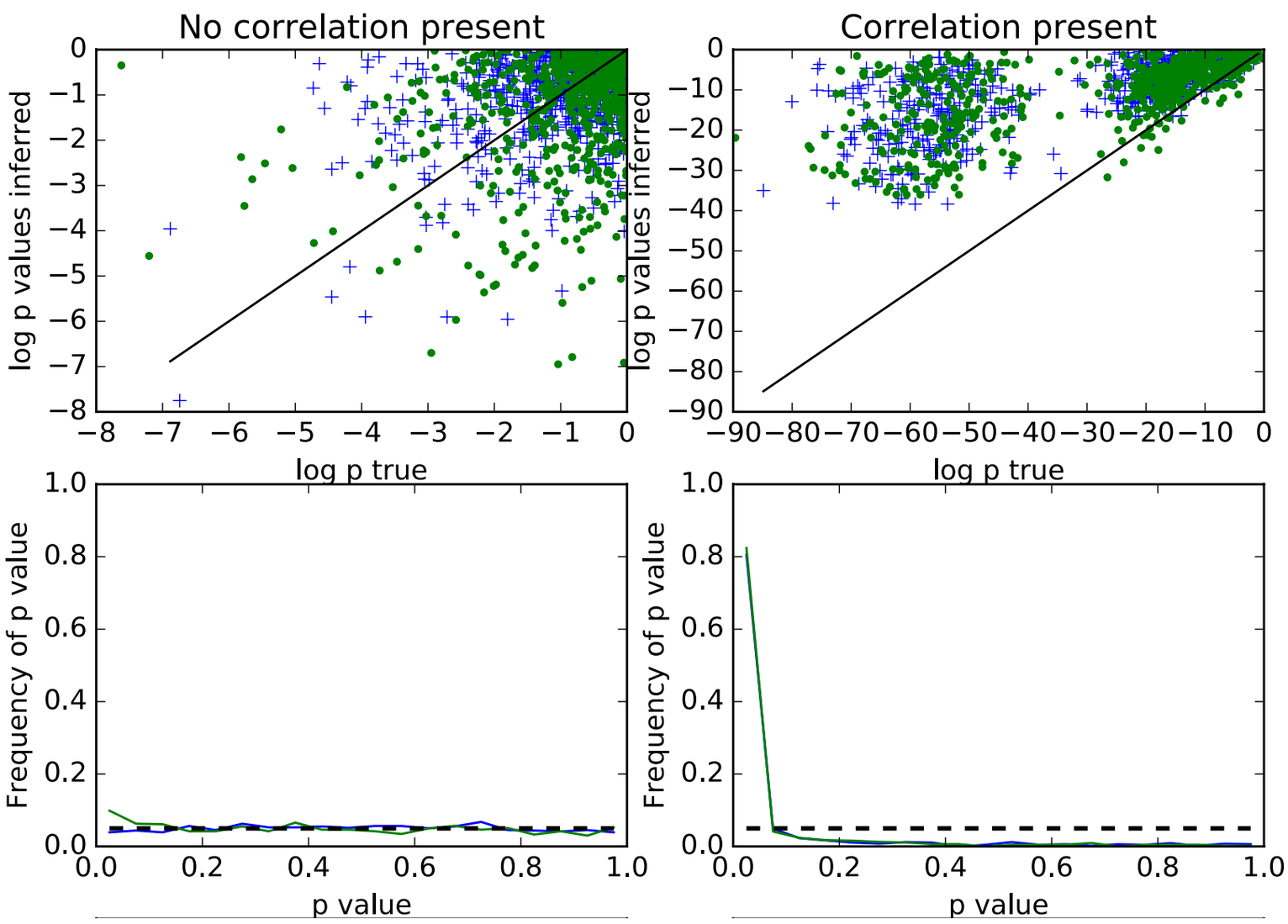
RW



Twostep



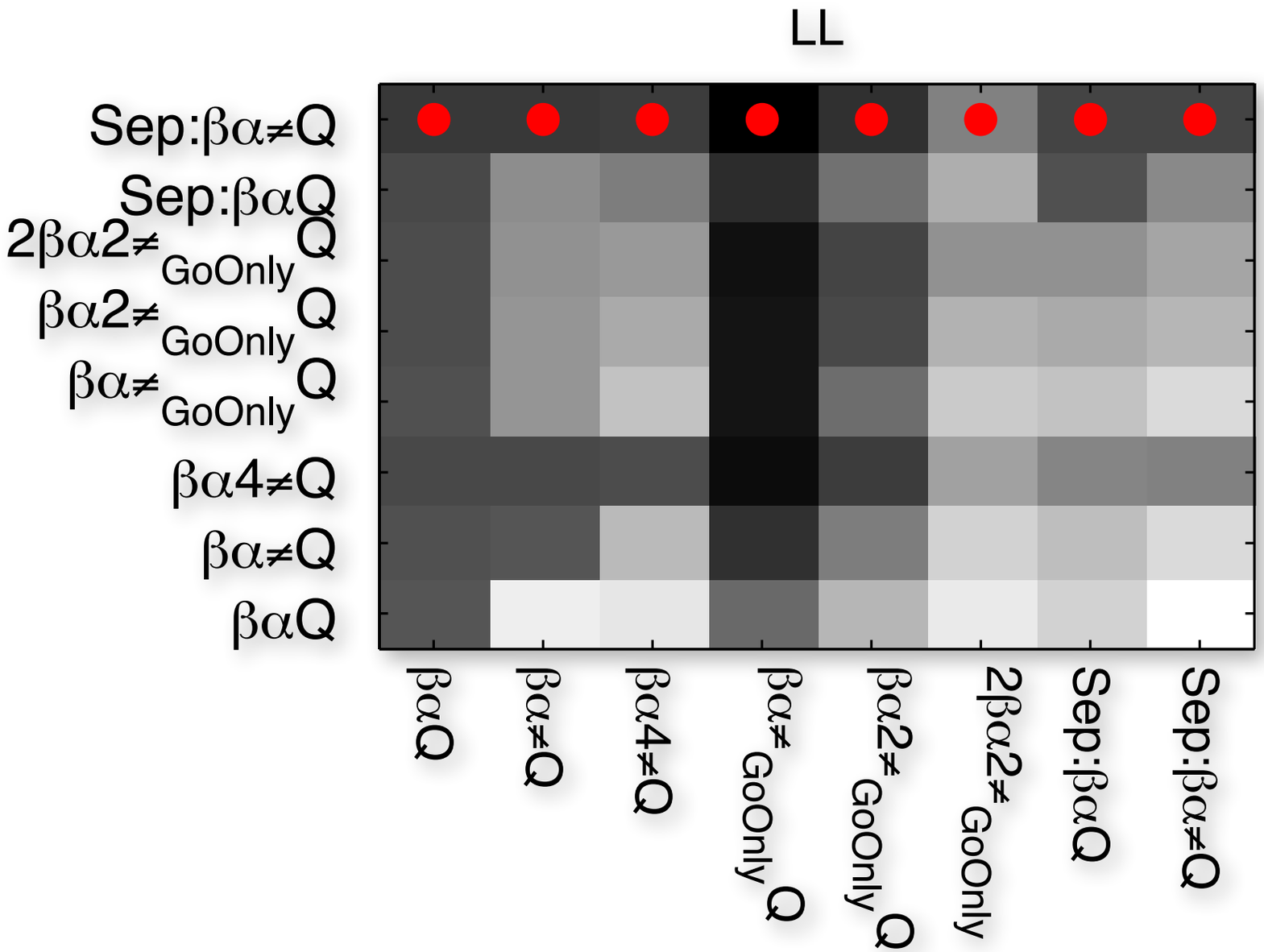
Forward differentiation



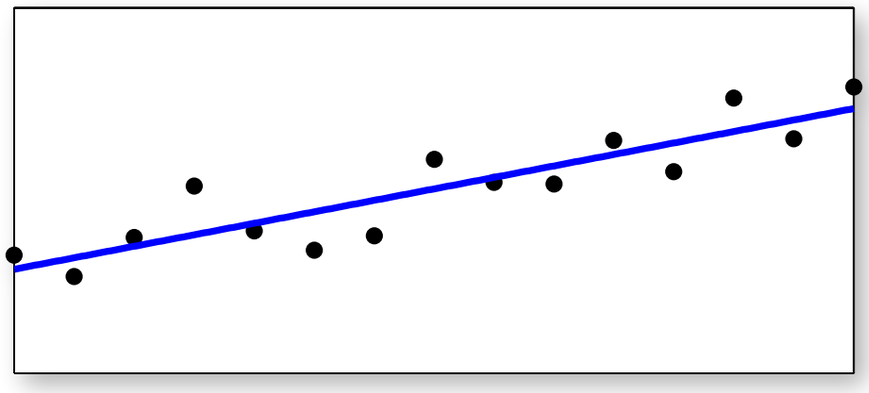
Hierarchical / random effects models

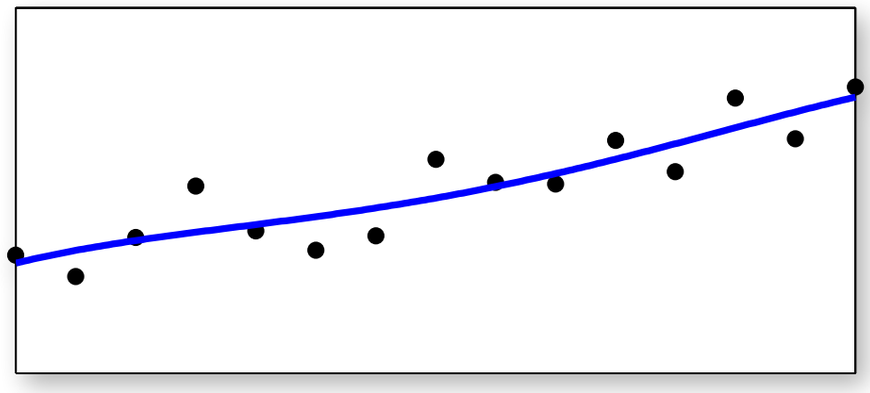
- Advantages
  - Accurate group-level mean and variance
  - Outliers due to weak likelihood are regularised
  - Strong outliers are not
  - Useful for model selection
- Disadvantages
  - Individual estimates θ_i depend on other data, i.e. on $\mathcal{A}_{j \neq i}$ and therefore need to be careful in interpreting these as summary statistics
