"Statistically optimal perception and learning: from behavior to neural representations."

Fiser, Berkes, Orban & Lengyel

Trends in Cognitive Sciences (2010)

"Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment."

Berkes, Orban, Lengyel, Fiser. *Science* (2011)

Jonathan Pillow

Computational & Theoretical Neuroscience Journal Club

June 22, 2011

Background

General Question: How does the brain represent probability distributions?

(also: *why would it want to?*)

Two basic schemes:
1) Probabilistic Population Coding (PPC) (last week)
2) "Sampling Hypothesis" (today, building on 2 weeks ago)

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues

- Ma et al, *Nat Neuro* 2006

$$r_i | s \sim \text{Poisson}(f_i(s))$$

i'th neuron response

stimulus (1D)

tuning curve

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues

- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response

stimulus (1D)

tuning curve

stimulus prior

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues

- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response

stimulus (1D)

tuning curve

stimulus prior

log TC

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z} \exp\left[\sum_i h_i(s) \cdot r_i\right]$$

posterior

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues

- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z} \exp\left[\sum_i h_i(s) \cdot r_i\right]$$

log TC; posterior

basic idea:

- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z} \exp\left[\sum_i h_i(s) \cdot r_i\right]$$

posterior; log TC

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996
"Sparse Coding Model" / ICA

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z} \exp\left[\sum_i h_i(s) \cdot r_i\right]$$

log TC; posterior

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996 "Sparse Coding Model" / ICA

$$P(r) = \lambda^n \prod_i e^{-\lambda|r_i|}$$

sparse prior over neural responses

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z}\exp\left[\sum_i h_i(s)\cdot r_i\right]$$

log TC; posterior

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996 "Sparse Coding Model" / ICA

$$P(r) = \lambda^n \prod_i e^{-\lambda|r_i|}$$

sparse prior over neural responses

$$P(S|r) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i\cdot r_i\right)^2\right]$$

posterior; stimulus (high-D); "receptive field"

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z}\exp\left[\sum_i h_i(s)\cdot r_i\right]$$

log TC; posterior

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996 "Sparse Coding Model" / ICA

$$P(r) = \lambda^n \prod_i e^{-\lambda|r_i|}$$

sparse prior over neural responses

$$P(S|r) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i \cdot r_i\right)^2\right]$$

posterior; stimulus (high-D); "receptive field"

$$\Longrightarrow \quad P(r|S) \propto P(S|r)P(r)$$

encoding distribution

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z}\exp\left[\sum_i h_i(s)\cdot r_i\right]$$

posterior; log TC

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996

"Sparse Coding Model" / ICA

their idea: neurons aim to represent S with few spikes

$$P(r) = \lambda^n \prod_i e^{-\lambda|r_i|}$$

sparse prior over neural responses

$$P(S|r) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i \cdot r_i\right)^2\right]$$

posterior; stimulus (high-D); "receptive field"

$$r = \arg\max_r P(r|S)$$

(MAP estimation)

$$\Longrightarrow \quad P(r|S) \propto P(S|r)P(r)$$

encoding distribution

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z}\exp\left[\sum_i h_i(s)\cdot r_i\right]$$

log TC; posterior

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996

"Sparse Coding Model" / ICA

their idea: neurons aim to represent S with few spikes

$$P(r) = \lambda^n \prod_i e^{-\lambda|r_i|}$$

sparse prior over neural responses

$$P(S|r) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i \cdot r_i\right)^2\right]$$

posterior; stimulus (high-D); "receptive field"

$$r = \arg\max_r P(r|S)$$

(MAP estimation)
- achieves sparsity
- *doesn't* represent probability

$$\Longrightarrow \quad P(r|S) \propto P(S|r)P(r)$$

encoding distribution

(1) "Probabilistic Population Codes"

Pouget, Beck, Ma, Latham & colleagues
- Ma et al, *Nat Neuro* 2006

$$r_i|s \sim \text{Poisson}(f_i(s)) \qquad P(s) = \text{flat}$$

i'th neuron response; stimulus (1D); tuning curve; stimulus prior

$$\Longrightarrow \quad P(s|r) = \frac{1}{Z}\exp\left[\sum_i h_i(s)\cdot r_i\right]$$

log TC; posterior

basic idea:
- "Poisson-like" neural noise consistent with a scheme for representing distributions by a sum of log-tuning-curves, weighted by neural responses
- makes it easy to do cue combination, readout, etc.

(2) Generative models / "Sampling Hypothesis"

- Olhausen & Field 1996 "Sparse Coding Model" / ICA

their idea: neurons aim to represent S with few spikes

$$P(r) = \lambda^n \prod_i e^{-\lambda|r_i|}$$

sparse prior over neural responses

$$P(S|r) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i \cdot r_i\right)^2\right]$$

posterior; stimulus (high-D); "receptive field"

$$r = \arg\max_r P(r|S)$$

(MAP estimation)
- achieves sparsity
- *doesn't* represent probability

$$\Longrightarrow \quad P(r|S) \propto P(S|r)P(r)$$

encoding distribution

- Berkes, Orban, Lengyel,Fiser (today): neurons represent P(S|r) with samples

$$r \sim P(r|S)$$

What does it mean "*to sample*"?

