Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

University of Zurich[UZH]

Bayesian hierarchical modeling of learning

Christoph Mathys

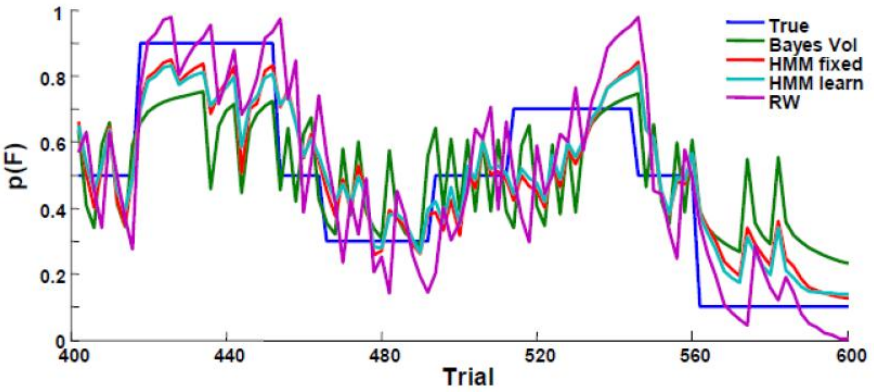
Translational Neuromodeling Unit (TNU)

University of Zurich and ETH Zurich

Schloss Ringberg, September 17, 2012

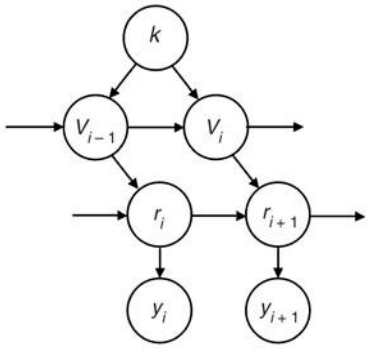
Context

- Hierarchical Bayesian models are natural candidates for explaining learning



den Ouden et al. (2010)

- However, their normative nature and computational cost pose problems

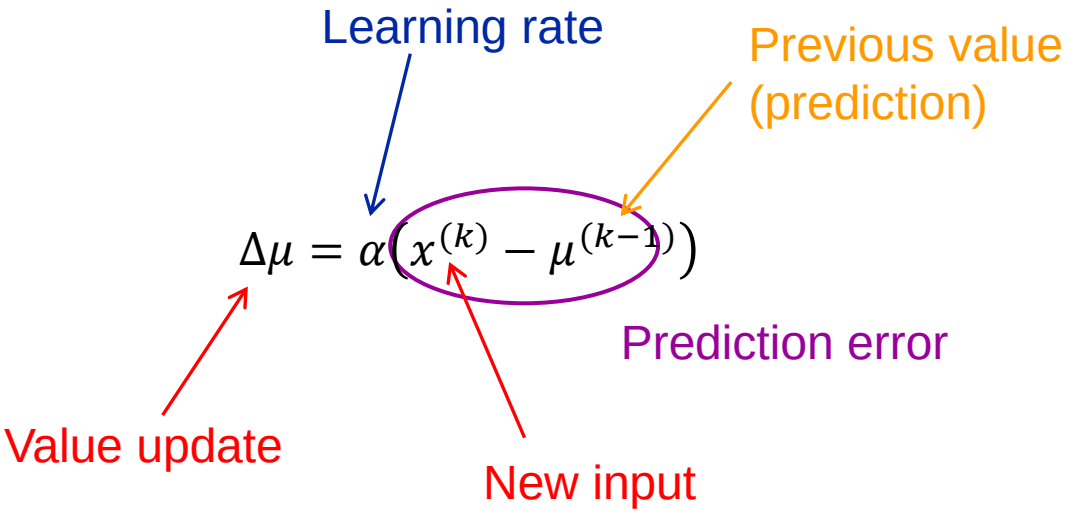


$$p(r_i, v_i, k \mid y_{\le i}) \propto p(k) \int \cdots \int p(r_1) p(v_1) \prod_{j=1}^{i} [p(y_j \mid r_j) p(r_j \mid r_{j-1}, v_j) p(v_j \mid v_{j-1}, k)] dr_{\le i-1} dv_{\le i-1}$$

Behrens et al. (2007)

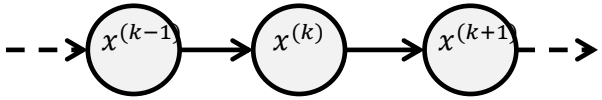
Context

Rescorla-Wagner learning:

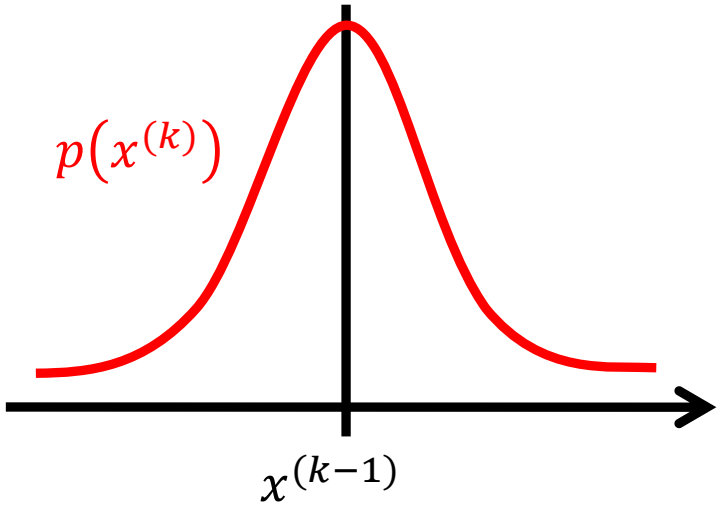


A generalized approach to learning

- A very general goal: to learn about a continuous quantity that changes



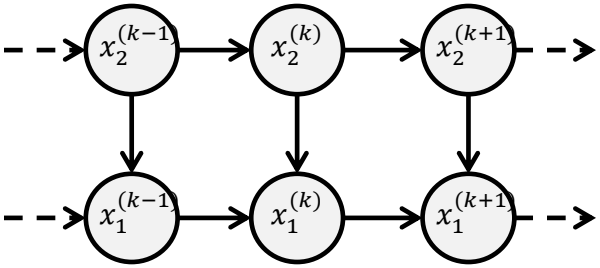
- Assumption: it performs a Gaussian random walk



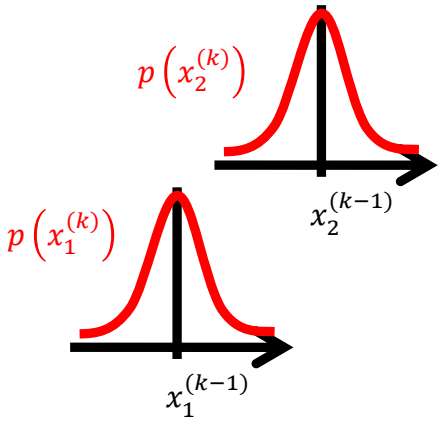
$$x^{(k)} \sim \mathcal{N}\left(x^{(k-1)}, \vartheta\right)$$

A generalized approach to learning

- To allow for changes in volatility, we take the variance of the random walk to be a positive function f of another state, x_2.



- We may then assume the volatility to perform its own Gaussian random walk.

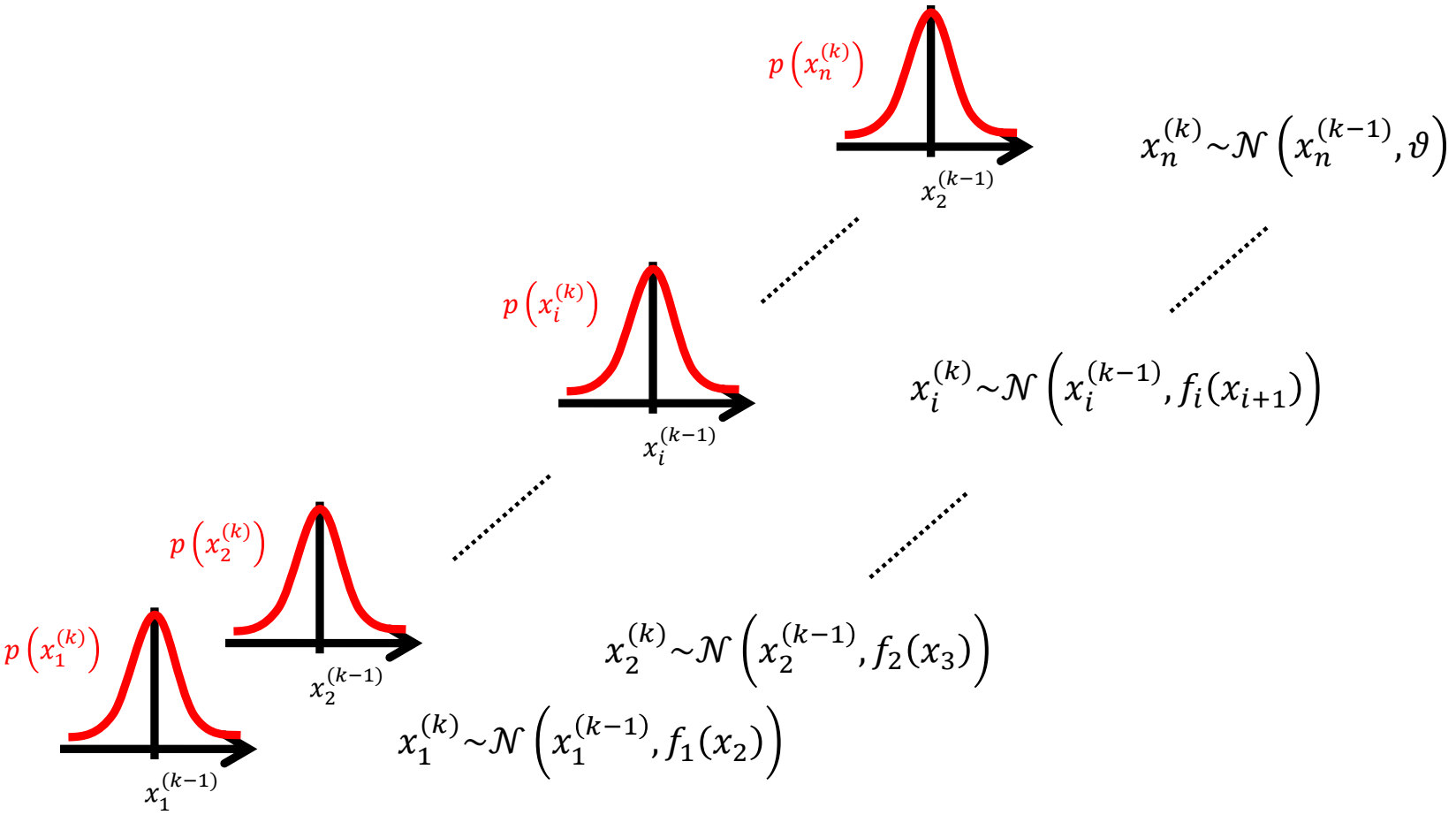


$$x_2^{(k)} \sim \mathcal{N}\left(x_2^{(k-1)}, \vartheta\right)$$

$$x_1^{(k)} \sim \mathcal{N}\left(x_1^{(k-1)}, f(x_2)\right)$$

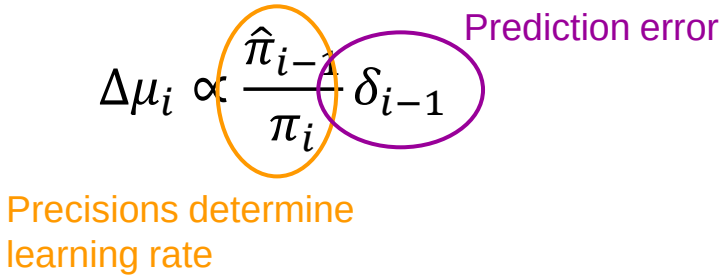
A generalized approach to learning

This can be continued *ad infinitum*. In practice, we stop at some level n, where we assume the volatility to be constant.



Variational inversion and update equations

- Inversion proceeds by introducing a mean field approximation and fitting quadratic approximations to the resulting variational energies (Mathys et al., 2011).
- This leads to **simple one-step update equations** for the sufficient statistics (mean and precision) of the approximate Gaussian posteriors of the states x_i.
- The updates of the means have the same structure as value updates in Rescorla-Wagner learning:



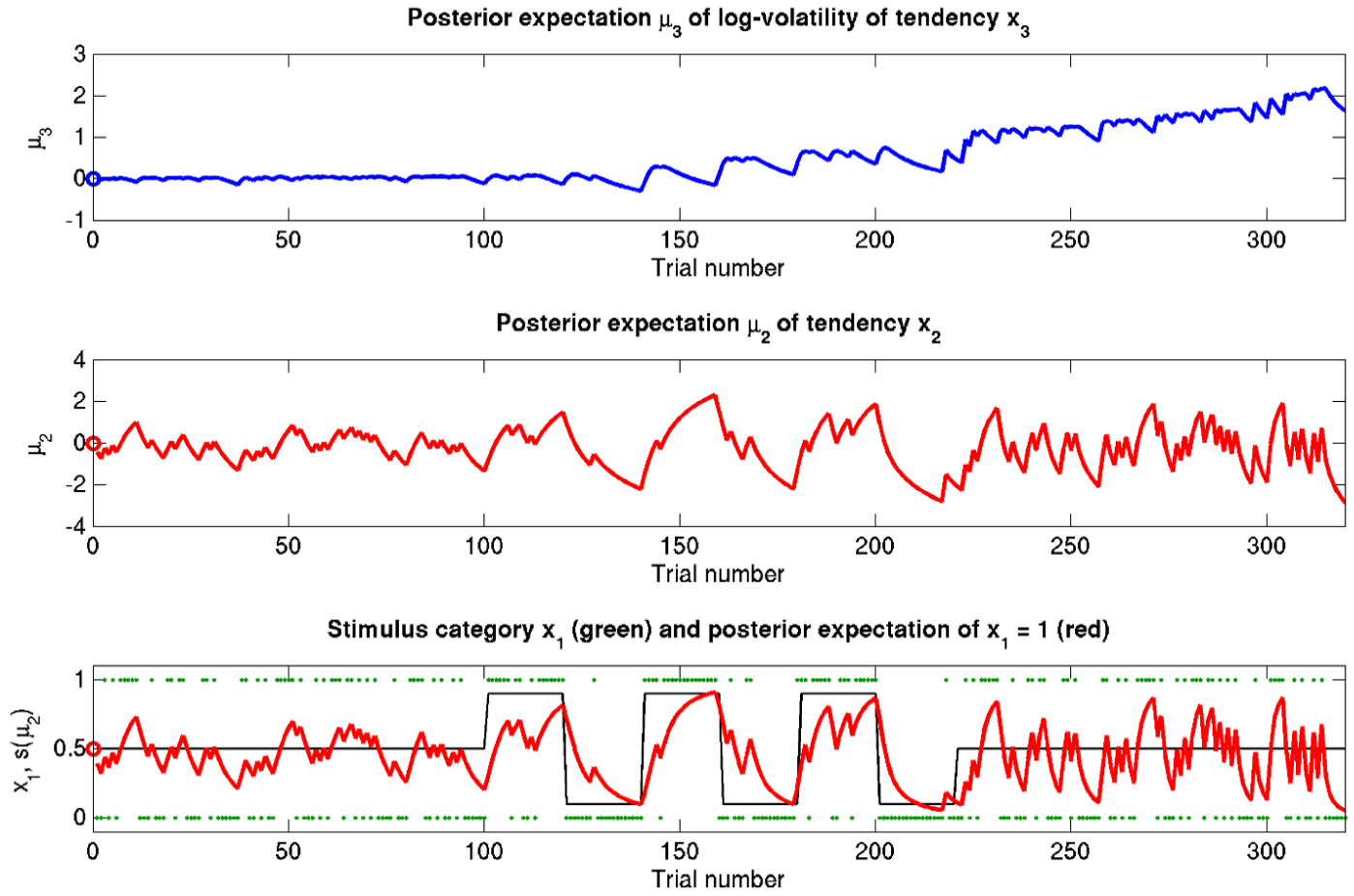
- Furthermore, the updates are **precision-weighted prediction errors**.

Practical uses

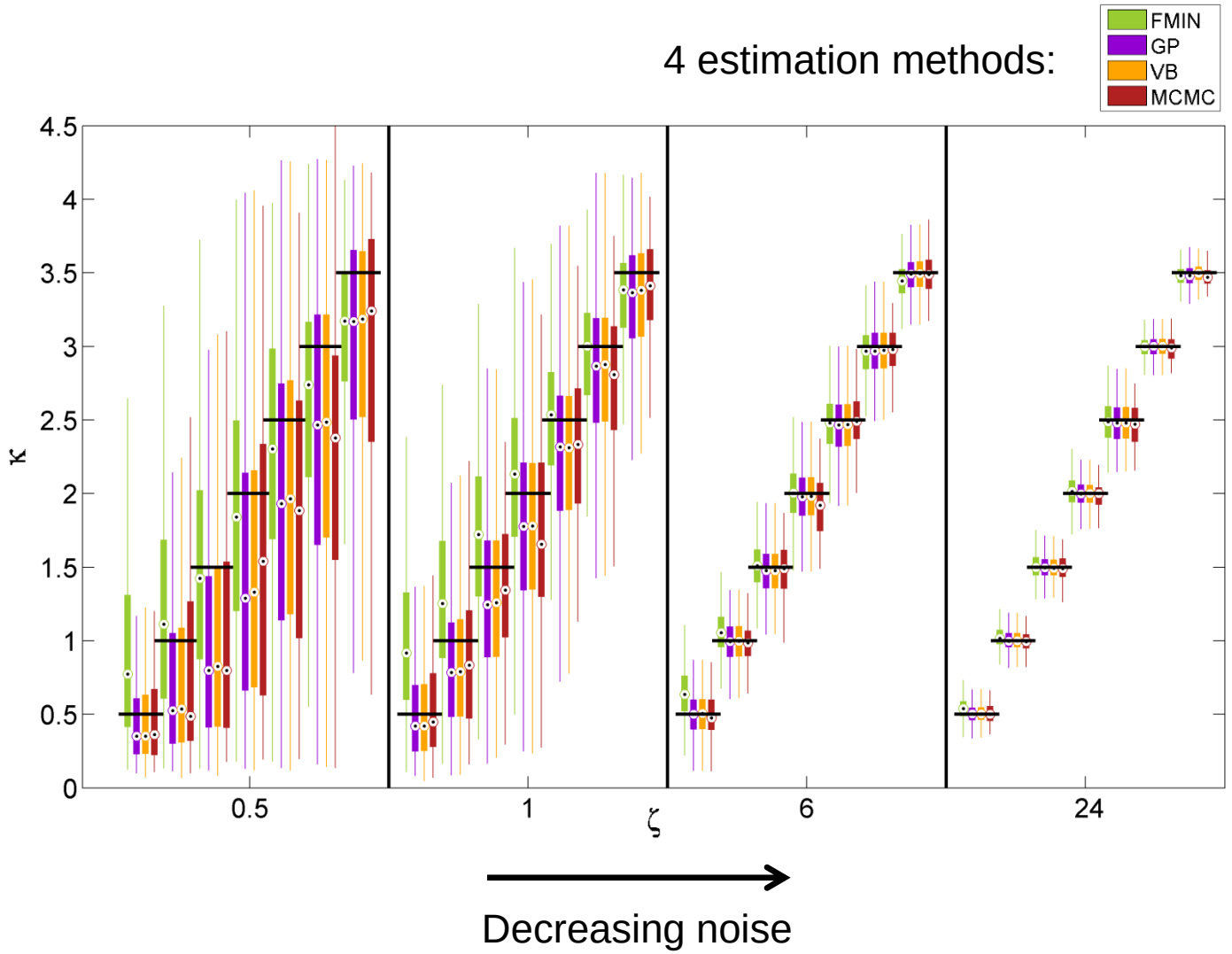
- Updates have a **general** and **interpretable** structure.
- They are **computationally** extremely **efficient.**
- They contain **parameters** that can differ from subject to subject and can be **individually estimated** from experimental data.
- This enables the **comparison of parameter estimates** between subjects and of **evolving beliefs on states** within subjects.
- Furthermore, it provides a basis for **model selection** on the basis of log-model evidence (e.g., comparison of **learning models** with different hierarchical depths, comparison of **decision models**).