  - More involved; less transparent
- Psychiatry
  - Groups often not well defined, covariates better
- fMRI
  - Shrink variance of ML estimates - fixed effects better still?

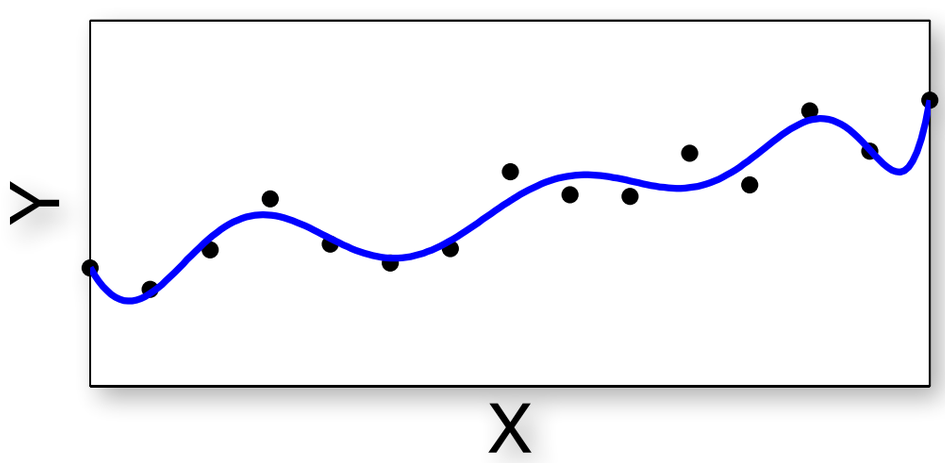
How does it do?

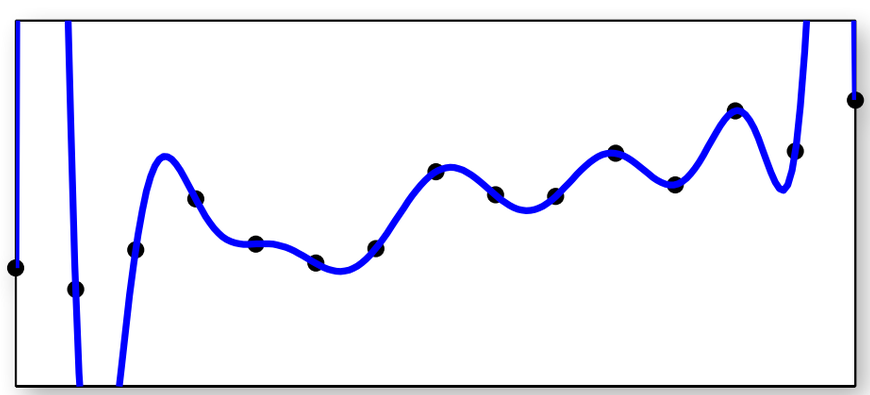


Overfitting









Model comparison

- A fit by itself is not meaningful
- Generative test
  - qualitative
- Comparisons
  - vs random
  - vs other model -> test specific hypotheses and isolate particular effects in a generative setting

Model comparison

- Averaged over its parameter settings, how well does the model fit the data?