Example: say we want to represent that we believe the probability of a reward on this trial is 1/3

- PPC scheme: neuron fires 33 spikes (out of max rate of 100)
- "Sampling" scheme: binary neuron responds with 1/3 probability (or, spikes stochastically 1/3 of the time)

What does it mean "*to sample*"?

Example: say we want to represent that we believe the probability of a reward on this trial is 1/3

- PPC scheme: neuron fires 33 spikes (out of max rate of 100)
- "Sampling" scheme: binary neuron responds with 1/3 probability (or, spikes stochastically 1/3 of the time)

Note that response variability has very different interpretations:

- PPC: different spike counts ⇒ different distributions
- Sampling: variability *required* to represent a distribution; variable responses across trials represent the *same* distribution. (but note you need many neurons or multiple trials to represent the distribution accurately!)

Other differences

Propose different semantics of neural responses:

- PPC: a neuron represents "bump" in the posterior distribution
- Sampling: neuron represents presence of a given *feature* in image

- PPC: has nonlinear "tuning curves" (extra layer of abstraction)
- Sparse Coding Model (SCM): projective fields are linearly related to image being represented

- PPC: low-dimensional stimulus (e.g., orientation of a bar)
- SCM: high-dimensional stimulus (e.g., image patch)

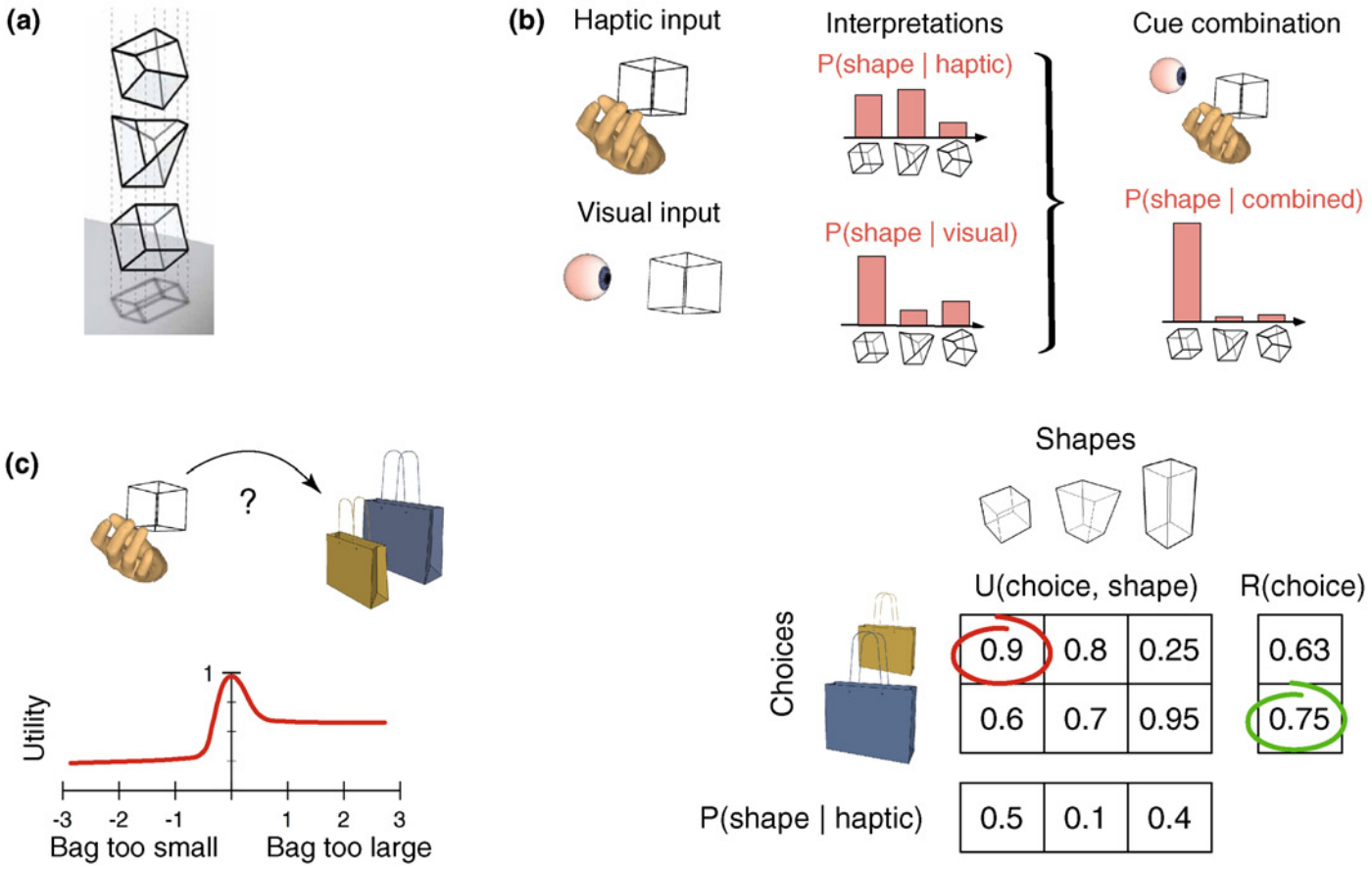
- PPC: responses are *parameters* of the distribution represented
- Sampling: responses are *random variables* drawn from a distribution that is to be represented