Context effects on the learning rate

Simulation: $\vartheta = 0.5,\ \omega = -2.2,\ \kappa = 1.4$



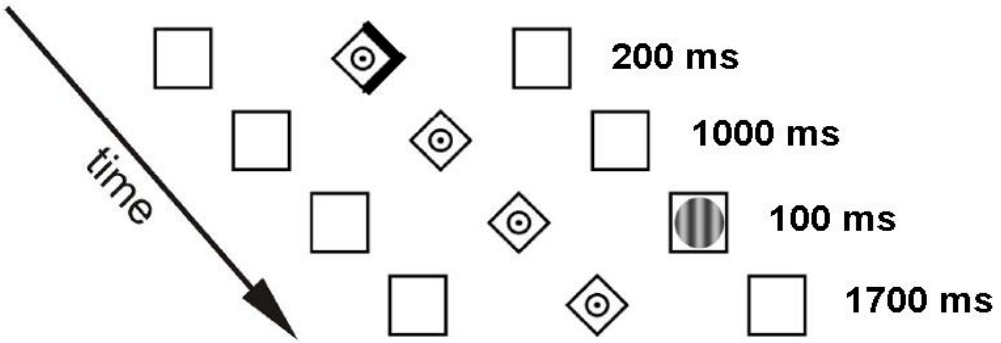
Parameter estimation

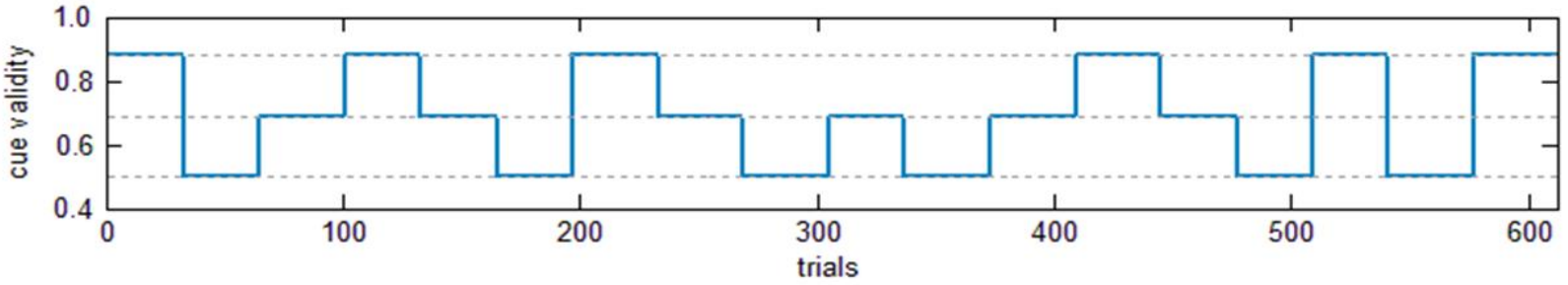


Applications (Vossel & Mathys et al., submitted)

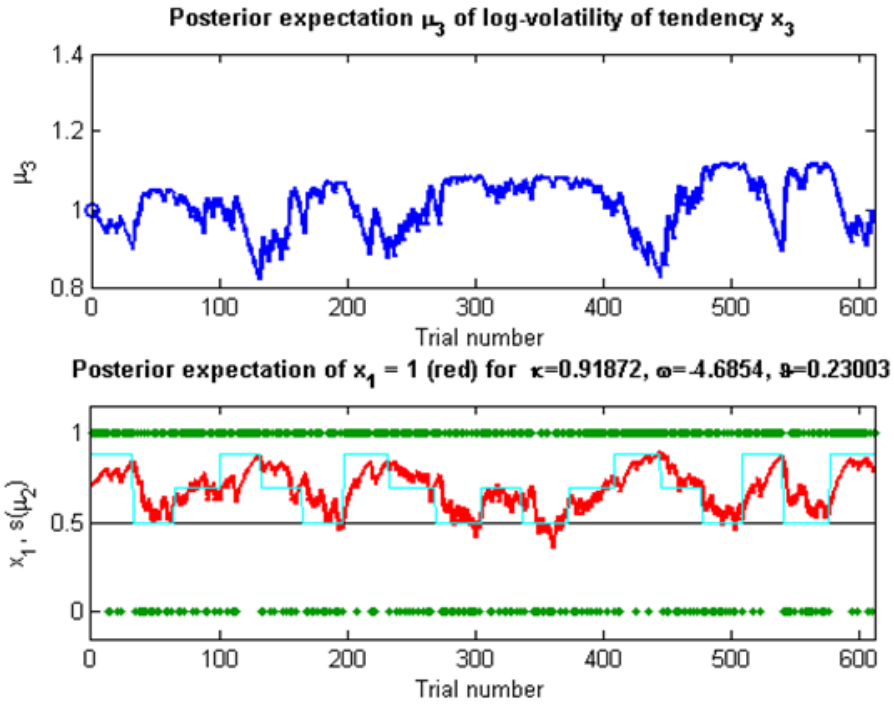
Measurement of saccadic reaction times in volatile Posner task

(15 healthy human subjects):

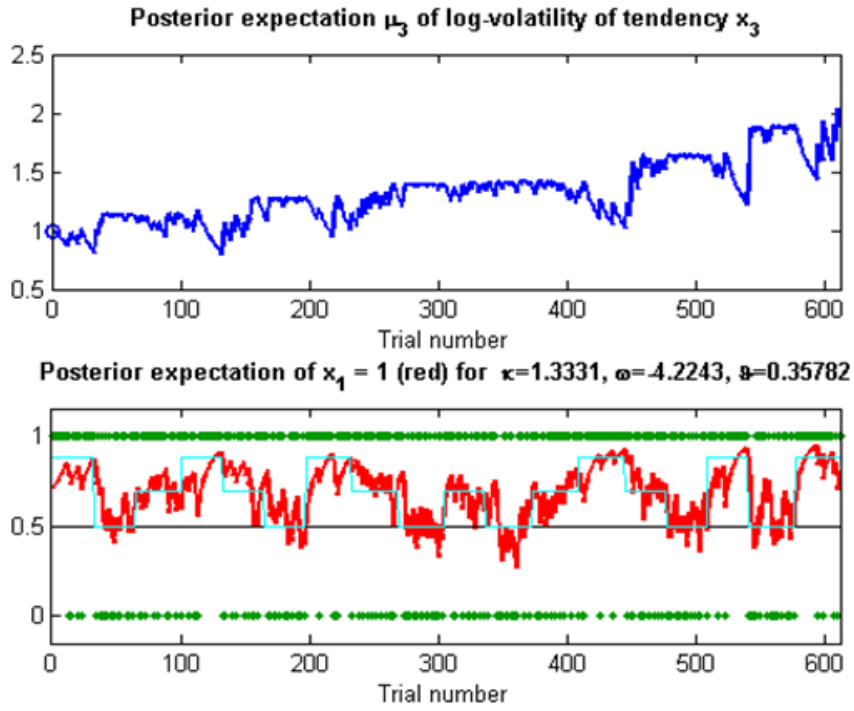




Applications (Vossel & Mathys et al., submitted)



ω=-5.84; ϑ=0.02



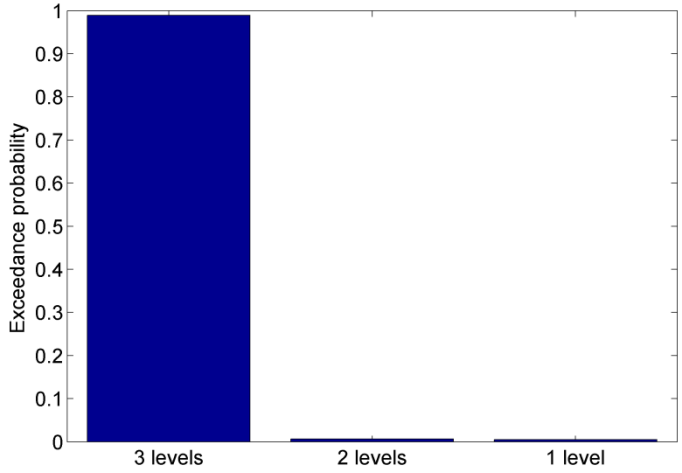
ω=-4.5; ϑ=0.98

Applications (Vossel & Mathys et al., submitted)

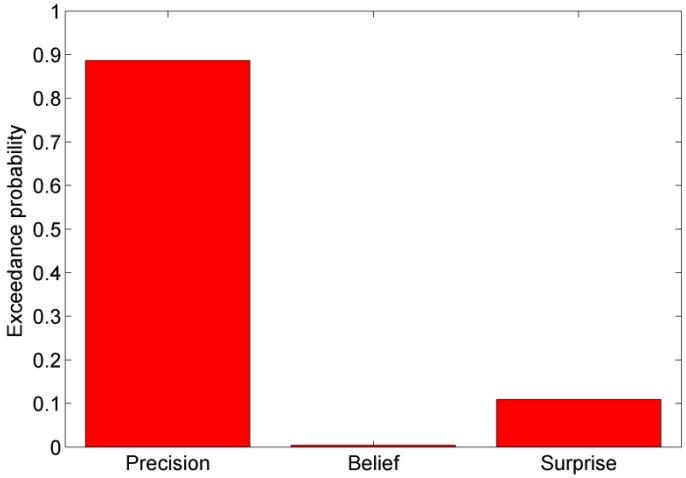
Decision

Learning	Precision	Belief	Surprise
1 level			
2 levels			
3 levels			

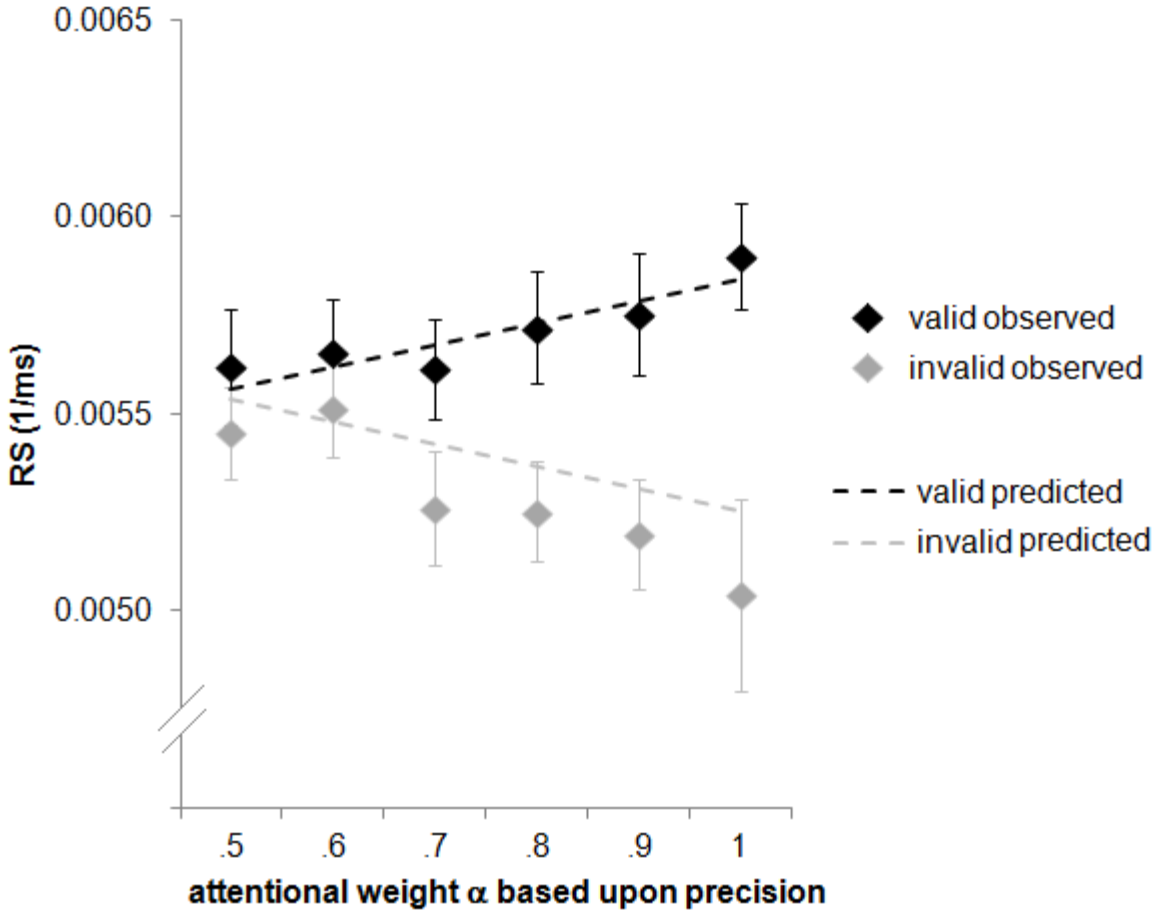
Learning models



Decision models

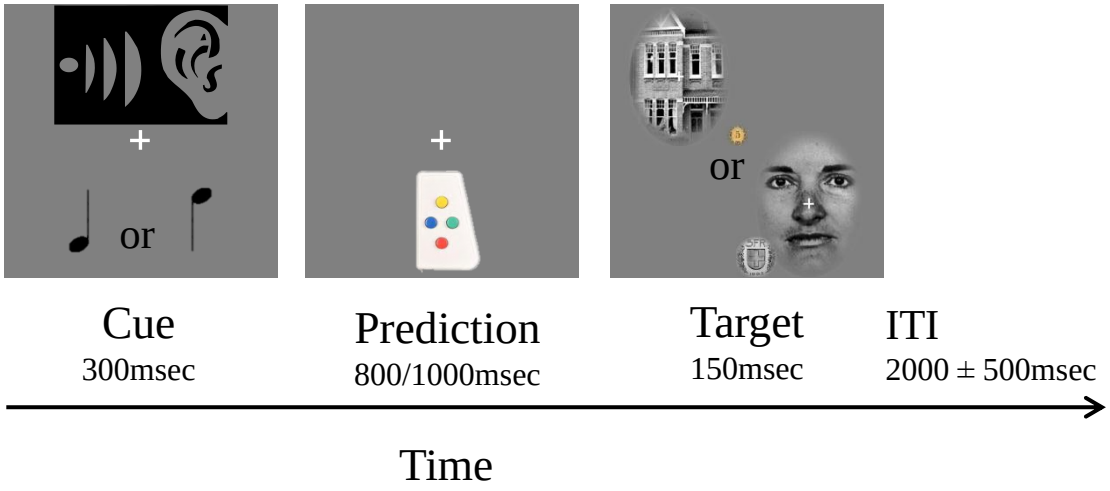


Applications (Vossel & Mathys et al., submitted)

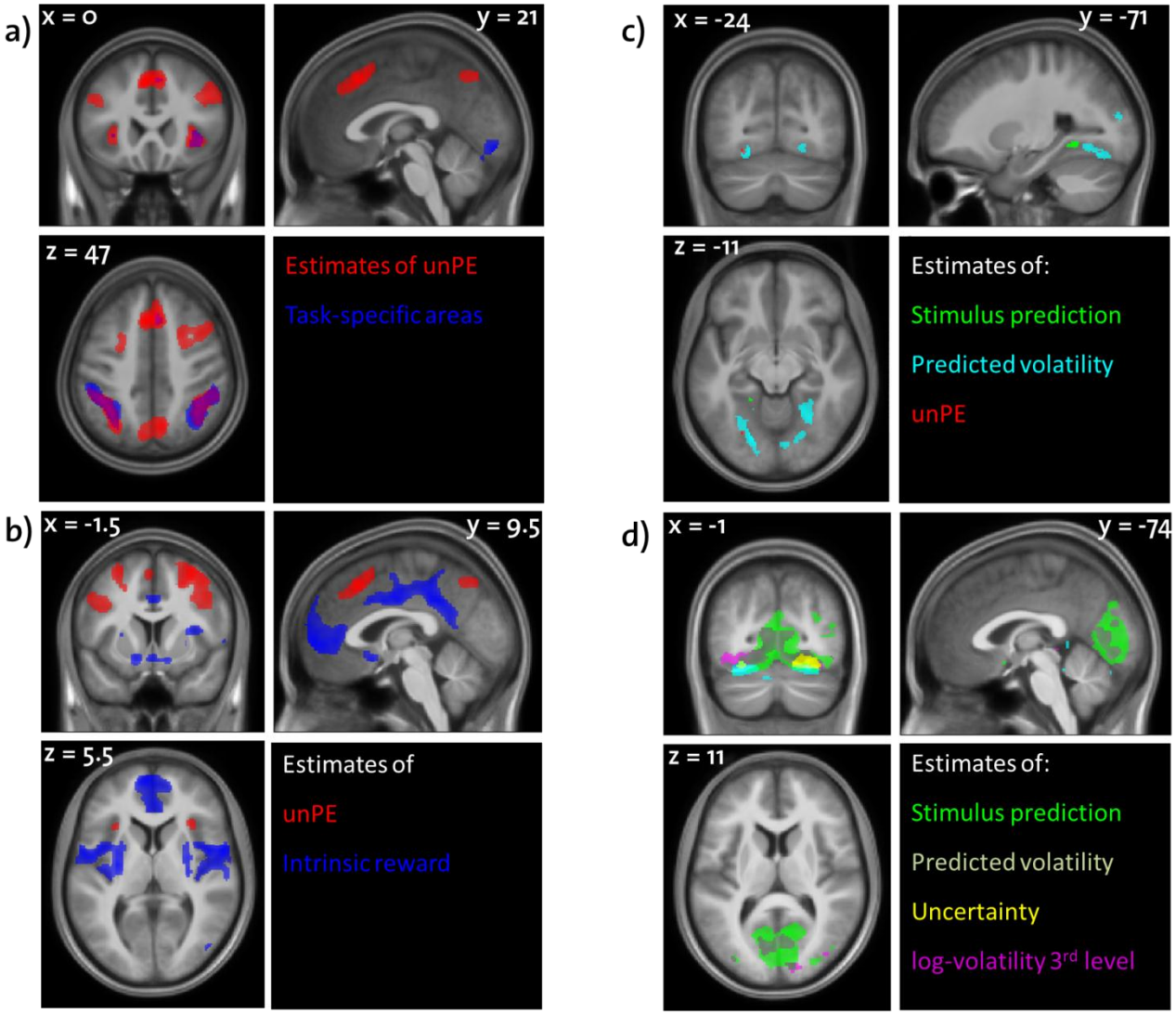


Applications (Iglesias et al., in prep.)

Associative learning task:

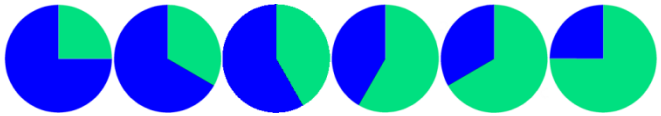


Applications (Iglesias et al., in prep.)