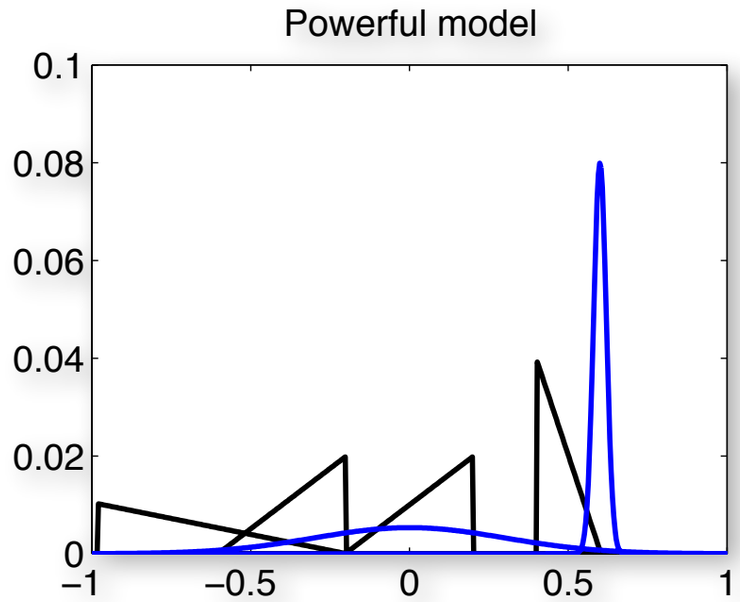
$$p(\mathcal{A}|\mathcal{M}) = \int d\theta \, p(\mathcal{A}|\theta) \, p(\theta|\mathcal{M})$$

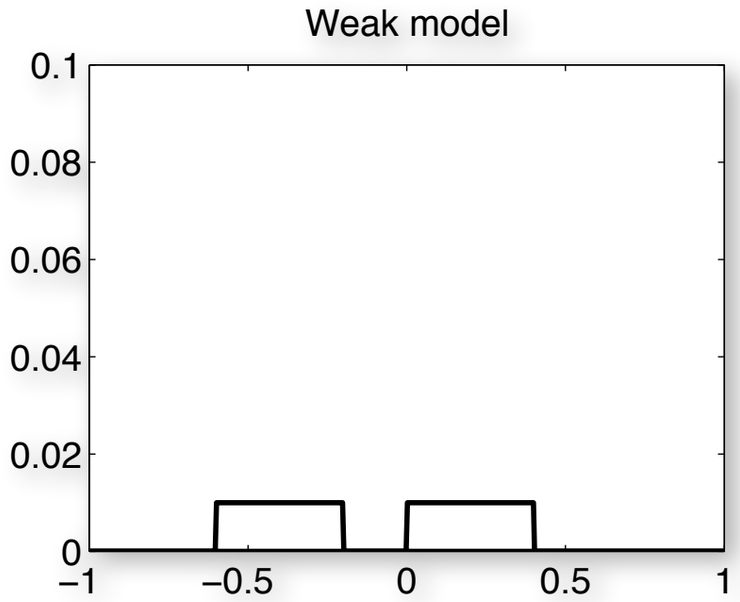
- Model comparison: Bayes factors

$$BF = \frac{p(\mathcal{A}|\mathcal{M}_1)}{p(\mathcal{A}|\mathcal{M}_2)}$$

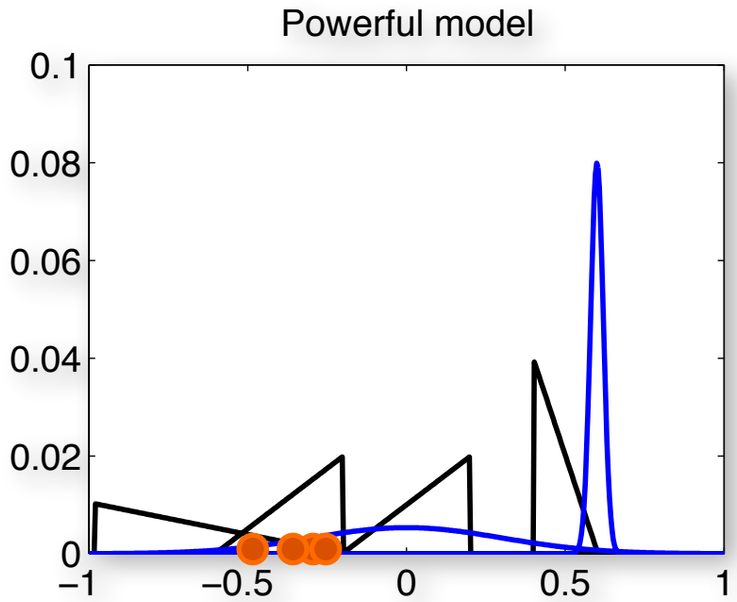
- Problem:
  - integral rarely solvable
  - approximation: Laplace, sampling, variational...

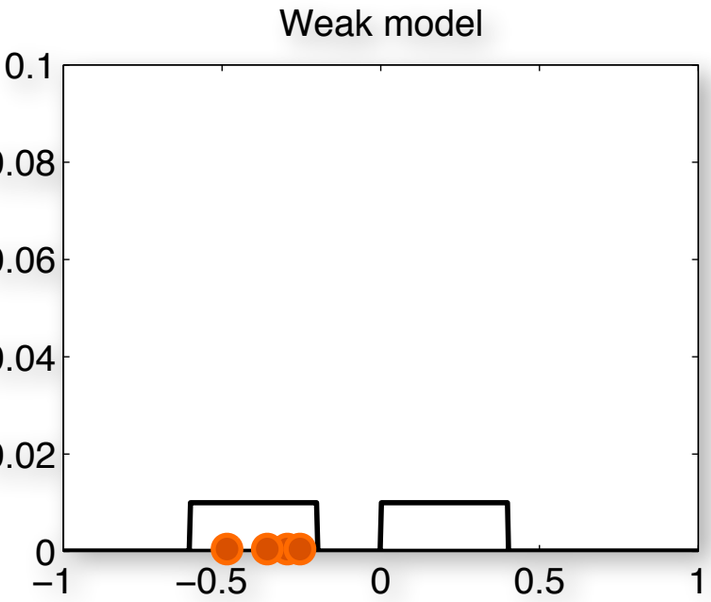
Why integrals? The God Almighty test



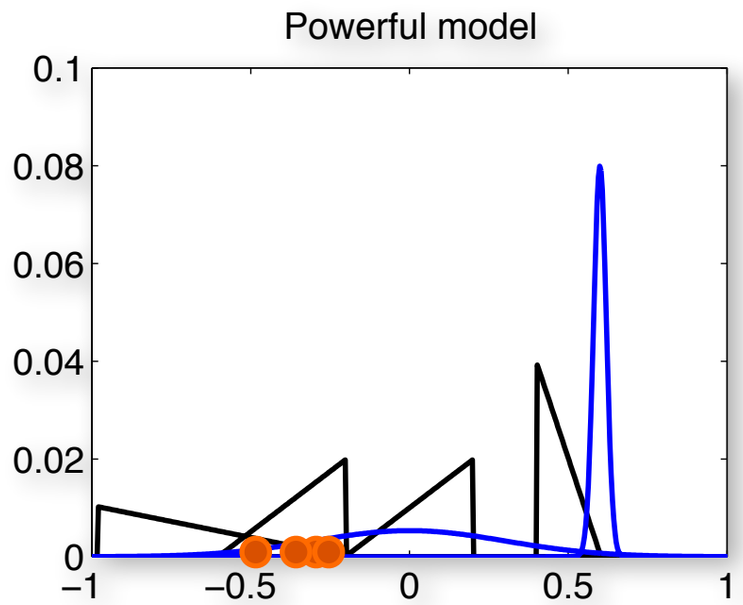


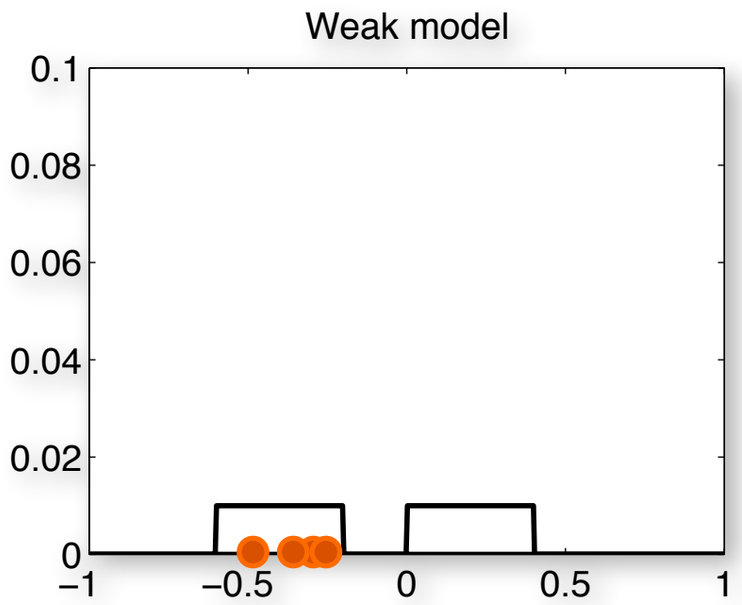
Why integrals? The God Almighty test





Why integrals? The God Almighty test





$$\frac{1}{N}(\mathbf{p}(\mathbf{X}|\boldsymbol{\theta_1}) + p(X|\theta_2) + \cdots)$$

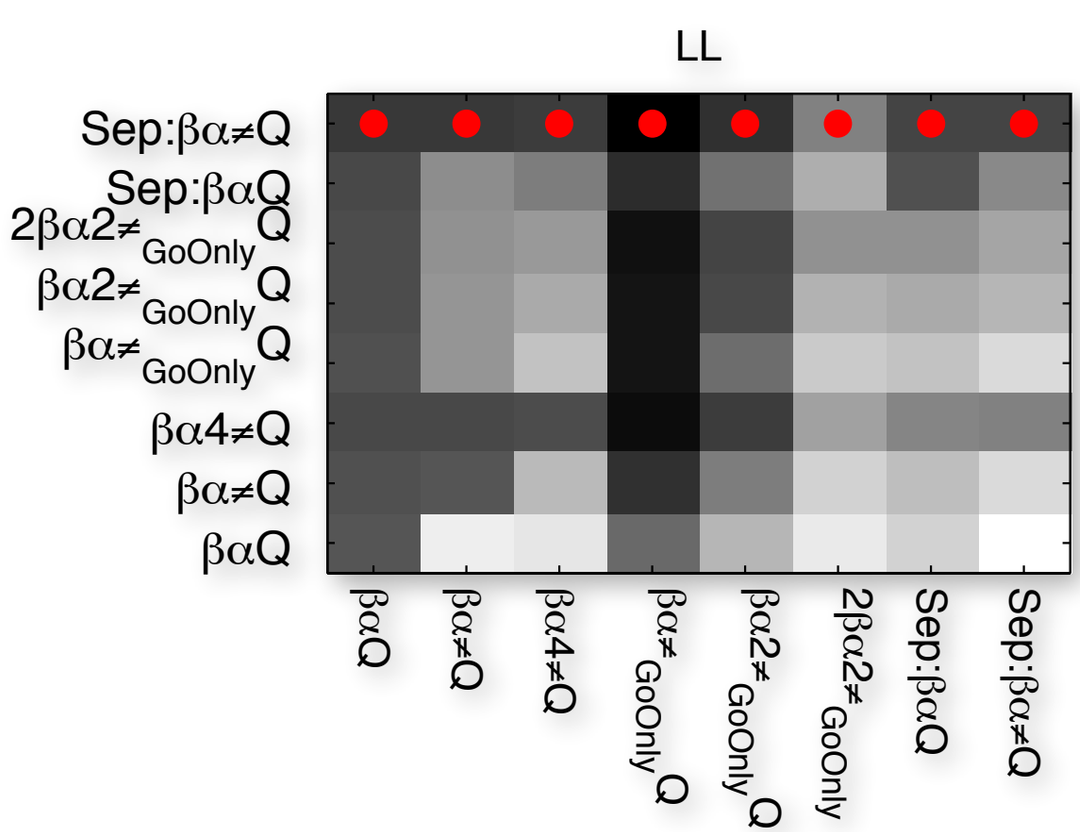
These two factors fight it out
Model complexity vs model fit