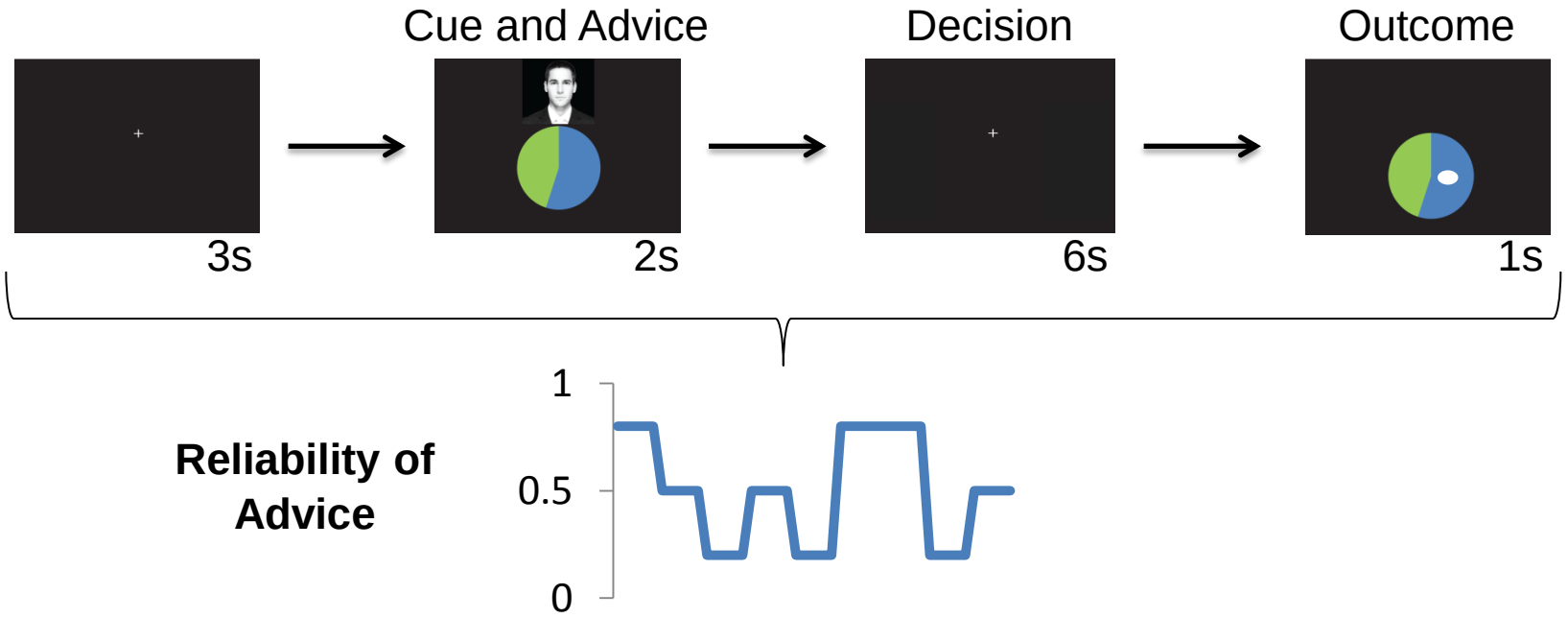
voxel-level threshold $p < 0.05$ FWE whole-brain corrected

Applications (Diaconescu et al., in prep.)

Cue: 

Advice: Video of adviser holding up blue or green card

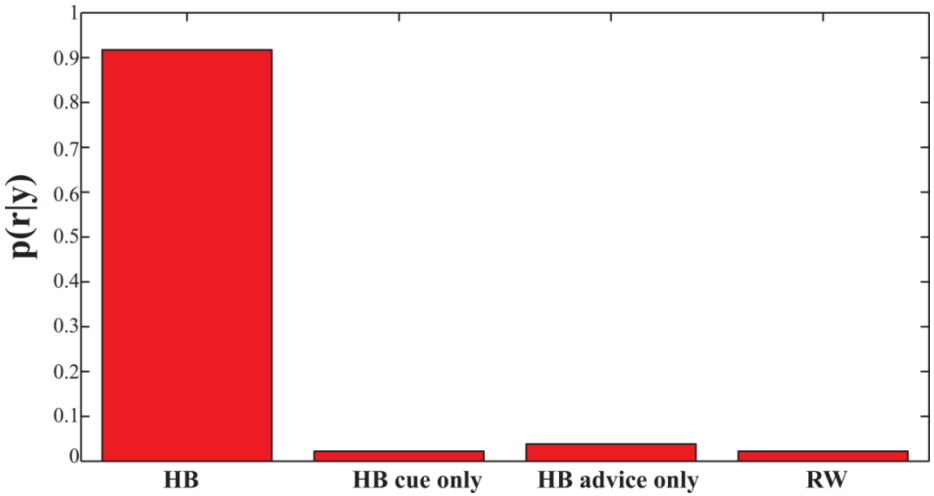
Subjects must compute: $p(\text{blue}|\ \text{cue} = 65\%\ \text{blue}, \text{advice} = \text{green}, \text{history of advice})$



Applications (Diaconescu et al., in prep.)

Bayesian model selection:

- HB model with decision based on cue + advice
- HB model with decision based on cue only
- HB model with decision based on advice only
- Rescorla-Wagner model

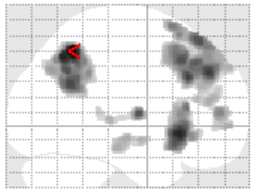


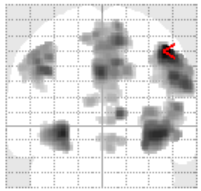
⟹ Use of volatility estimates in adjusting learning

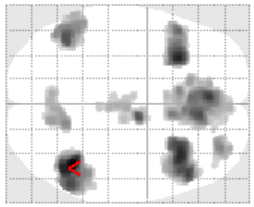
⟹ Combination of both social and non-social cues in decision-making

Applications (Diaconescu et al., in prep.)

Correlation with absolute prediction error on CUE

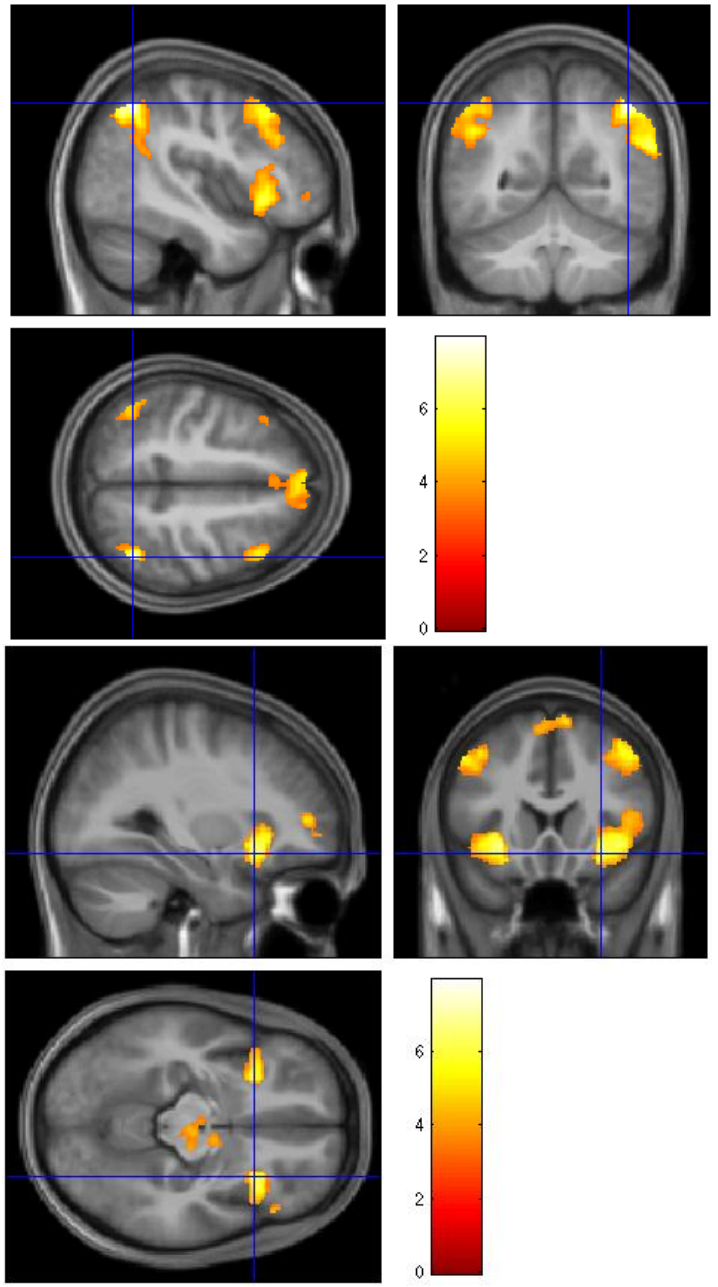






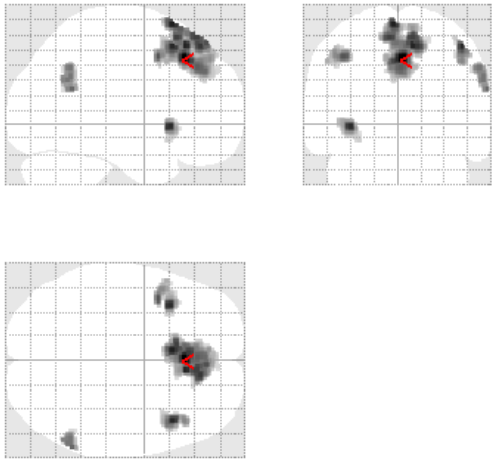
Cluster level p < 0.05 FWE corrected:

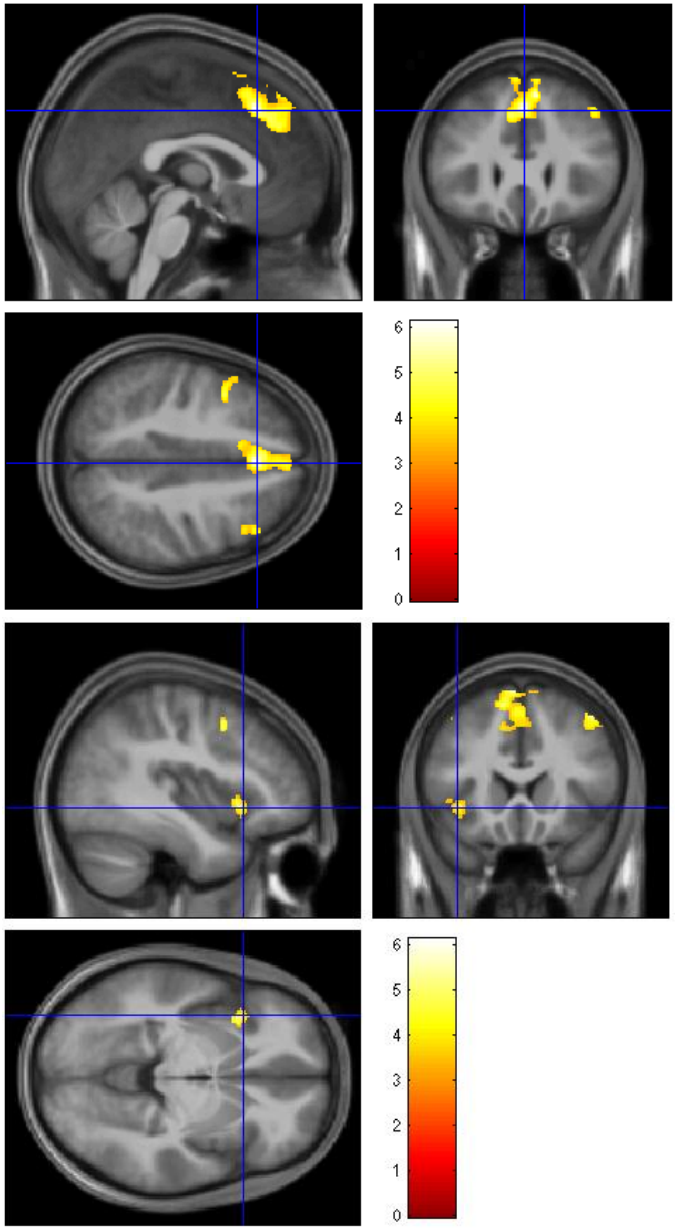
- R midbrain
- B anterior insula
- B dorsomedial prefrontal cortex
- R orbitofrontal cortex
- B inferior parietal cortex



Applications (Diaconescu et al., in prep.)

Correlation with absolute prediction error on ADVICE



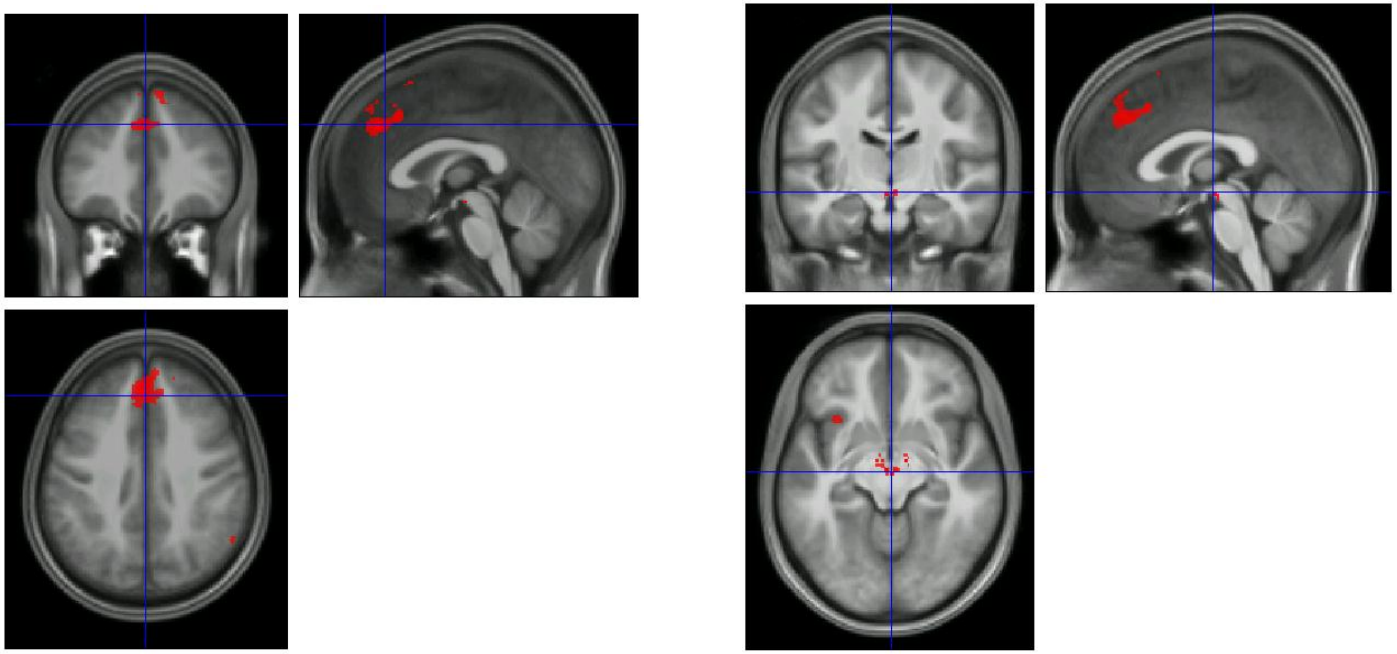


Cluster level p < 0.05 FWE corrected:

- B superior medial frontal cortex
- R dorsomedial prefrontal cortex
- L anterior insula
- R superior temporal cortex
- R inferior parietal cortex

Applications (Diaconescu et al., in prep.)

Conjunction of absolute prediction error on CUE AND ADVICE



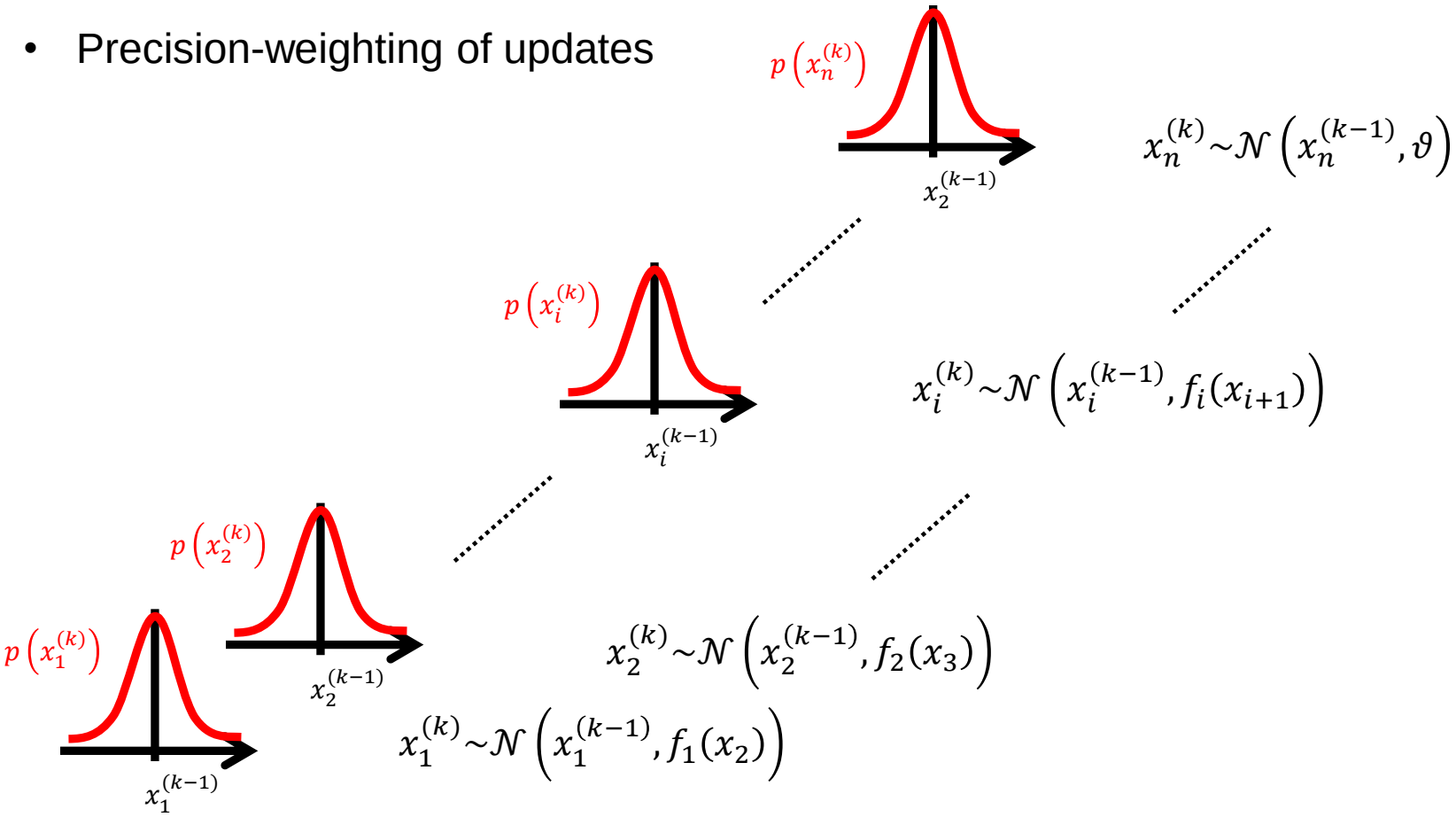
Whole-brain p < 0.05 FWE corrected:

- B dorsomedial prefrontal cortex
- L anterior insula
- B VTA
- B substantia nigra

The model in more detail

A closer look at:

- The coupling between levels
- The variational inversion
- Precision-weighting of updates



Coupling between levels

- At each level, the coupling to the next highest level is given by a positive function $f_i(x_{i+1})$.

- In what follows, we define this as the exponential of an affine function of x_{i+1}:

$$f_i(x_{i+1}) \stackrel{\text{def}}{=} \exp(\kappa_i x_{i+1} + \omega_i)$$

- This choice of coupling has the advantage of enabling the derivation of simple one-step update equations for an approximate inversion of the model.

- In some situations, even simpler choices of coupling (e.g., eliminating the κ_i by setting them to 1) can be appropriate.