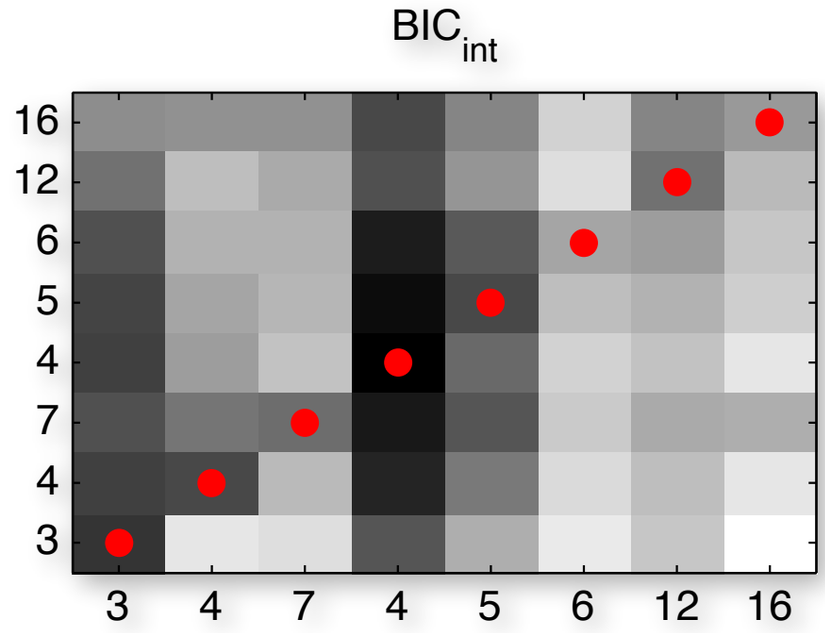
Group-level BIC

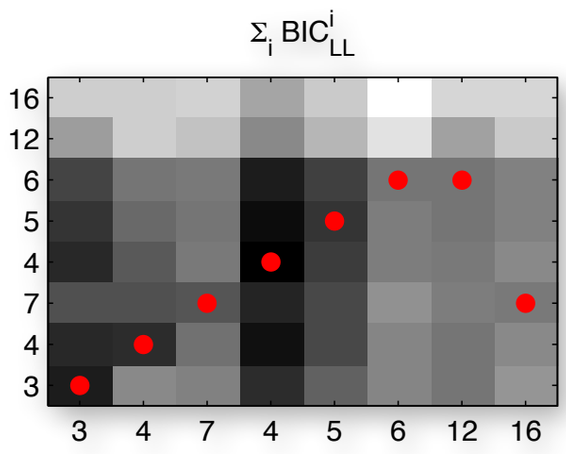
$$
\begin{aligned}
\log p(\mathcal{A}|\mathcal{M}) &= \int d\boldsymbol{\zeta}\, p(\mathcal{A}|\boldsymbol{\zeta})\, p(\boldsymbol{\zeta}|\mathcal{M}) \\
&\approx -\frac{1}{2}\mathsf{BIC}_{\mathsf{int}} \\
&= \log \hat{p}(\mathcal{A}|\hat{\boldsymbol{\zeta}}^{ML}) - \frac{1}{2}|\mathcal{M}| \log(|\mathcal{A}|)
\end{aligned}
$$

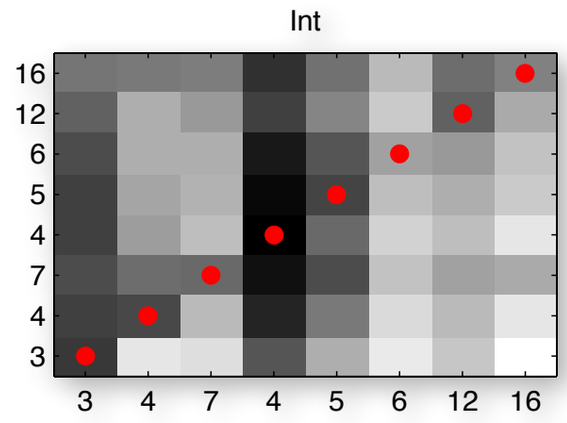
- Very simple
  - 1) EM to estimate group prior mean & variance
    - simply done using fminunc, which provides Hessians
  - 2) Sample from estimated priors
  - 3) Average

How does it do?