Coupling between levels

Since f has to be everywhere positive, we cannot approximate it by expanding in powers. Instead, we expand its logarithm.

$$f(x) > 0 \; \forall \, x \implies \exists \, g : f(x) = \exp\big(g(x)\big) \; \forall \, x$$

$$g(x) = g(a) + g'(a) \cdot (x - a) + O(2) = \log f(x) =$$

$$= \log f(a) + \frac{f'(a)}{f(a)} \cdot (x - a) + O(2) =$$

$$= \underbrace{\frac{f'(a)}{f(a)}}_{\stackrel{\text{def}}{=} \kappa} \cdot x + \underbrace{\log f(a) - a \cdot \frac{f'(a)}{f(a)}}_{\stackrel{\text{def}}{=} \omega} + O(2) =$$

$$= \kappa x + \omega + O(2)$$

$$\implies f(x) \approx \exp(\kappa x + \omega)$$

Variational inversion

- Mean-field approximation:

$$p\left(x^{(k)},\chi \mid u^{(1\ldots k)}\right)=p\left(x^{(k)} \mid \chi,u^{(1\ldots k)}\right)p\left(\chi \mid u^{(1\ldots k)}\right)$$

$$p\left(x^{(k)} \mid \chi,u^{(1\ldots k)}\right)\approx\prod_{i=1}^{n}\hat{q}\left(x_i^{(k)}\right) \qquad (11)$$

$$q\left(x_i^{(k)}\right)\approx\hat{q}\left(x_i^{(k)}\right)$$

- Maximum-entropy form for approximate posteriors under parametric assumptions:

$$q\left(x_i^{(k)}\right)=\mathcal{N}\left(x_i^{(k)};\mu_i^{(k)},\sigma_i^{(k)}\right)$$

Variational inversion

The variational energy gives posteriors under the mean-field approximation that still violate our Gaussian assumptions:

$$\hat{q}\left(x_i^{(k)}\right) = \frac{1}{\mathcal{Z}_i} \exp\left(I\left(x_i^{(k)}\right)\right) \tag{15}$$

$$I\left(x_i^{(k)}\right) = \int_{X_{\setminus i}} q\left(x_{\setminus i}^{(k)}\right) \ln p\left(u^{(k)}, x_i^{(k)}, x_{\setminus i}^{(k)}, \chi \mid u^{(1\ldots k-1)}\right) \mathrm{d}x_{\setminus i}^{(k)}$$

$$q\left(x_{\setminus i}\right) = \prod_{j \neq i} q\left(x_j\right) \tag{16}$$

Variational inversion

This violation of distributional assumptions is obvious from the non-quadratic terms in the variational energy:

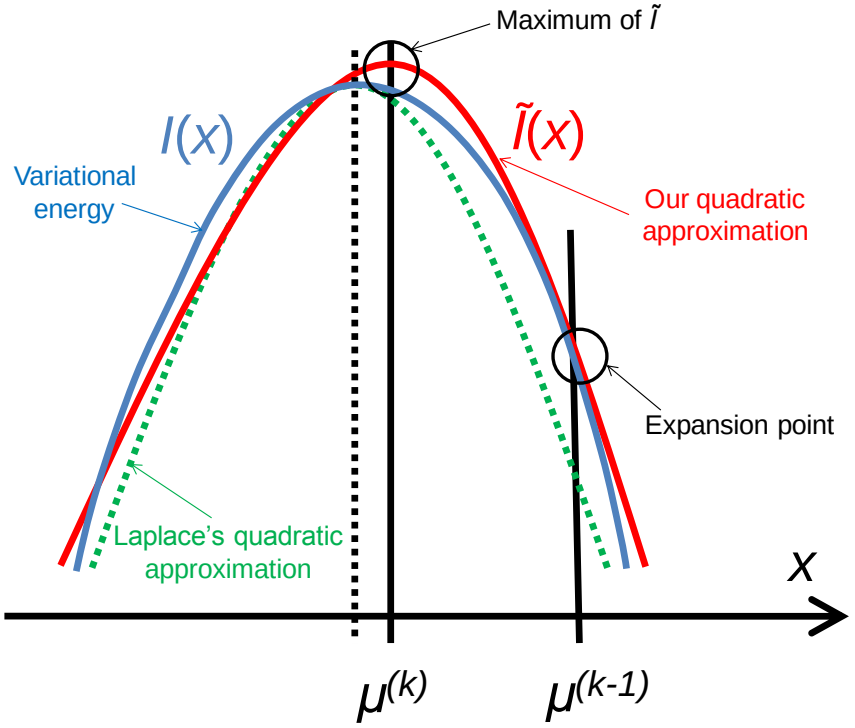
$$I\left(x_i^{(k)}\right) = \frac{1}{2}\ln \hat{\pi}_{i-1}^{(k)} - \frac{1}{2}\sigma_{i-1}^{+(k)}\hat{\pi}_{i-1}^{(k)} - \frac{1}{2}\hat{\pi}_i^{(k)}\left(x_i^{(k)} - \mu_i^{(k-1)}\right)^2$$

$$\hat{\pi}_i^{(k)} = \frac{1}{\sigma_i^{(k-1)} + \exp\left(\kappa_i x_{i+1}^{(k)} + \omega_i\right)}$$

$$\sigma_i^{+(k)} = \sigma_i^{(k)} + \left(\mu_i^{(k)} - \mu_i^{(k-1)}\right)^2$$

Variational inversion

- A quadratic approximation is found by expanding to second order about the expectation $\mu^{(k-1)}$.
- The update in the sufficient statistics of the approximate posterior is then performed by analytically finding the maximum of the quadratic approximation.



$$\sigma_i^{(k)} = -\frac{1}{\partial^2 I\left(\mu_i^{(k-1)}\right)}$$

$$\mu_i^{(k)} = \mu_i^{(k-1)} - \frac{\partial I\left(\mu_i^{(k-1)}\right)}{\partial^2 I\left(\mu_i^{(k-1)}\right)} = \mu_i^{(k-1)} + \sigma_i^{(k)} \partial I\left(\mu_i^{(k-1)}\right)$$

Mathys et al. (2011). *Front. Hum. Neurosci.*, 5:39.

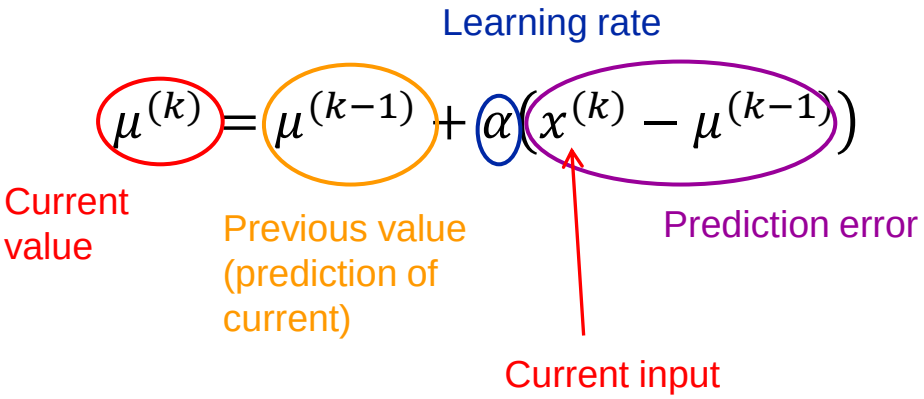
Comparison to Rescorla-Wagner

The resulting update equations for the posterior means of the x_i are similar to those of Rescorla-Wagner learning:

HB:

$$\mu_i^{(k)} = \mu_i^{(k-1)} + \frac{1}{2}\kappa_{i-1}\, v_{i-1}^{(k)} \cdot \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \cdot \delta_{i-1}^{(k)}$$

RW:



Furthermore, the updates are **precision-weighted prediction errors**.

Precision-weighting of updates

- Updates are weighted by belief precisions.
- To see this, first consider a simple non-hierarchical model with one parameter ϑ.
- Likelihood and prior are Gaussian, therefore the posterior also:

$$p(\vartheta) = \mathcal{N}(\vartheta; \mu_\vartheta, \pi_\vartheta) \qquad \textbf{Prior}$$

$$p(y|\vartheta) = \mathcal{N}(y; \vartheta, \pi_\varepsilon) \qquad \textbf{Likelihood}$$

$$\Longrightarrow \quad p(\vartheta|y) = \mathcal{N}\left(\vartheta; \mu_{\vartheta|y}, \pi_{\vartheta|y}\right) \qquad \textbf{Posterior}$$

- The exact Bayesian update then is the precision-weighted prediction error:

$$\pi_{\vartheta|y} = \pi_\vartheta + \pi_\varepsilon \qquad \mu_{\vartheta|y} = \mu_\vartheta + \frac{\pi_\varepsilon}{\pi_{\vartheta|y}}(y - \mu_\vartheta)$$

Precision-weighted prediction error

Precision-weighting of updates

Comparison to the simple non-hierarchical Bayesian update:

HB:

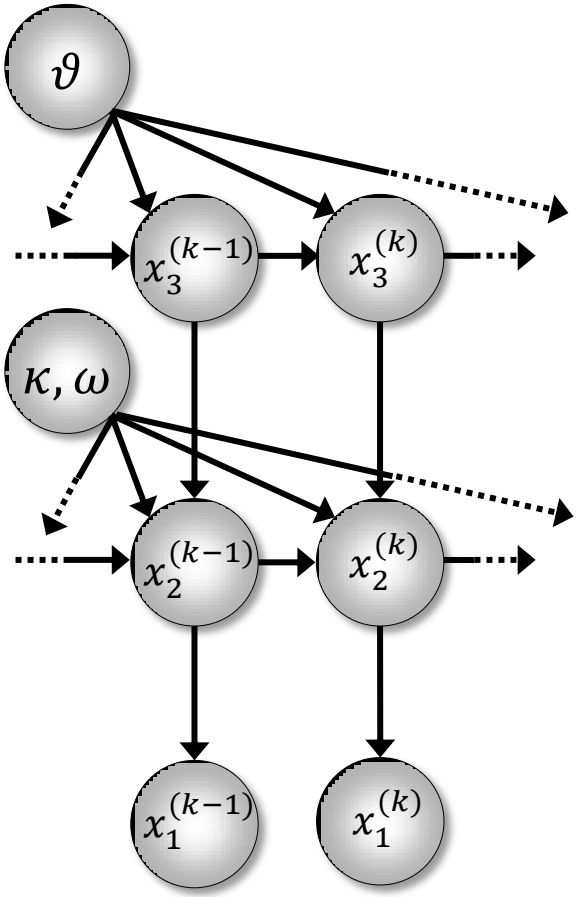
$$\mu_i^{(k)} = \mu_i^{(k-1)} + \frac{1}{2}\kappa_{i-1}\, v_{i-1}^{(k)} \cdot \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \cdot \delta_{i-1}^{(k)}$$

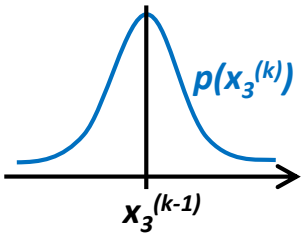
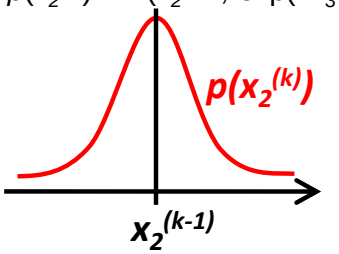
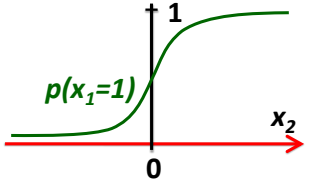
Precision-weighted prediction error

nHB:

$$\mu_{\vartheta|y} = \mu_\vartheta + \frac{\pi_\varepsilon}{\pi_{\vartheta|y}}(y - \mu_\vartheta)$$

Application to binary data



State of the world	Model	
Log-volatility x_3 of tendency	Gaussian random walk with constant step size ϑ	$p(x_3^{(k)}) \sim N(x_3^{(k-1)}, \vartheta)$ 
Tendency x_2 towards category "1"	Gaussian random walk with step size $\exp(\kappa x_3+\omega)$	$p(x_2^{(k)}) \sim N(x_2^{(k-1)}, \exp(\kappa x_3+\omega))$ 
Stimulus category x_1 ("0" or "1")	Sigmoid trans-formation of x_2	$p(x_1=1) = s(x_2)$ $p(x_1=0) = 1-s(x_2)$ 

Mathys et al. (2011). *Front. Hum. Neurosci.*, 5:39.

Update equation for binary observations

- $x_1 \in \{0,1\}$ is observed by the agent. Each observation leads to an update in the belief on x_2, x_3, ..., and so on up the hierarchy.
- The updates for x_2 can be derived in the same manner as above.

$$I\left(x_2^{(k)}\right) = \ln s\left(x_2^{(k)}\right) + x_2^{(k)}\left(x_1^{(k)} - 1\right) - \frac{1}{2}\hat{\pi}_2^{(k)}\left(x_2^{(k)} - \mu_2^{(k-1)}\right)^2$$

$$\mu_2^{(k)} = \mu_2^{(k-1)} + \sigma_2^{(k)}\delta_1^{(k)}$$

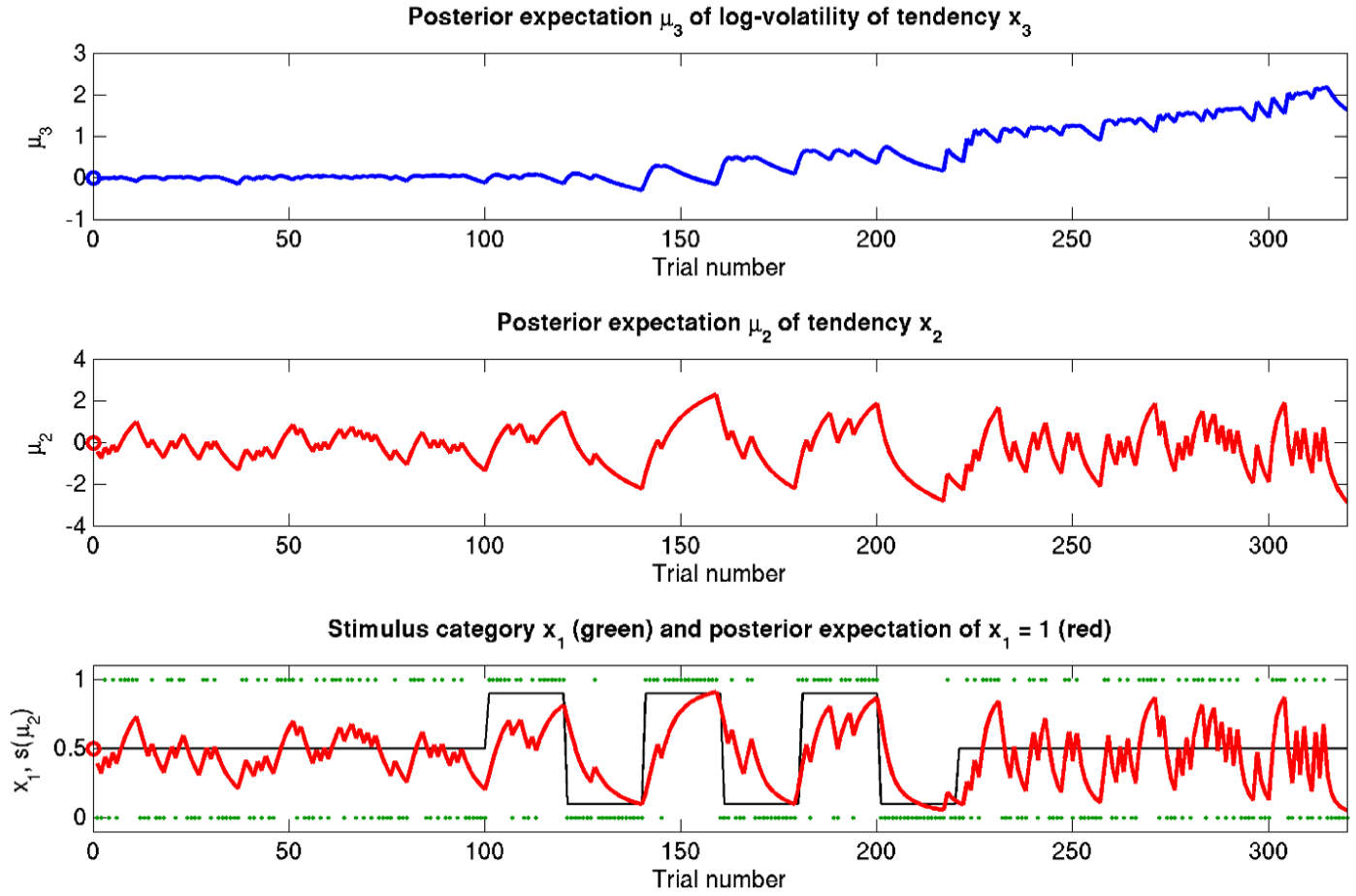
- At first, this simply looks like an uncertainty-weighted update. However, when we unpack σ_2 and do a Taylor expansion in powers of $\hat{\pi}_1$, we see that it is again proportional to the precision of the prediction on the level below:

$$\sigma_2^{(k)} = \frac{\hat{\pi}_1^{(k)}}{\hat{\pi}_2^{(k)}\hat{\pi}_1^{(k)} + 1} = \hat{\pi}_1^{(k)} - \hat{\pi}_2^{(k)}\left(\hat{\pi}_1^{(k)}\right)^2 + \left(\hat{\pi}_2^{(k)}\right)^2\left(\hat{\pi}_1^{(k)}\right)^3 + O(4)$$

- At all higher levels, the updates are as previously derived.

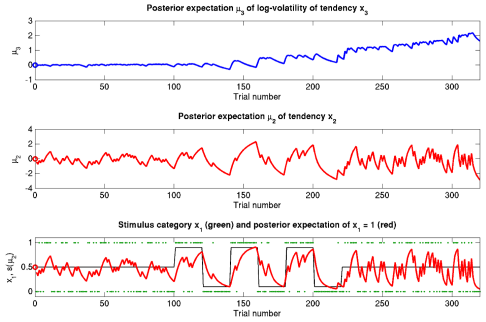
Simulation

The reference scenario: $\vartheta = 0.5,\ \omega = -2.2,\ \kappa = 1.4$

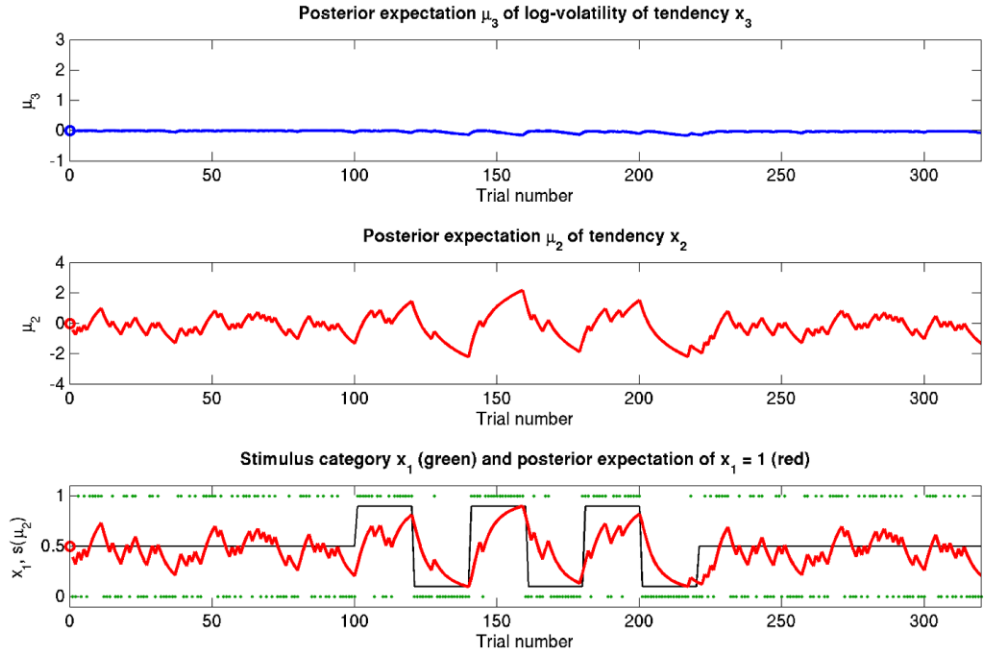


Simulations

Reference

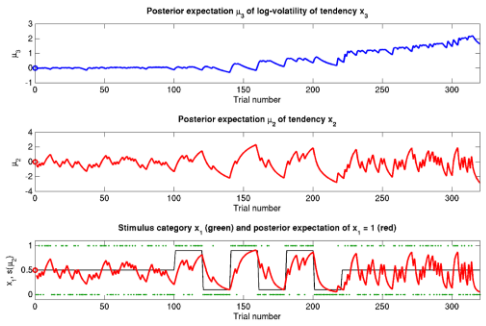


Reduced ϑ: $\vartheta = 0.05,\ \omega = -2.2,\ \kappa = 1.4$

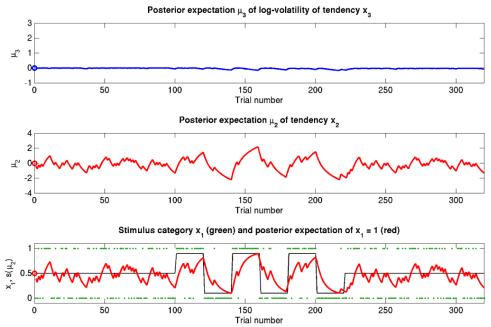


Simulations

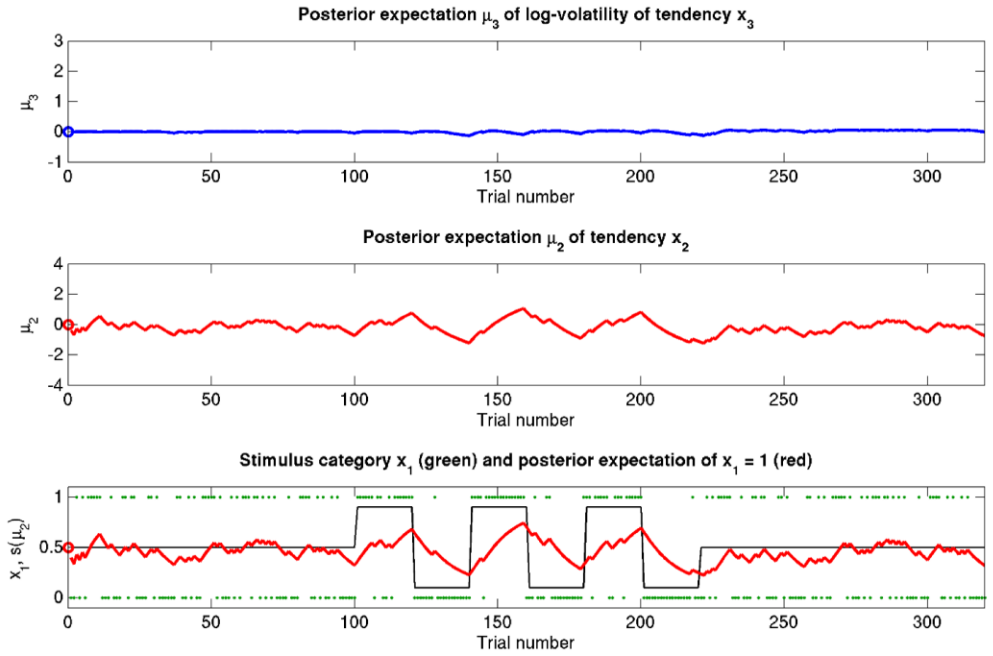
Reference



Reduced ϑ

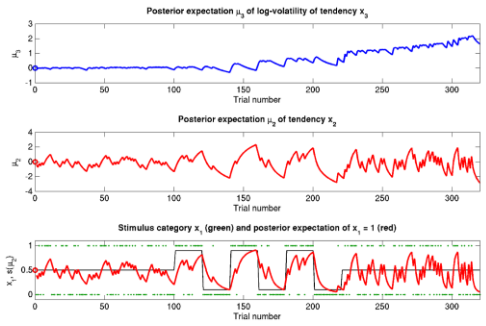


Reduced ω: $\vartheta = 0.5,\ \omega = -4,\ \kappa = 1.4$

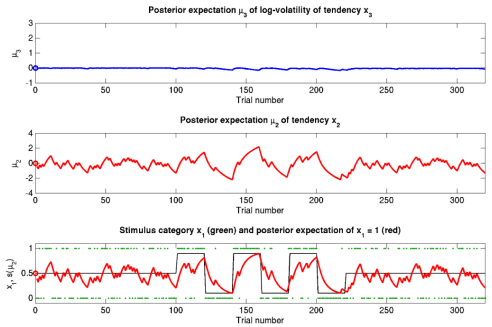


Simulations

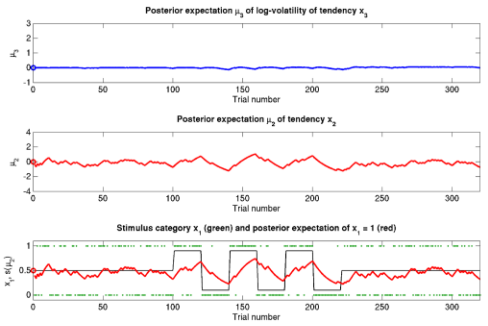
Reference



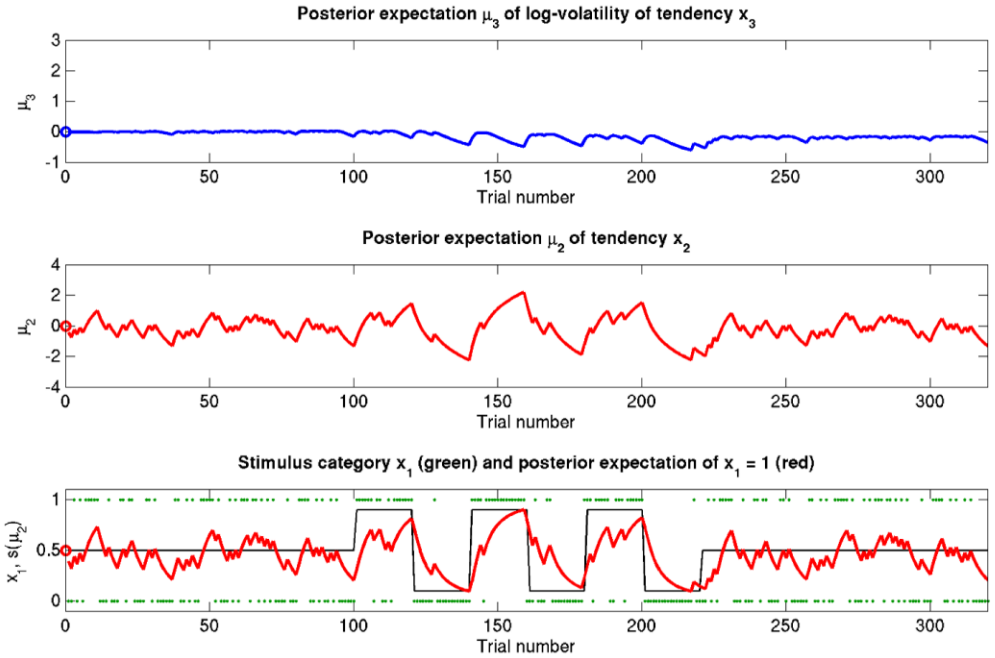
Reduced ϑ



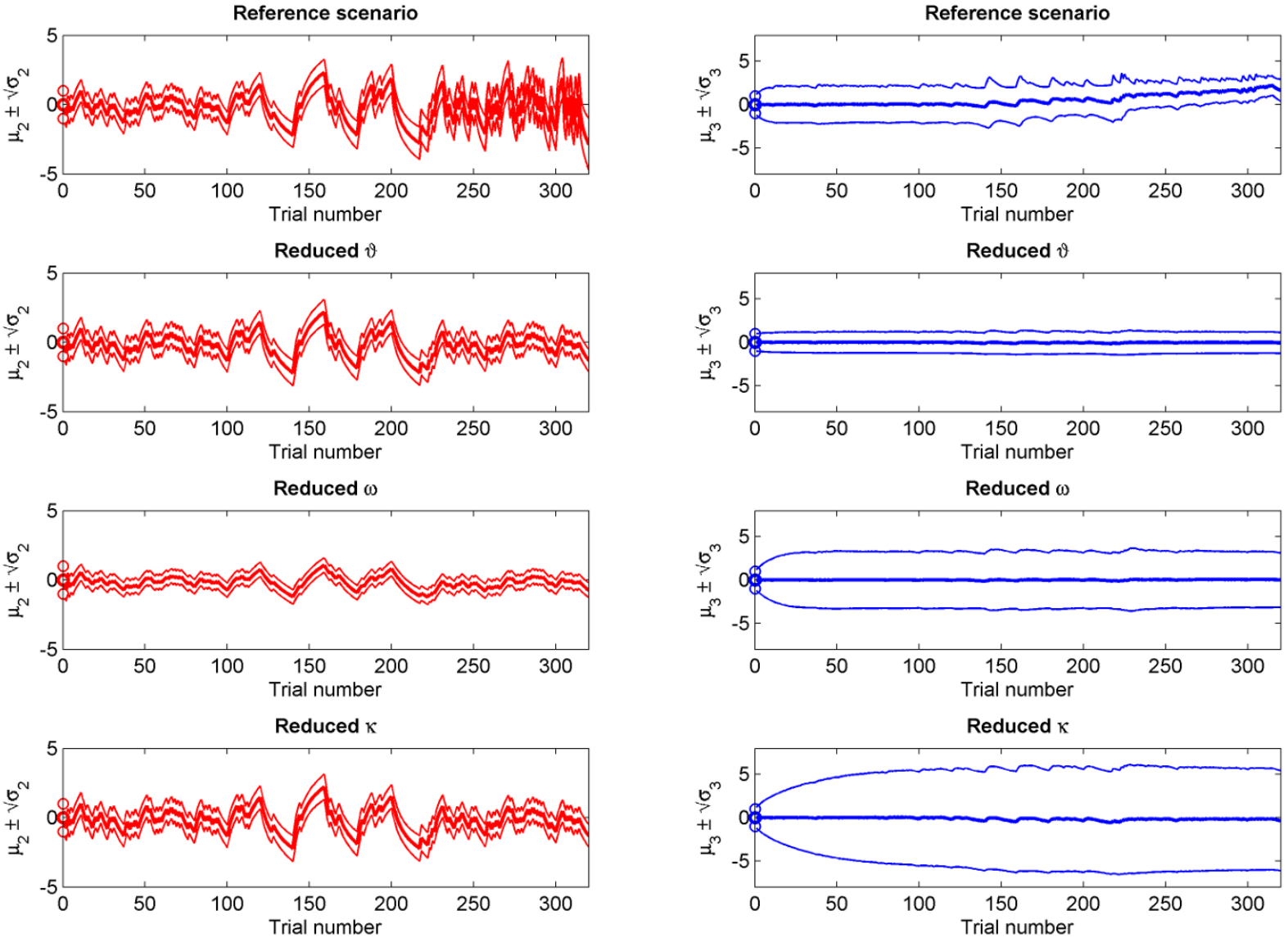
Reduced ω



Reduced κ: $\vartheta = 0.5,\ \omega = -4,\ \kappa = 0.2$



Simulations

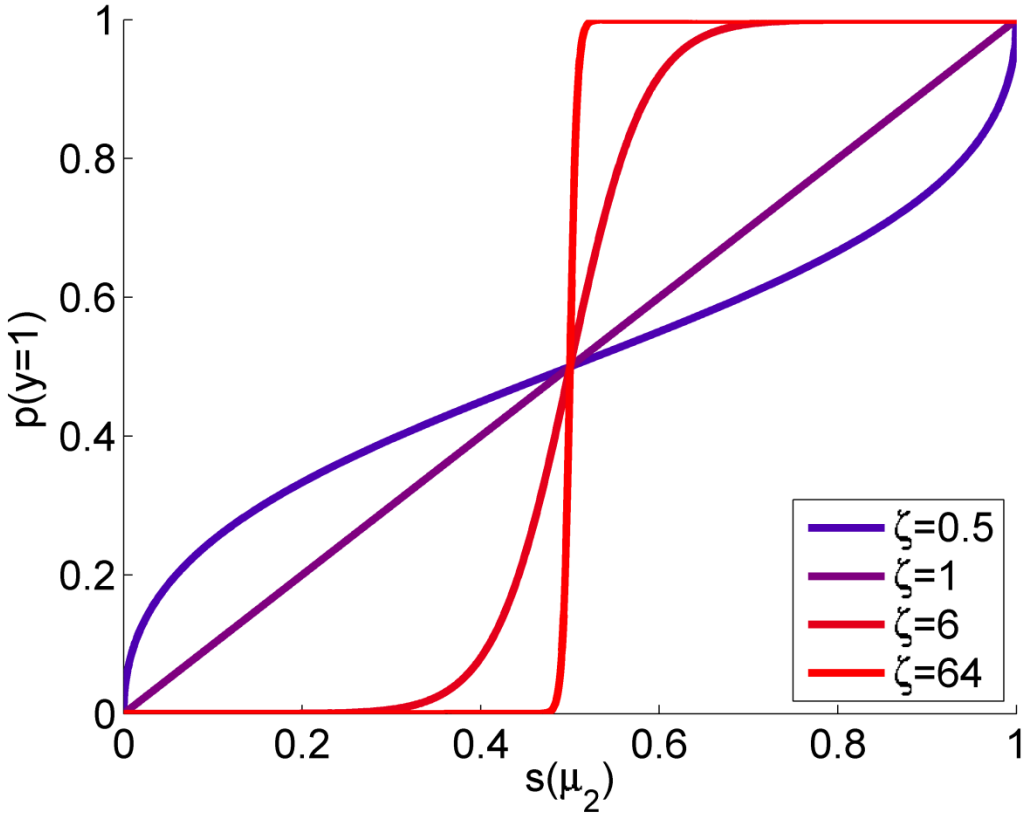


Decision models

- Softmax decision rule
- Curve shape is determined by the parameter ζ
- Translates beliefs into decision probabilities

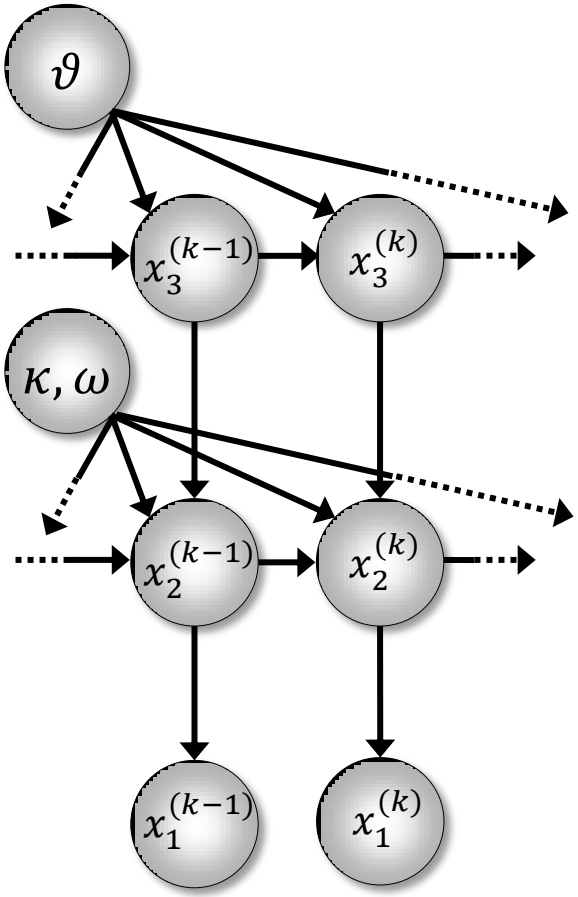
Probability of decision “1”,

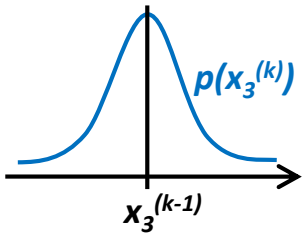
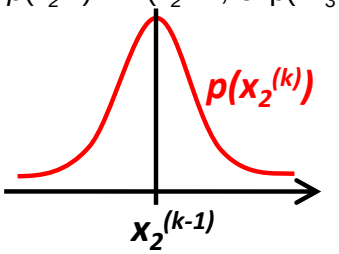
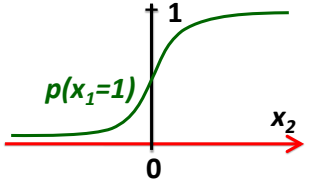
(i.e., of betting on “1”)



Prediction that next stimulus is “1”

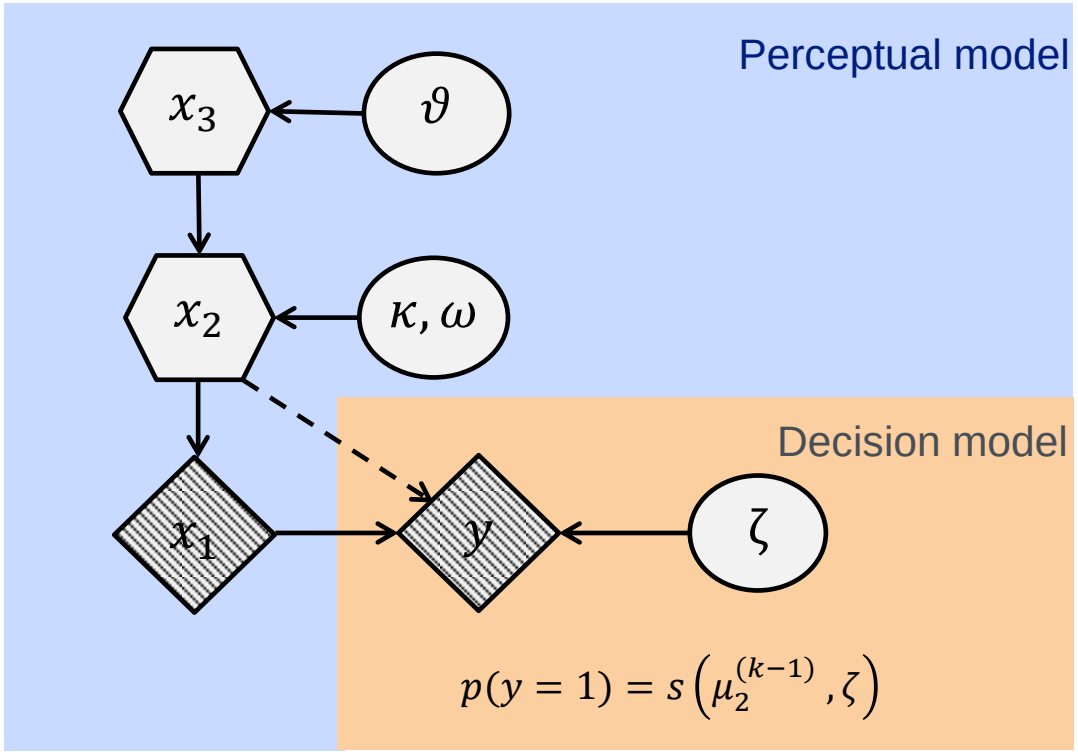
The perceptual model (reminder)



State of the world	Model	
Log-volatility x_3 of tendency	Gaussian random walk with constant step size ϑ	$p(x_3^{(k)}) \sim N(x_3^{(k-1)}, \vartheta)$ 
Tendency x_2 towards category "1"	Gaussian random walk with step size $\exp(\kappa x_3+\omega)$	$p(x_2^{(k)}) \sim N(x_2^{(k-1)}, \exp(\kappa x_3+\omega))$ 
Stimulus category x_1 ("0" or "1")	Sigmoid trans-formation of x_2	$p(x_1=1) = s(x_2)$ $p(x_1=0) = 1-s(x_2)$ 

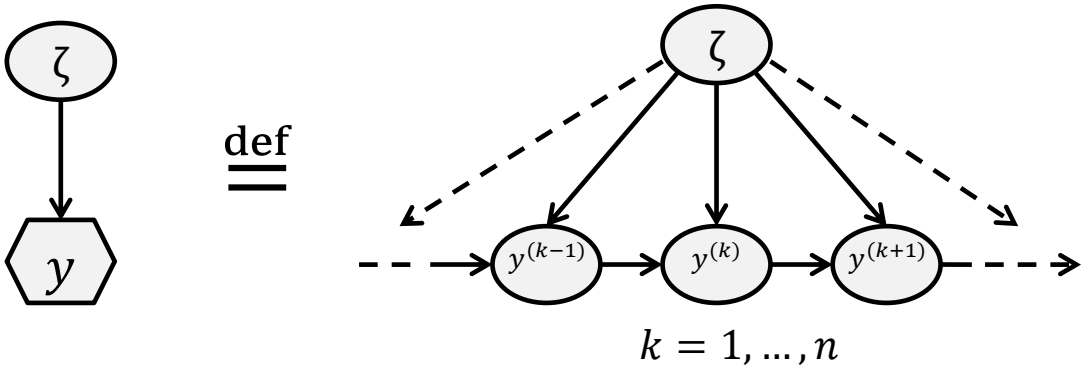
Mathys et al. (2011). *Front. Hum. Neurosci.*, 5:39.