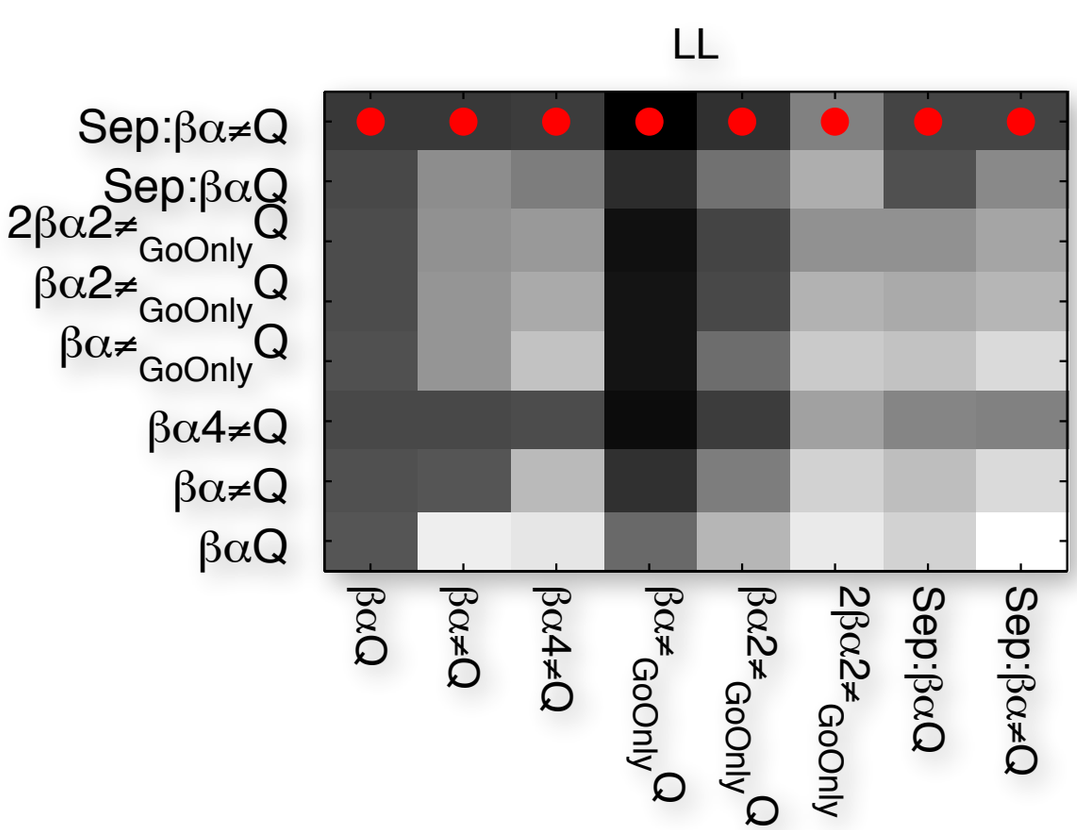


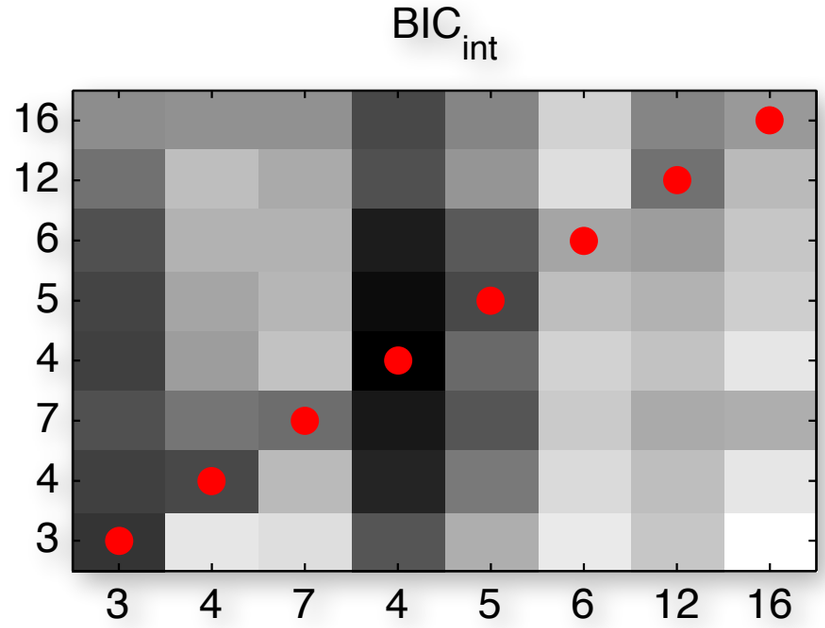




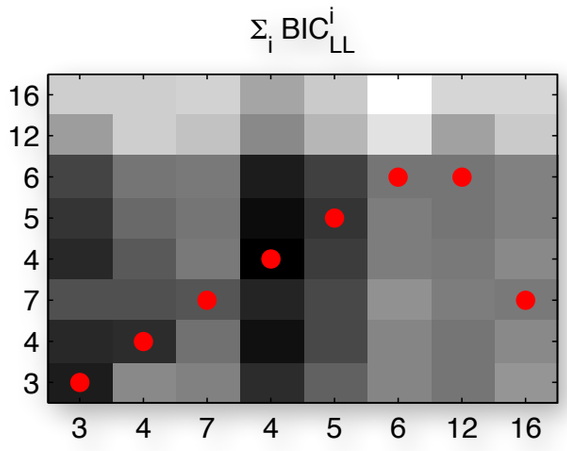


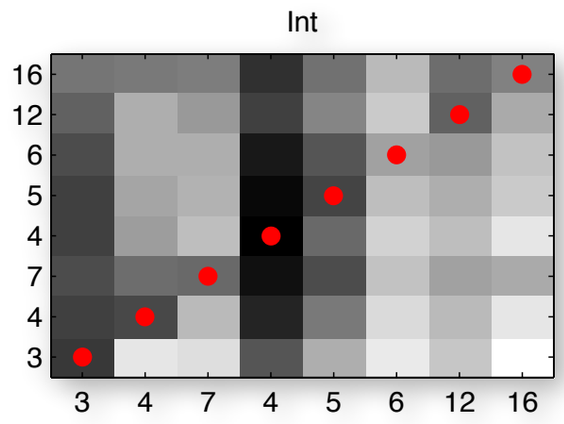
How does it do?





Fitted by EM...
too nice?