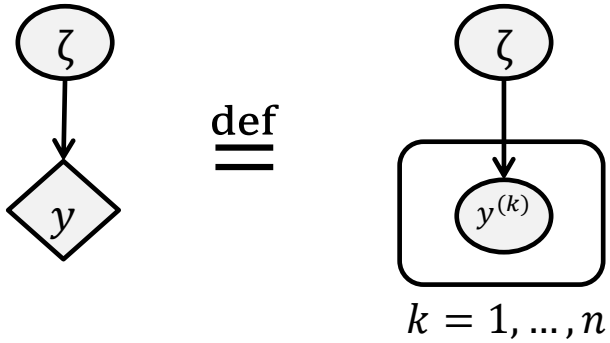
Taking it all together: perception and decision

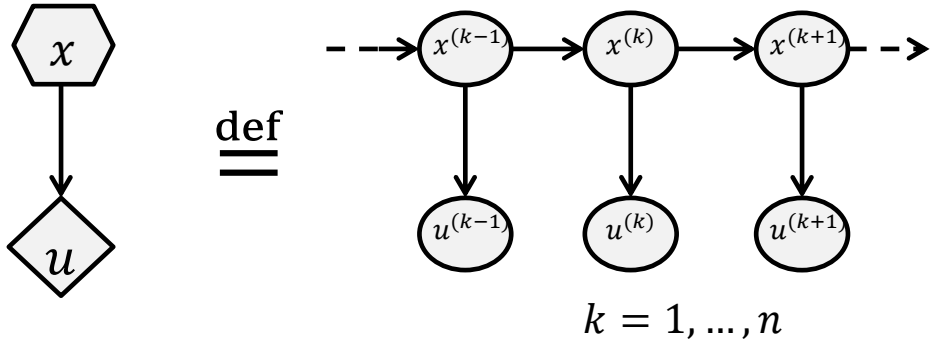


cf. Daunizeau et al. (2010a,b)

Taking it all together: notation

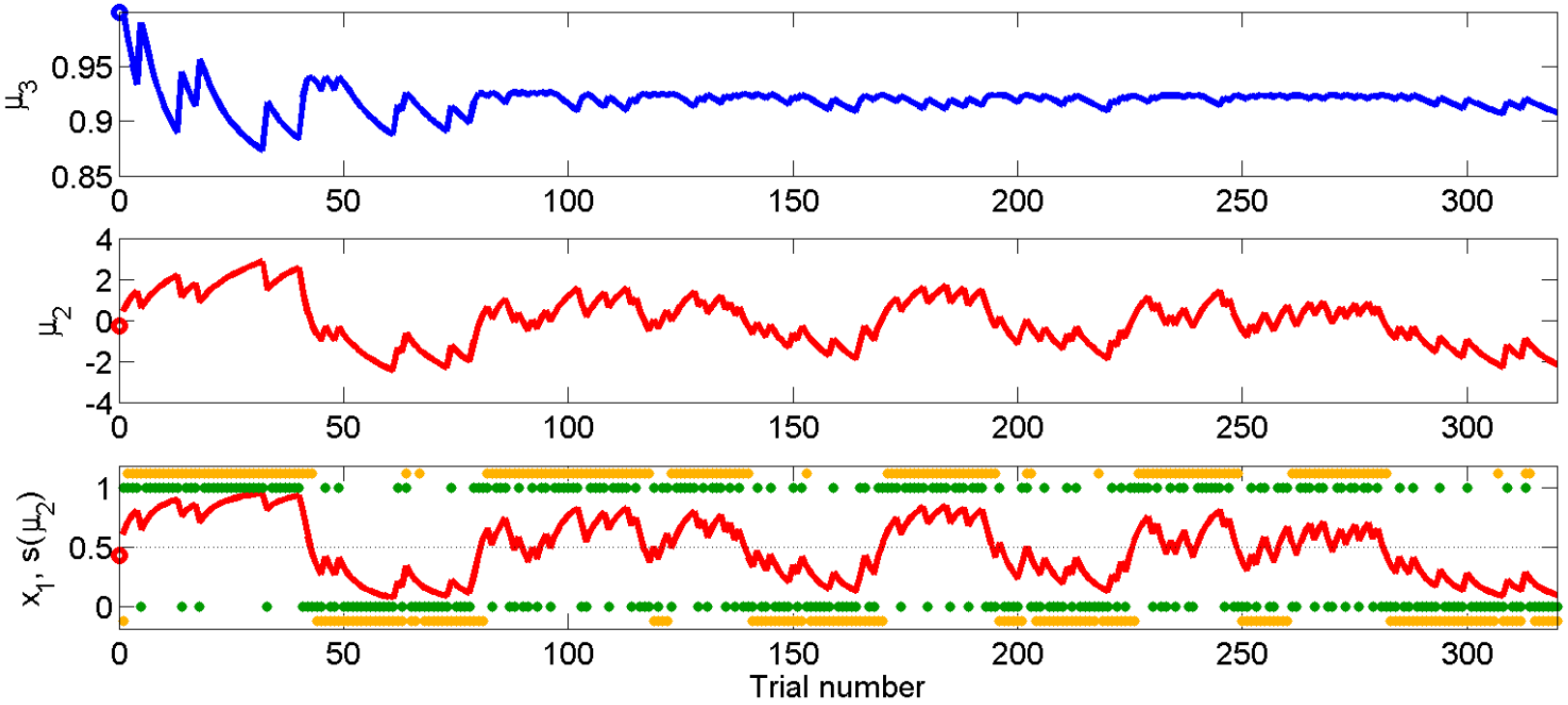






Parameter estimation

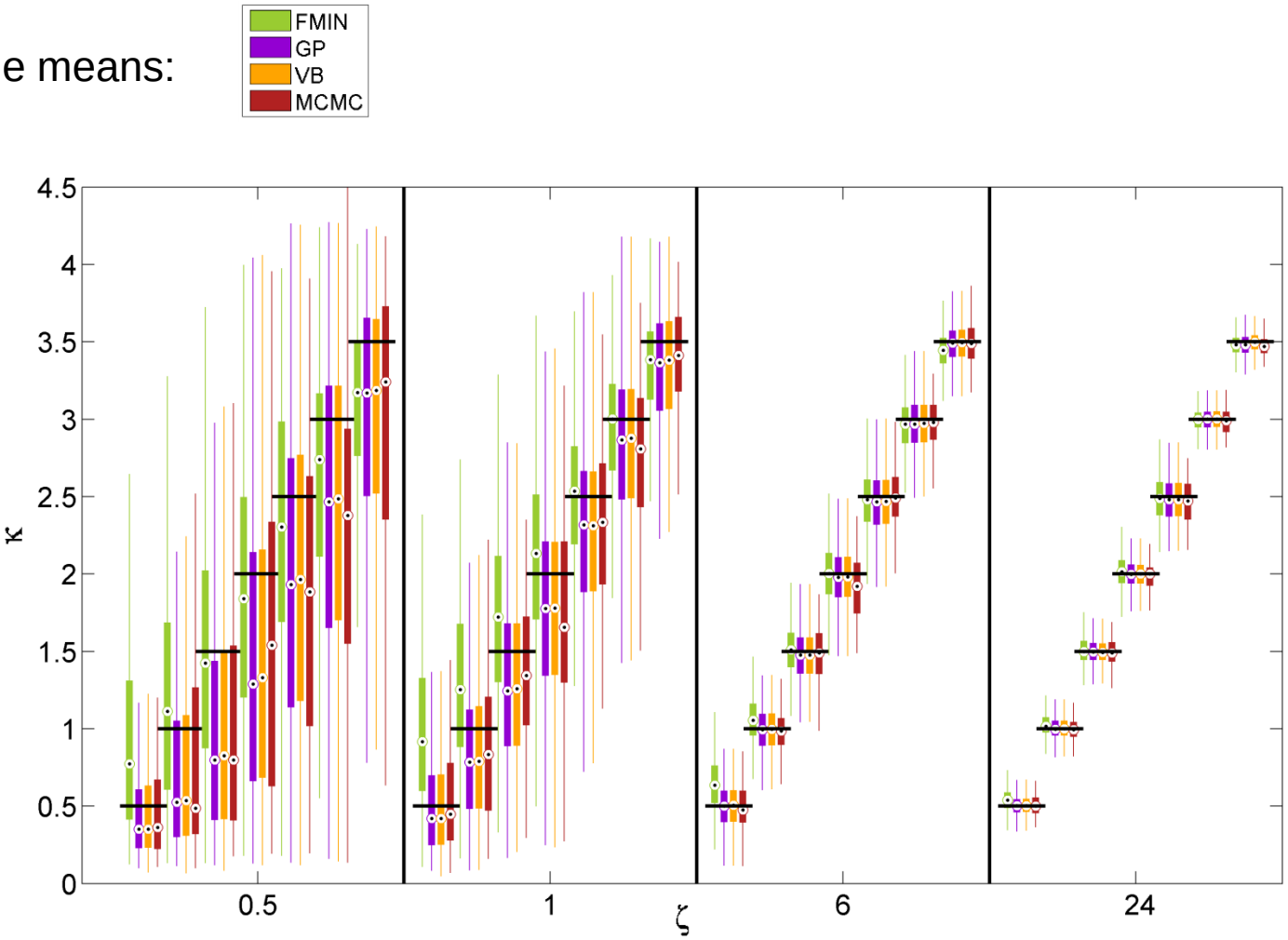
The goal:



$$\kappa = 2.49$$
$$\sigma_3^{(0)} = 0.988$$
$$\vartheta = 0.000592$$

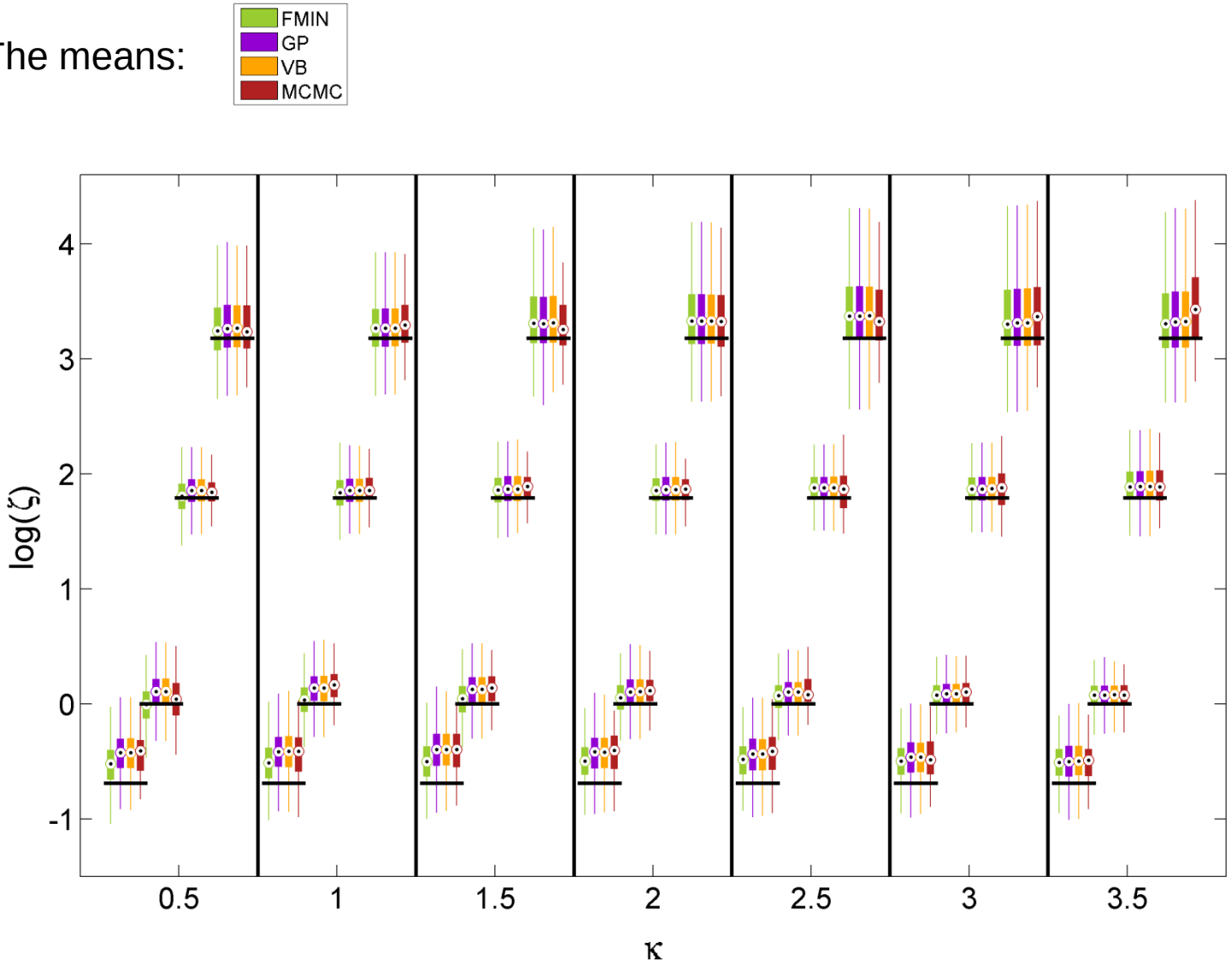
Parameter estimation

The means:

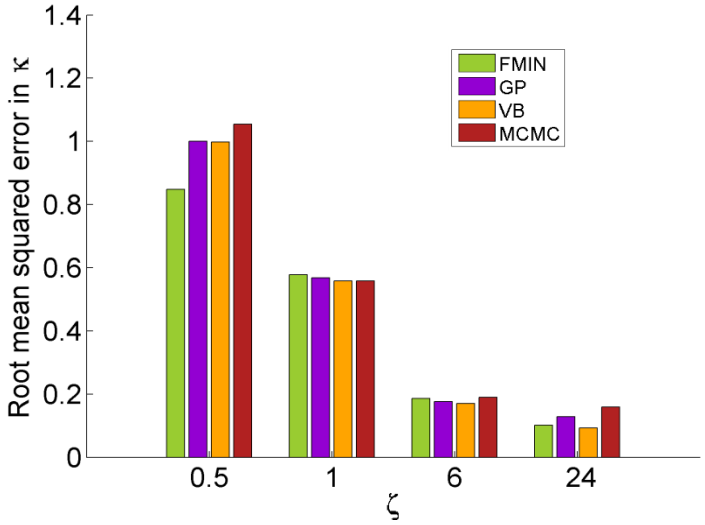


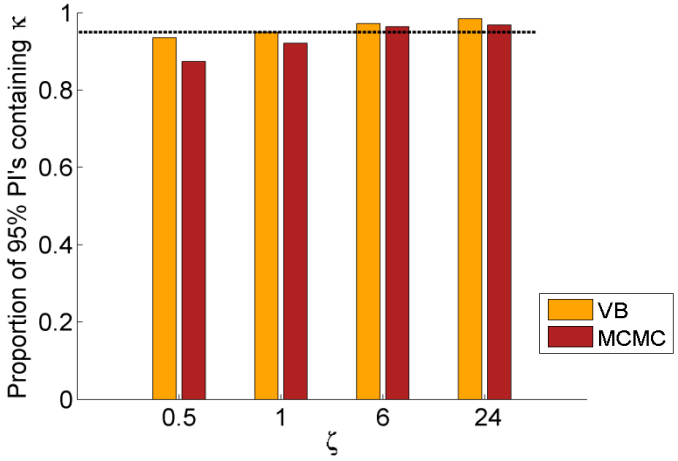
Parameter estimation

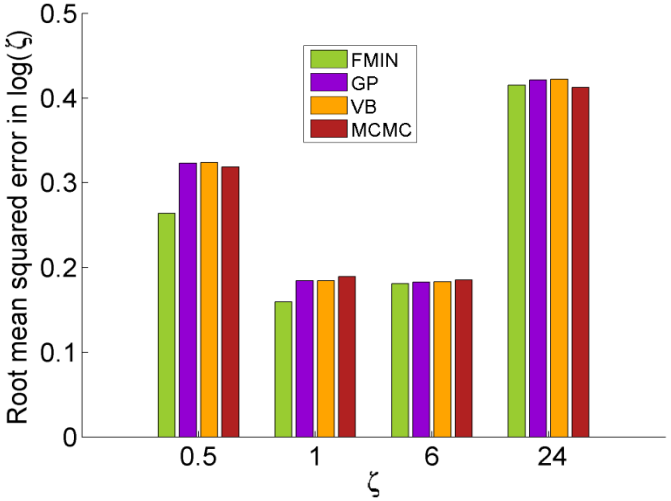
The means:

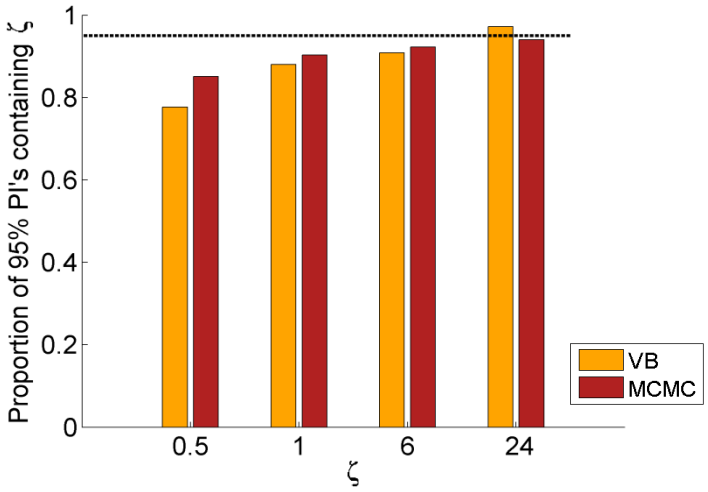


Parameter estimation



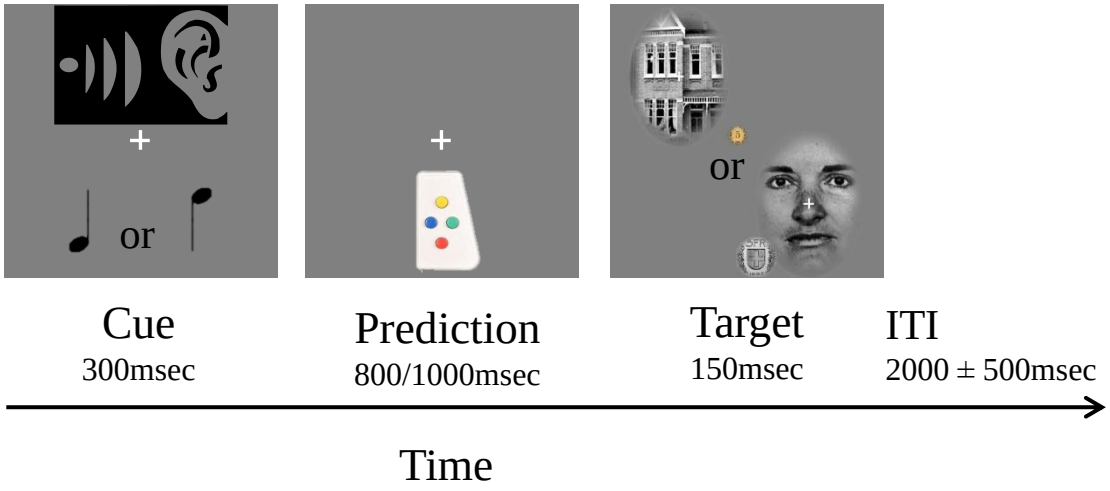




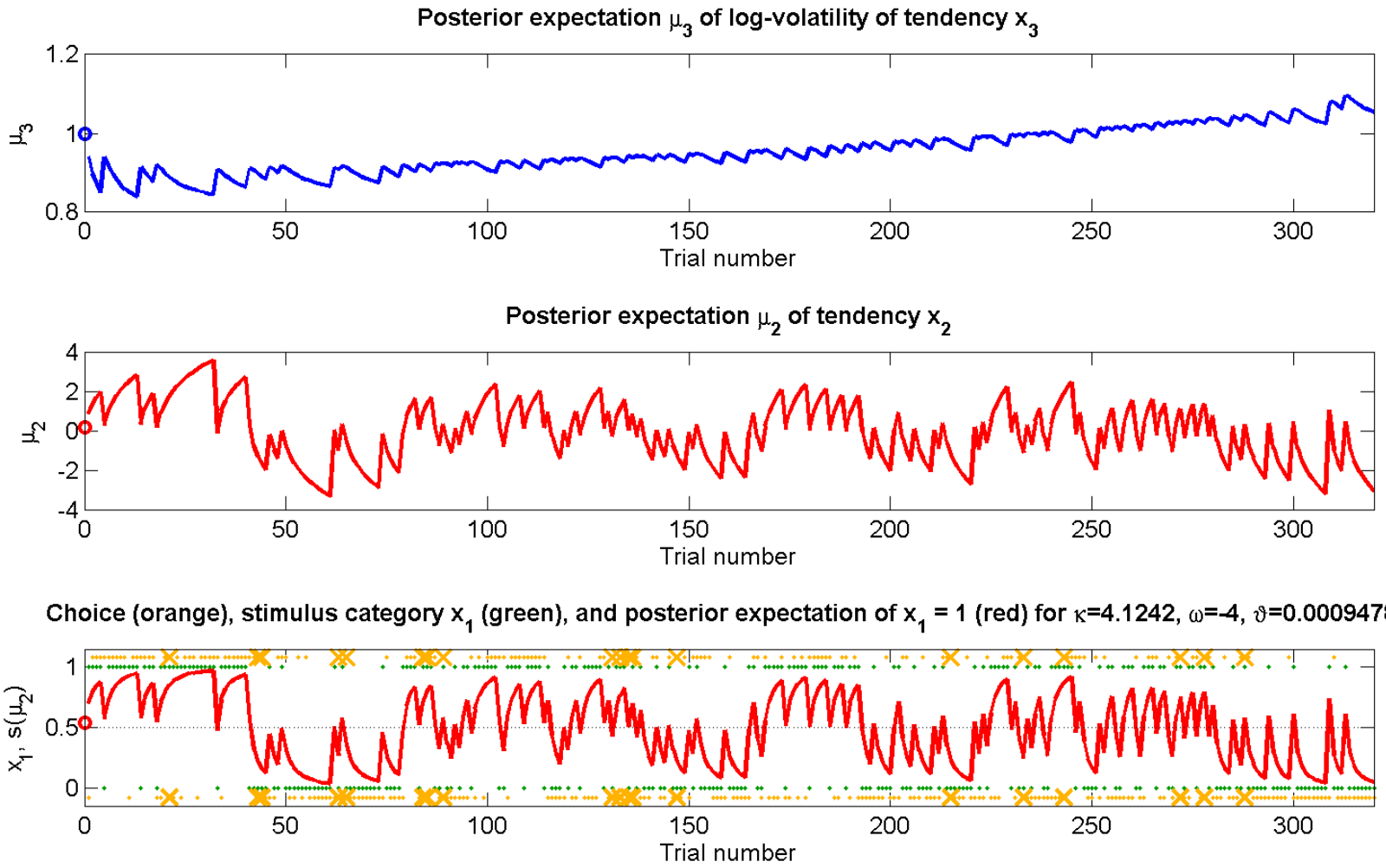


Applications (Iglesias et al., in prep.)