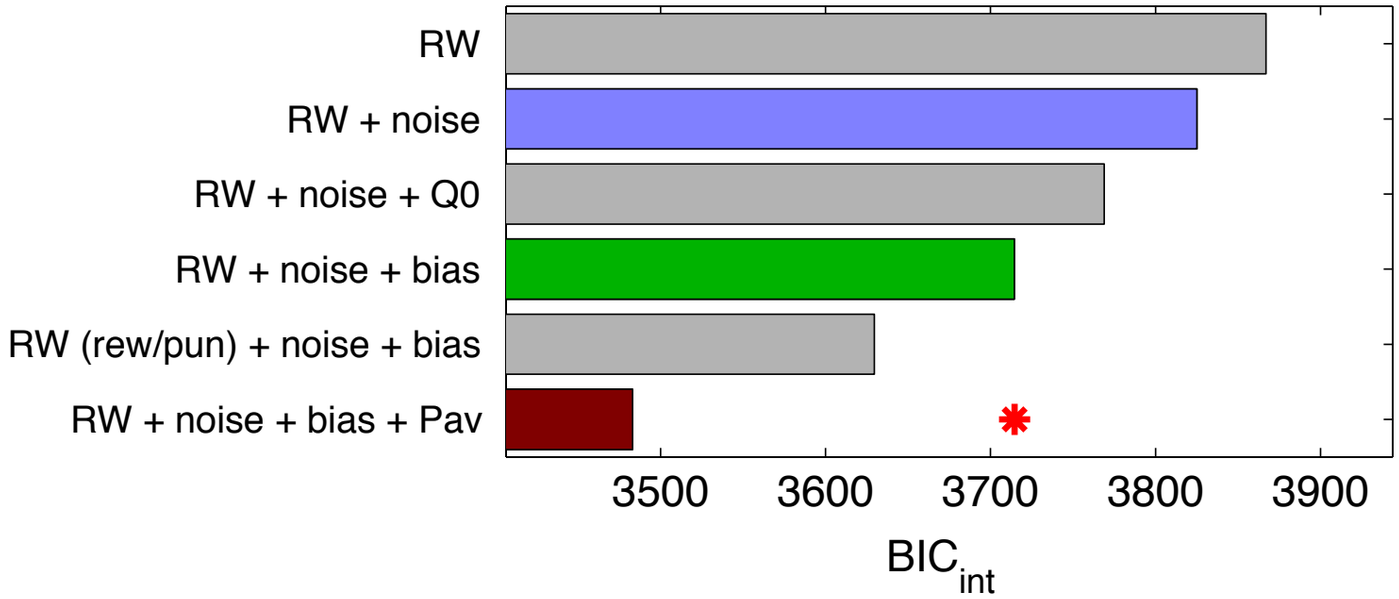


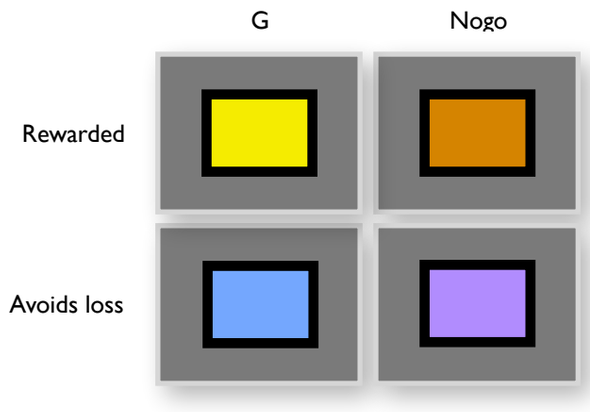


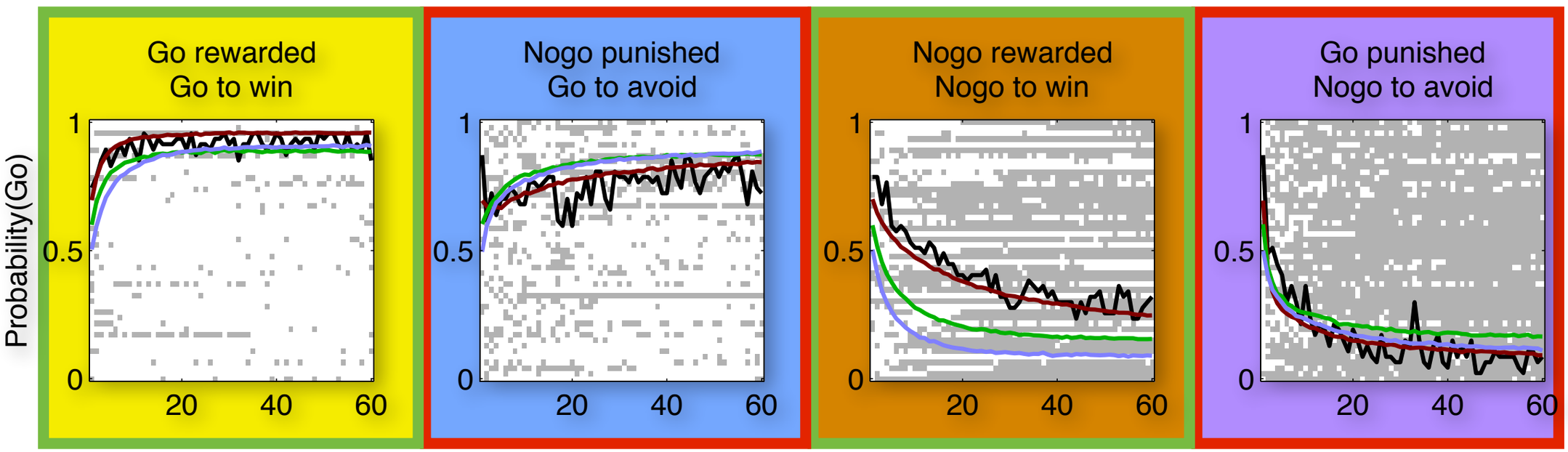
Group Model selection

Integrate out your parameters

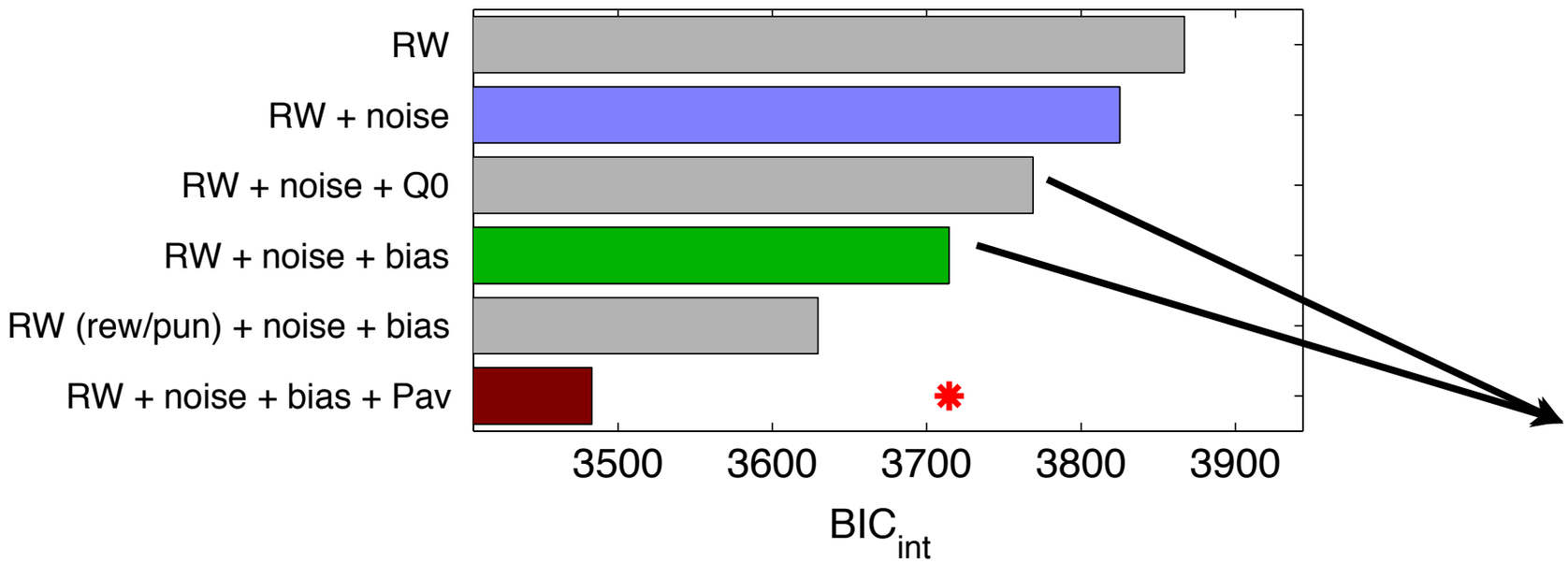
Model comparison: overfitting?

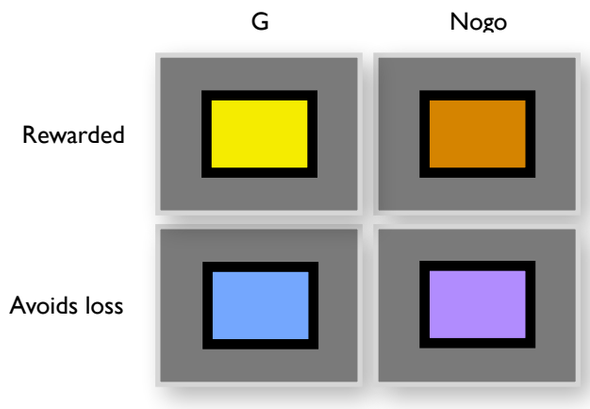




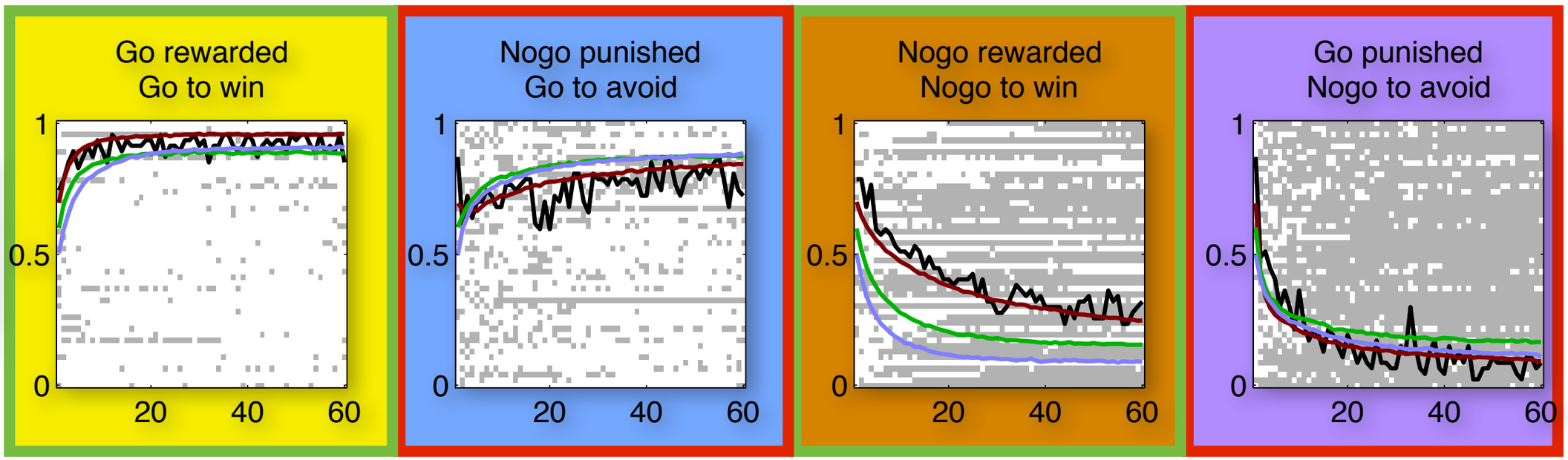


Model comparison: overfitting?