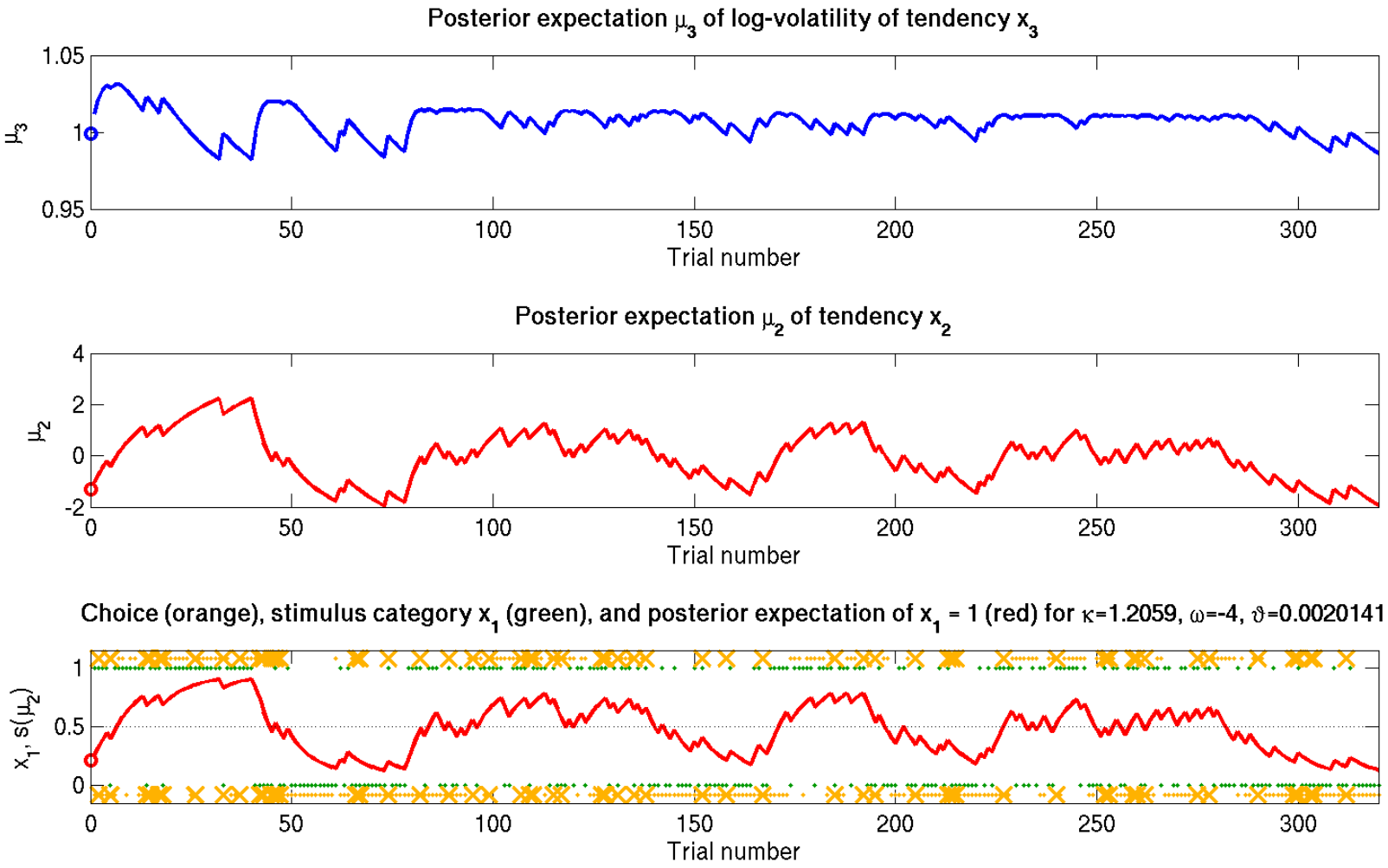
Associative learning task:



Applications (Iglesias et al., in prep.)

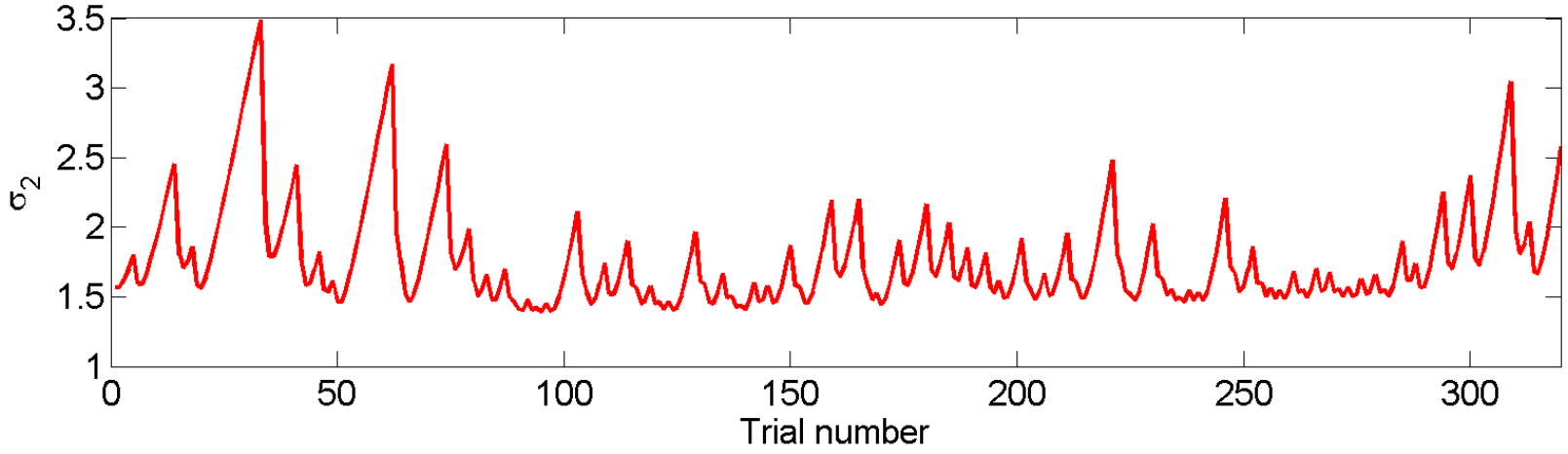


Applications (Iglesias et al., in prep.)



Applications (Iglesias et al., in prep.)

Regressor: uncertainty (i.e., σ_2)



Update equation for binary observations (reminder)

- $x_1 \in \{0,1\}$ is observed by the agent. Each observation leads to an update in the belief on x_2, x_3, ..., and so on up the hierarchy.
- The updates for x_2 can be derived in the same manner as above.

$$I\left(x_2^{(k)}\right) = \ln s\left(x_2^{(k)}\right) + x_2^{(k)}\left(x_1^{(k)} - 1\right) - \frac{1}{2}\hat{\pi}_2^{(k)}\left(x_2^{(k)} - \mu_2^{(k-1)}\right)^2$$

$$\mu_2^{(k)} = \mu_2^{(k-1)} + \sigma_2^{(k)}\delta_1^{(k)}$$

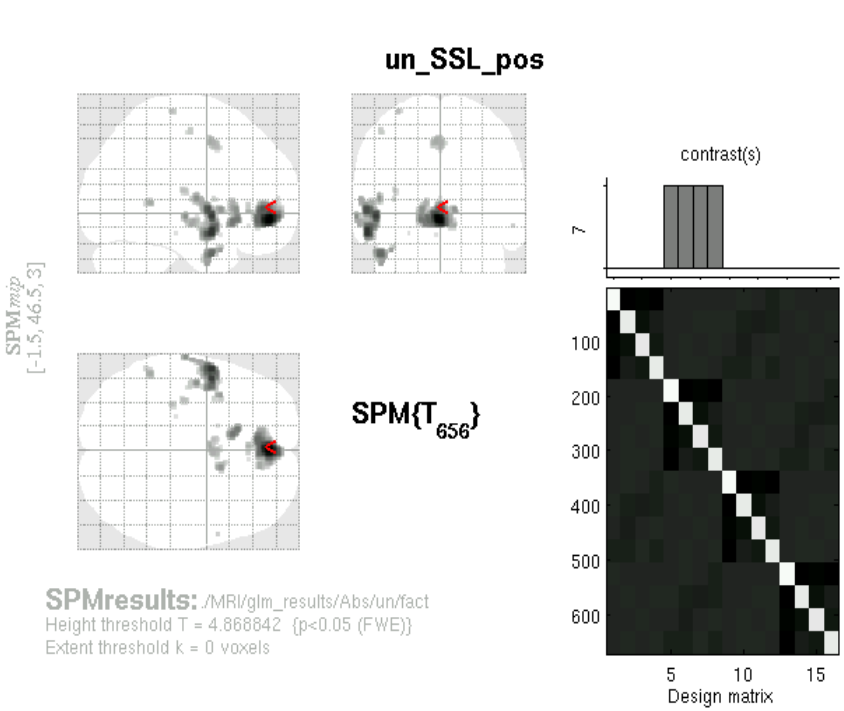
- At first, this simply looks like an uncertainty-weighted update. However, when we unpack σ_2 and do a Taylor expansion in powers of $\hat{\pi}_1$, we see that it is again proportional to the precision of the prediction on the level below:

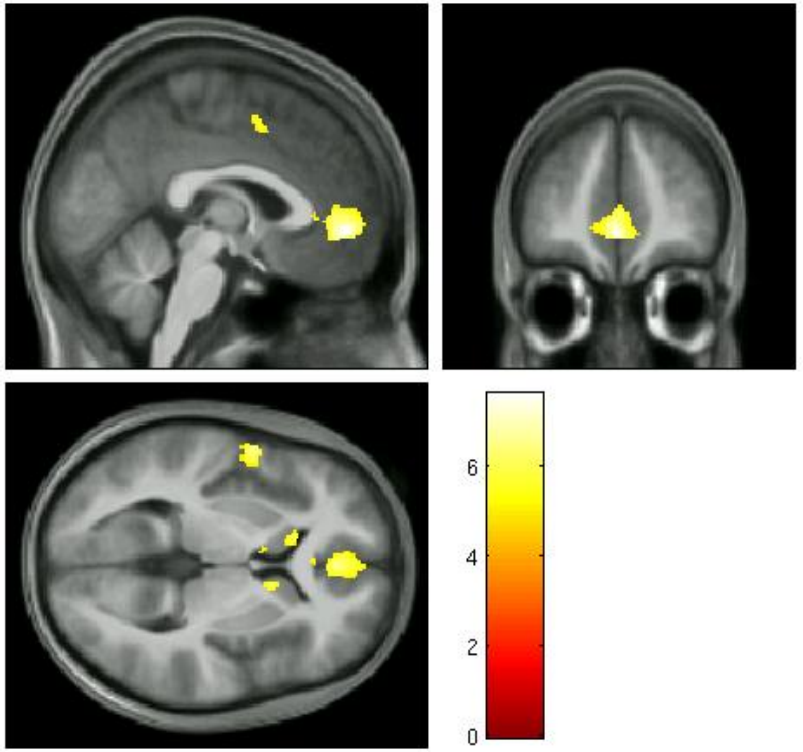
$$\sigma_2^{(k)} = \frac{\hat{\pi}_1^{(k)}}{\hat{\pi}_2^{(k)}\hat{\pi}_1^{(k)} + 1} = \hat{\pi}_1^{(k)} - \hat{\pi}_2^{(k)}\left(\hat{\pi}_1^{(k)}\right)^2 + \left(\hat{\pi}_2^{(k)}\right)^2\left(\hat{\pi}_1^{(k)}\right)^3 + O(4)$$

- At all higher levels, the updates are as previously derived.

Applications (Iglesias et al., in prep.)

Effect of uncertainty



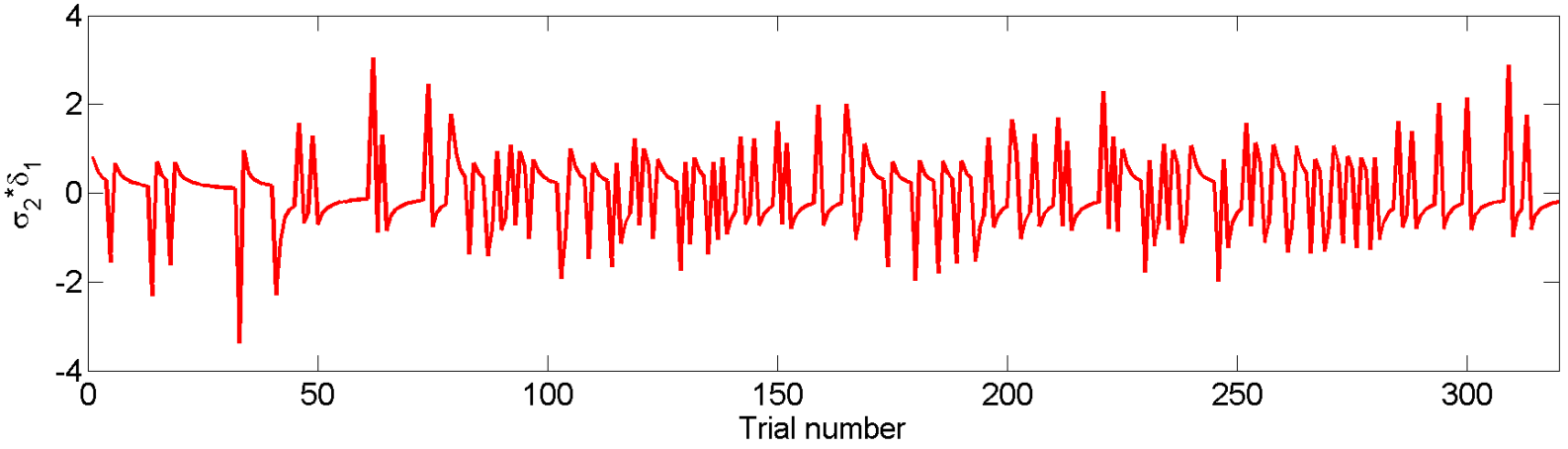


N=42, p<0.05 (FWE, whole brain)

x=-1.5, y=47, z=3

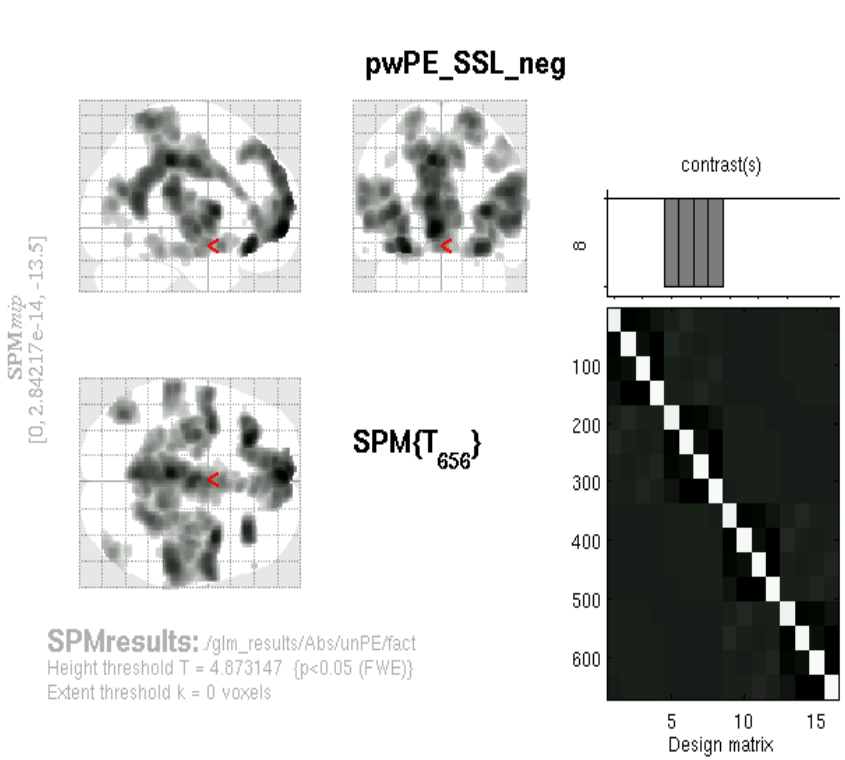
Applications (Iglesias et al., in prep.)

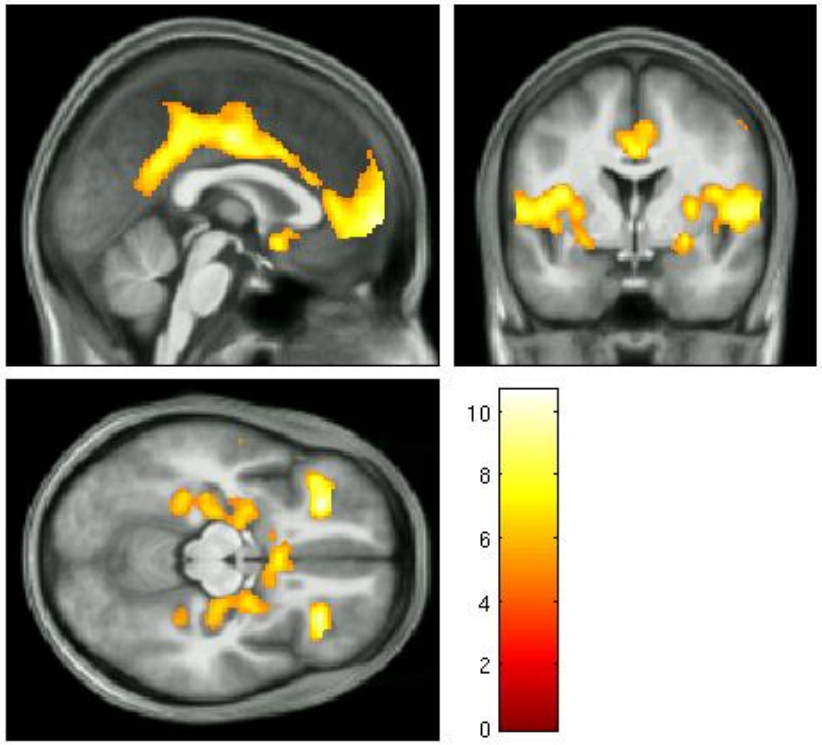
Regressor: uncertainty-weighted prediction error (i.e., $\sigma_2 \cdot \delta_1$)



Applications (Iglesias et al., in prep.)