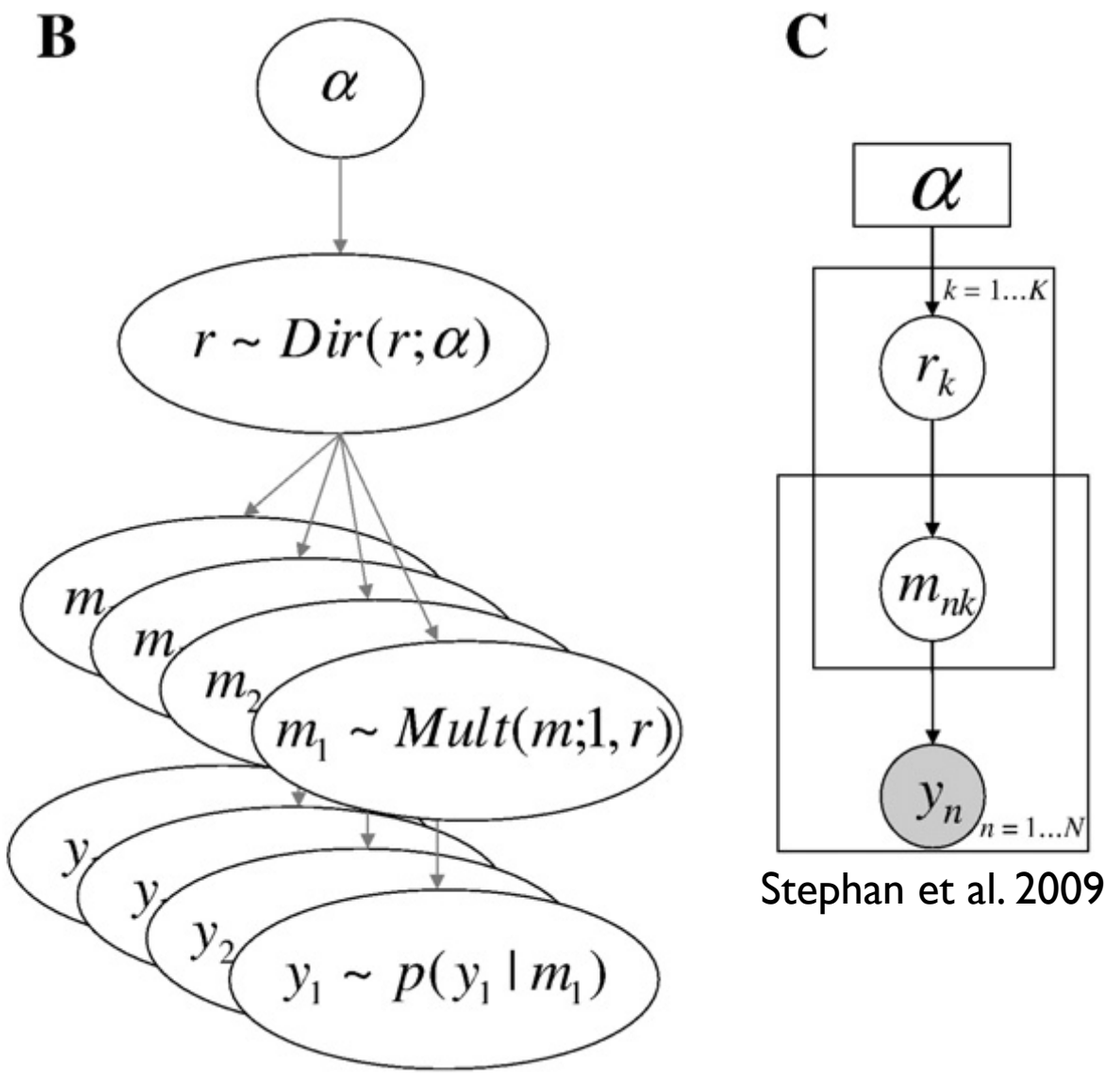


Note: same number of parameters



Posterior distribution on models

- Generative model for models



Stephan et al. 2009

Bayesian model selection - equations

- Write down joint distribution of generative model
- Variational approximations lead to set of very simple update equations
  - start with flat prior over model probabilities

$$\alpha = \alpha_0$$

  - then update

$$u_k^i = \left(\int d\theta_i \, p(\mathcal{A}_i, \theta_i | \mathcal{M}_k) \right) \exp\left(\Psi(\alpha_k) - \Psi\left(\sum_k \alpha_k \right) \right)$$

$$\alpha_k \leftarrow \alpha_{0,k} + \sum_i \frac{u_k^i}{\sum_k u_k^i}$$

emfit

- www.quentinhuys.com/pub/emfit
- six model classes
  - basic RW - to learn
  - Affective Go/Nogo (Guitart et al., 2012)
  - Probabilistic Reward (Pizzagalli et al., 2005)
  - Twostep (Daw et al., 2011)
  - Effort (Gold et al., 2013)
  - Pruning (Huys et al., 2012)
- ML, MAP and EM-MAP
- GLM (no implicit differentiation, so only for small models!)
- many models for each task - standardly fit all
- model comparison
- data generation and plotting of standard sanity checks

emfit

- Run example
  - generate sample dataset
  - fit all models
  - do model comparison, output as figure & latex file
  - plot sanity / generative checks
  - batchRunEMfit('mTASKNAME')
- Fit dataset
  - check data format in mTASKNAME/dataformat.txt or look at example dataset generated above in fitResults/Data.mat
  - batchRunEMfit('mTASKNAME',Data.mat)

How to fit a model and believe the results

Model building The first step is to build a series of models. Each contains an internal process by which different choice options are valued, and a link function which describes how preferences turn into observed decisions. At least two models should be built: a model M0 of 'no interest' that performs the task, but without involving the process of interest, and a model M1 that does contain the process of interest.

Validation on surrogate data

1. **Data generation**: Run each model on the experiment from which data will be examined. Do the generated data look reasonable?
2. **Surrogate model fitting**: Fit each model to the data generated from it. Are the true parameters readily recovered? Are some parameters not identifiable?
3. **Surrogate model comparison**: Does the model comparison procedure correctly identify the data generated by each model?

Real data analysis

1. **Real model fitting**: Fit each model to the real data.
2. **Real model validation**: Run each model with the fitted parameters on the exact experimental instance presented to that particular subject. Are the key features of the real data captured reasonably?
3. **Real model comparison**: choose the least complex model that best accounts for the data.
4. **Parameter examination**: only at this point should the parameters of the model be examined, and only the parameters of the most parsimonious model should be ascribed meaning.

Huys 2017

Behavioural data modelling

- Are no panacea
  - statistics about specific aspects of decision machinery
  - only account for part of the variance

- Model needs to match experiment
  - ensure subjects actually do the task the way you wrote it in the model
  - model comparison

- Model = Quantitative hypothesis
  - strong test
  - need to compare **models**, not **parameters**

Thanks

- Nathaniel Daw
- Peter Dayan
- Daniel Schad

- SNF
- DFG