Negative correlation with uncertainty-weighted prediction error



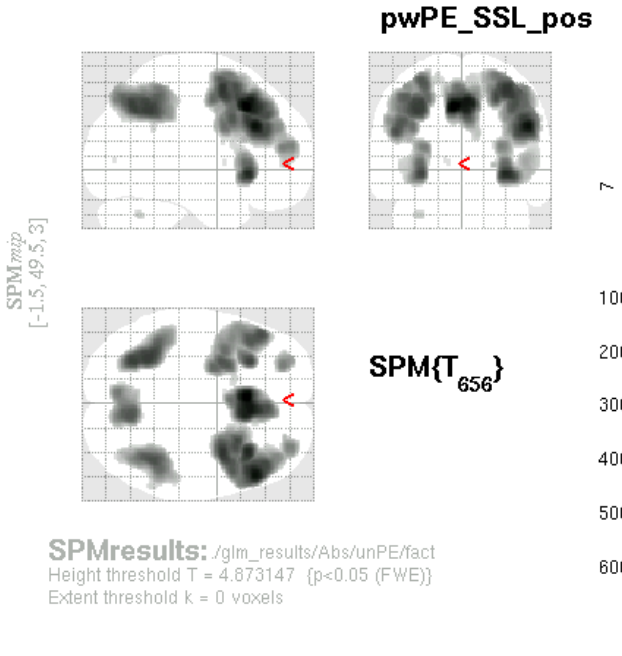


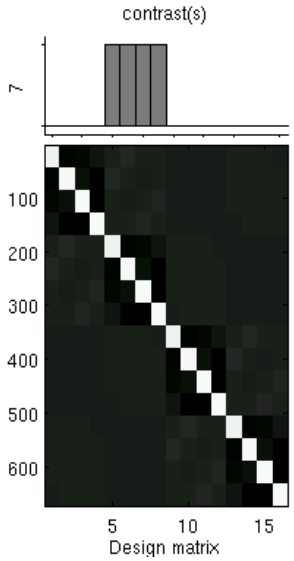
N=42, p<0.05 (FWE, whole brain)

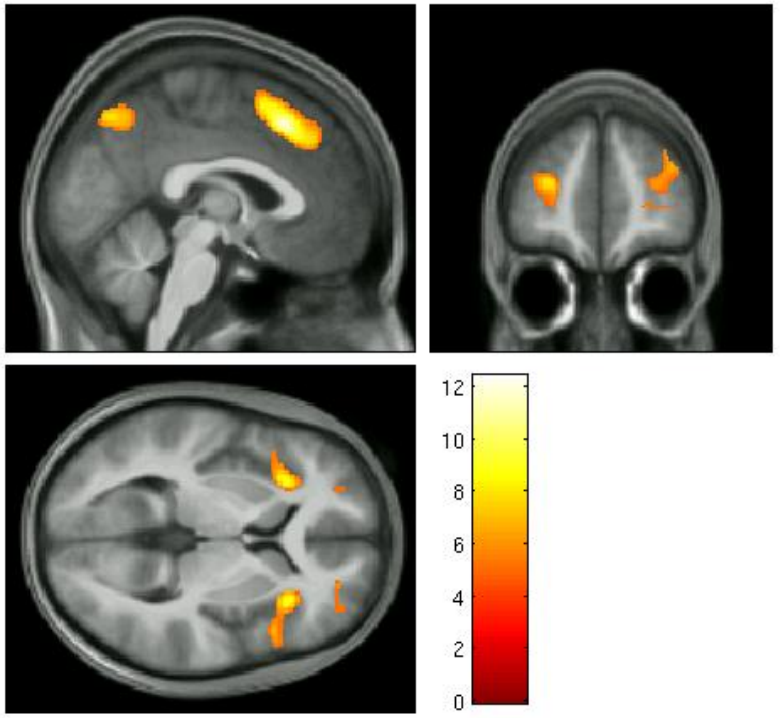
x=0, y=0, z=-14

Applications (Iglesias et al., in prep.)

Positive correlation with uncertainty-weighted prediction error



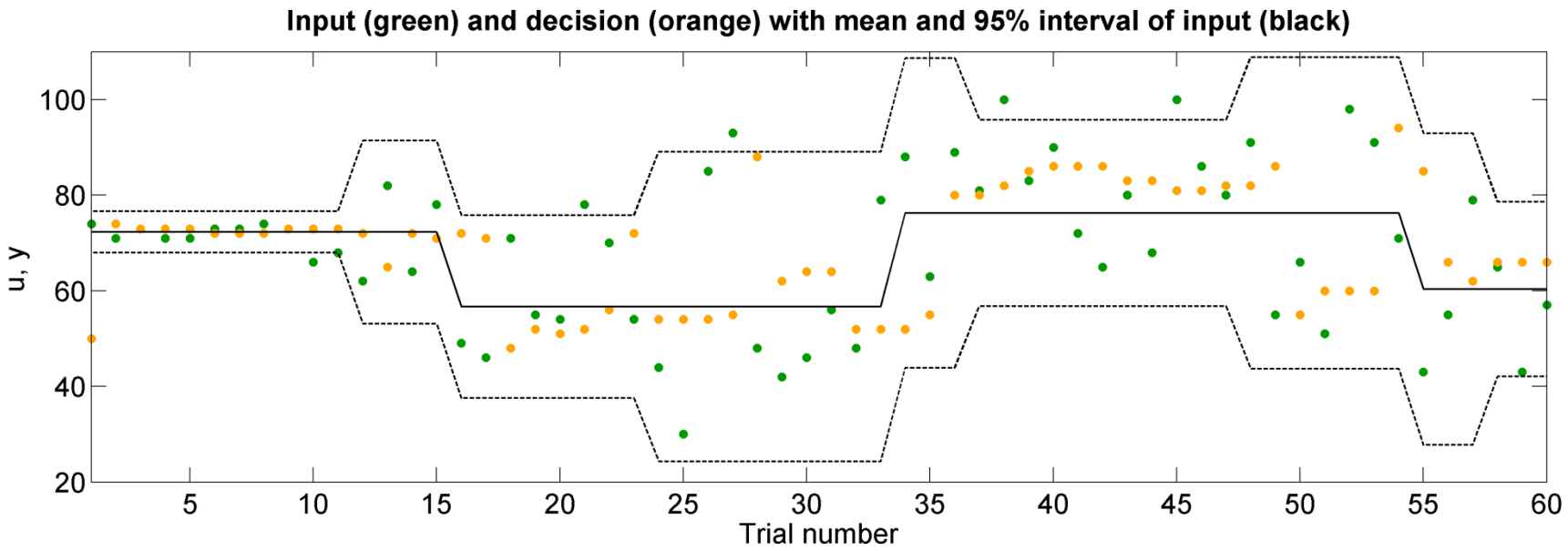




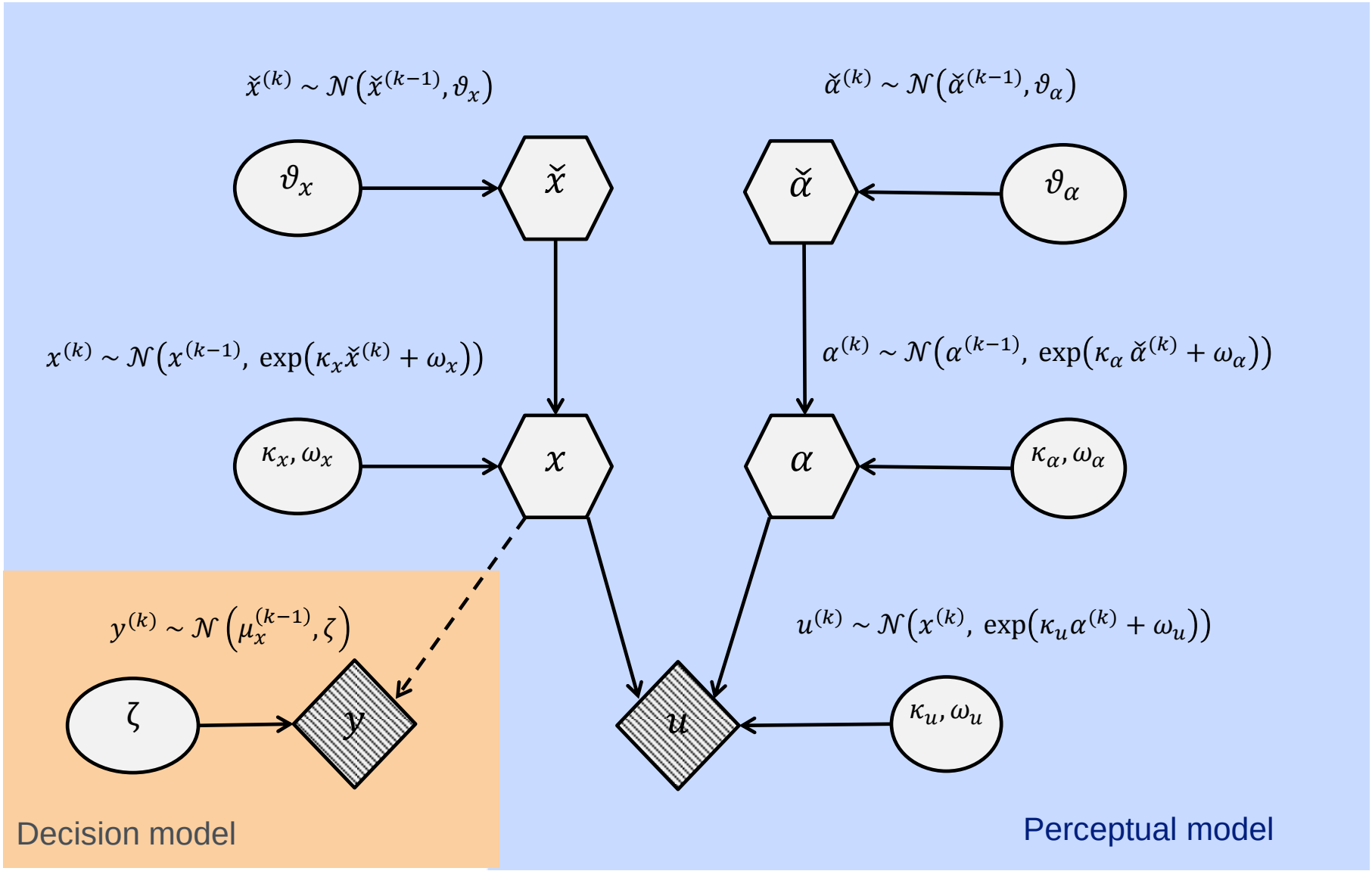
N=42, p<0.05 (FWE, whole brain)

x=-1.5, y=49, z=3

Extensions (Guo et al., in prep.)



Extensions


$\check{x}^{(k)} \sim \mathcal{N}\left(\check{x}^{(k-1)}, \vartheta_x\right)$
$\check{\alpha}^{(k)} \sim \mathcal{N}\left(\check{\alpha}^{(k-1)}, \vartheta_\alpha\right)$
ϑ_x
$\check{x}$
$\check{\alpha}$
ϑ_α
$x^{(k)} \sim \mathcal{N}\left(x^{(k-1)}, \exp\left(\kappa_x \check{x}^{(k)} + \omega_x\right)\right)$
$\alpha^{(k)} \sim \mathcal{N}\left(\alpha^{(k-1)}, \exp\left(\kappa_\alpha \check{\alpha}^{(k)} + \omega_\alpha\right)\right)$
κ_x, ω_x
x
α
$\kappa_\alpha, \omega_\alpha$
$y^{(k)} \sim \mathcal{N}\left(\mu_x^{(k-1)}, \zeta\right)$
$u^{(k)} \sim \mathcal{N}\left(x^{(k)}, \exp\left(\kappa_u \alpha^{(k)} + \omega_u\right)\right)$
ζ
y
u
κ_u, ω_u
Decision model
Perceptual model

Extensions

$$\pi_{\breve{x}}^{(k)} = \hat{\pi}_{\breve{x}}^{(k)} + \frac{\kappa_x^2}{2} w_x^{(k)} \left(w_x^{(k)} + r_x^{(k)} \delta_x^{(k)} \right)$$

$$\mu_{\breve{x}}^{(k)} = \mu_{\breve{x}}^{(k-1)} + \frac{\kappa_x}{2} \frac{w_x^{(k)}}{\pi_{\breve{x}}^{(k)}} \delta_x^{(k)}$$

$$\pi_{\breve{\alpha}}^{(k)} = \hat{\pi}_{\breve{\alpha}}^{(k)} + \frac{\kappa_\alpha^2}{2} w_\alpha^{(k)} \left(w_\alpha^{(k)} + r_\alpha^{(k)} \delta_\alpha^{(k)} \right)$$

$$\mu_{\breve{\alpha}}^{(k)} = \mu_{\breve{\alpha}}^{(k-1)} + \frac{\kappa_\alpha}{2} \frac{w_\alpha^{(k)}}{\pi_{\breve{\alpha}}^{(k)}} \delta_\alpha^{(k)}$$

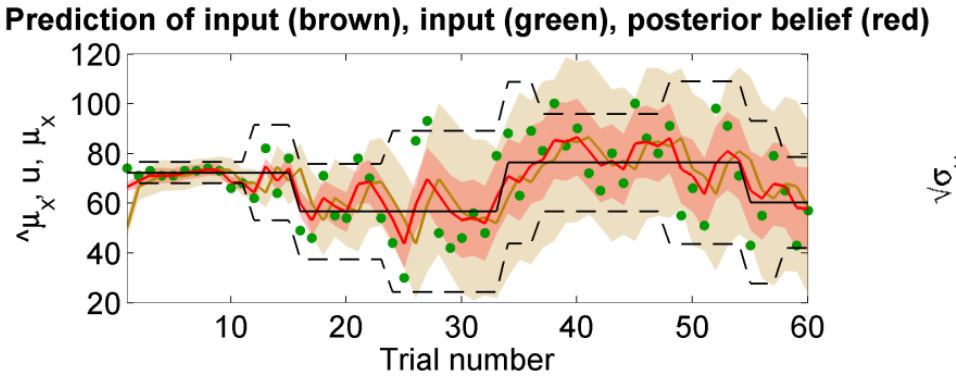
$$\pi_x^{(k)} = \hat{\pi}_x^{(k)} + \hat{\pi}_u^{(k)}$$

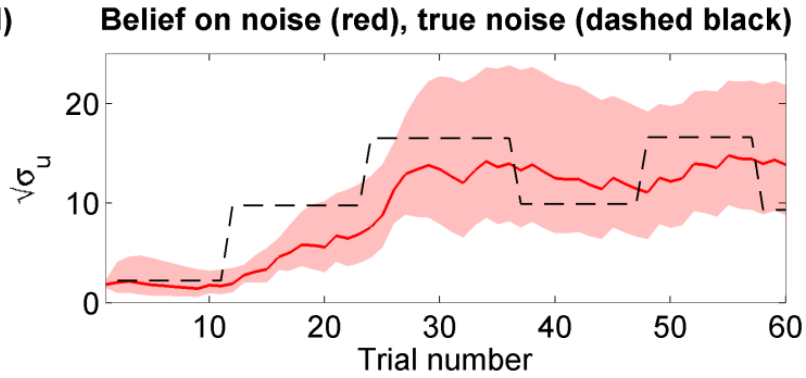
$$\mu_x^{(k)} = \mu_x^{(k-1)} + \frac{\hat{\pi}_u^{(k)}}{\pi_x^{(k)}} \delta_{ux}^{(k)}$$

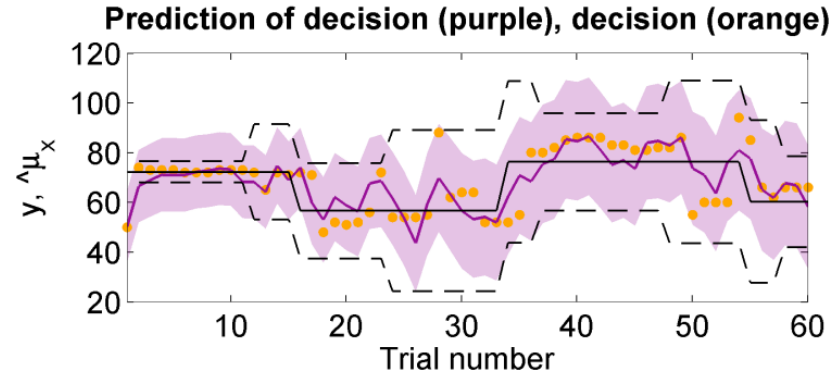
$$\pi_\alpha^{(k)} = \hat{\pi}_\alpha^{(k)} + \frac{\kappa_u^2}{2} \left(1 + \delta_{u\alpha}^{(k)} \right)$$

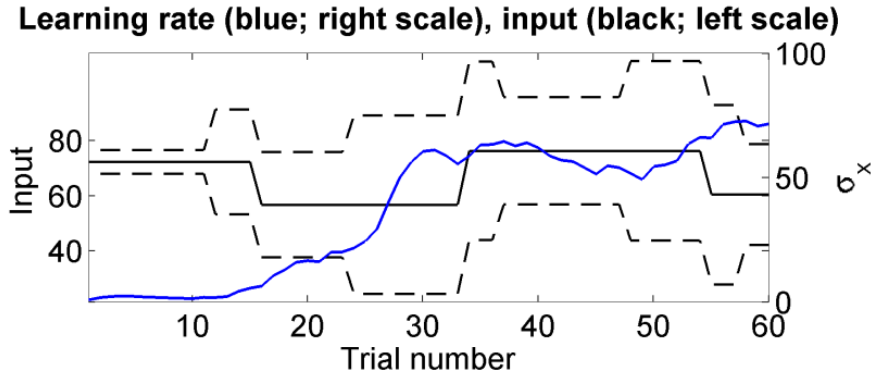
$$\mu_\alpha^{(k)} = \mu_\alpha^{(k-1)} + \frac{\kappa_u}{2} \frac{1}{\pi_\alpha^{(k)}} \delta_{u\alpha}^{(k)}$$

Extensions


Prediction of input (brown), input (green), posterior belief (red)
$\hat{\mu}_x$, u, μ_x
Trial number


Belief on noise (red), true noise (dashed black)
$\sqrt{\sigma_u}$
Trial number


Prediction of decision (purple), decision (orange)
y, $\hat{\mu}_x$
Trial number


Learning rate (blue; right scale), input (black; left scale)
Input
σ_x
Trial number

Summary

- We have developed a **general Bayesian model** for the learning of any changing quantity on the basis of a hierarchy of Gaussian random walks.

- We can derive **one-step updates** that are interpretable, have the structure of **precision-weighted prediction errors**, and can be understood in terms of Rescorla-Wagner learning.

- The resulting model is **modular** and **scalable**, and it can be combined with **many different decision models**.

- The parameters of the learning model **can reliably be estimated** by at least four methods.

- This model **provides a way to explain context effects on learning**, and its three-level version outperforms simplified versions of itself as well as Rescorla-Wagner in all practical applications tested so far.

Thanks

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

University of Zurich[UZH]

UCL

Kay Brodersen

Jean Daunizeau

Andreea Diaconescu

Chaohui Guo

Karl Friston

Sandra Iglesias

Ekaterina Lomakina

Klaas Enno Stephan

Simone Vossel

Lilian Weber