Fiser *et al*, TICS 2010

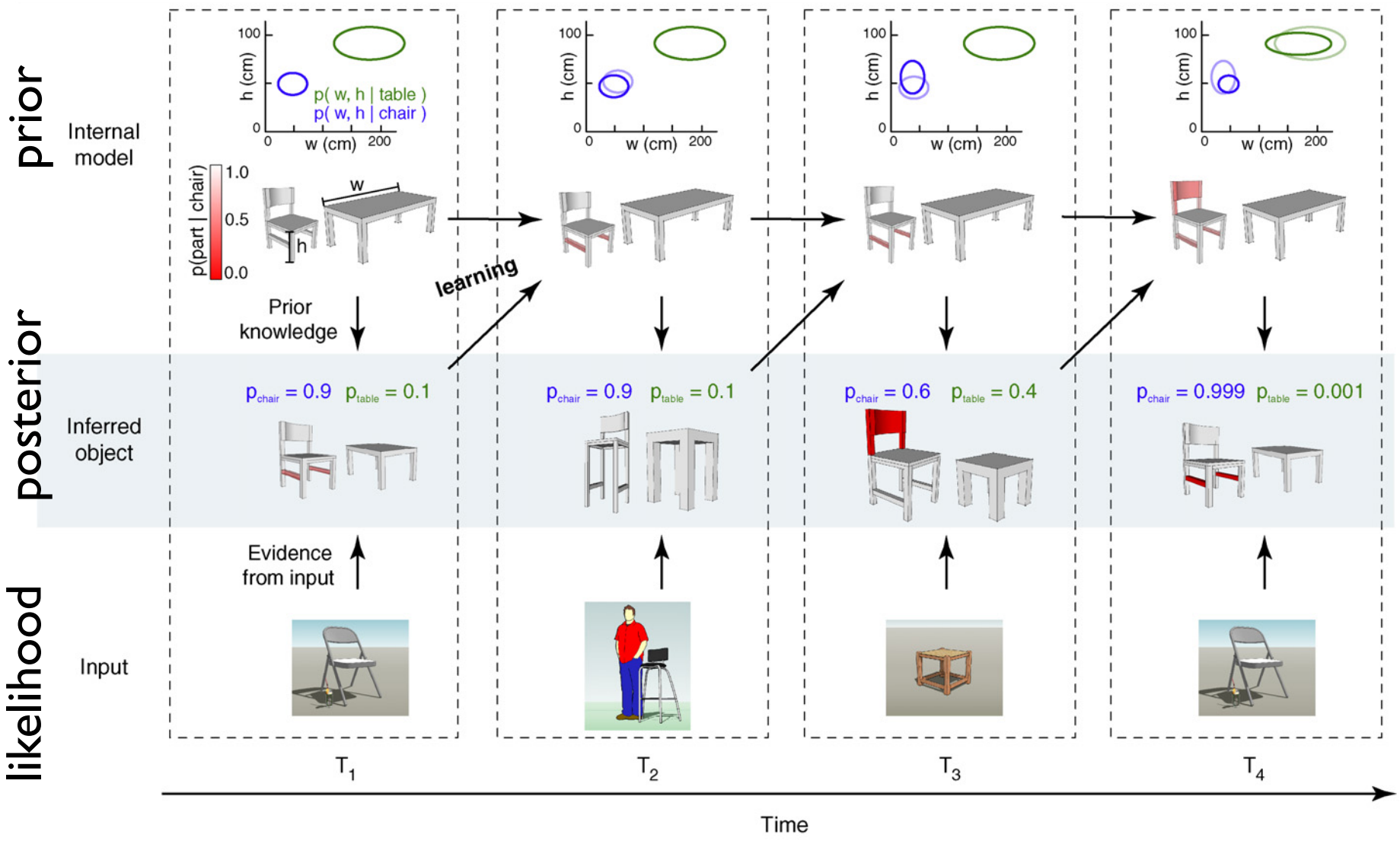
Stated goal:

- why use probabilistic representations
- unify "inference" and "learning"

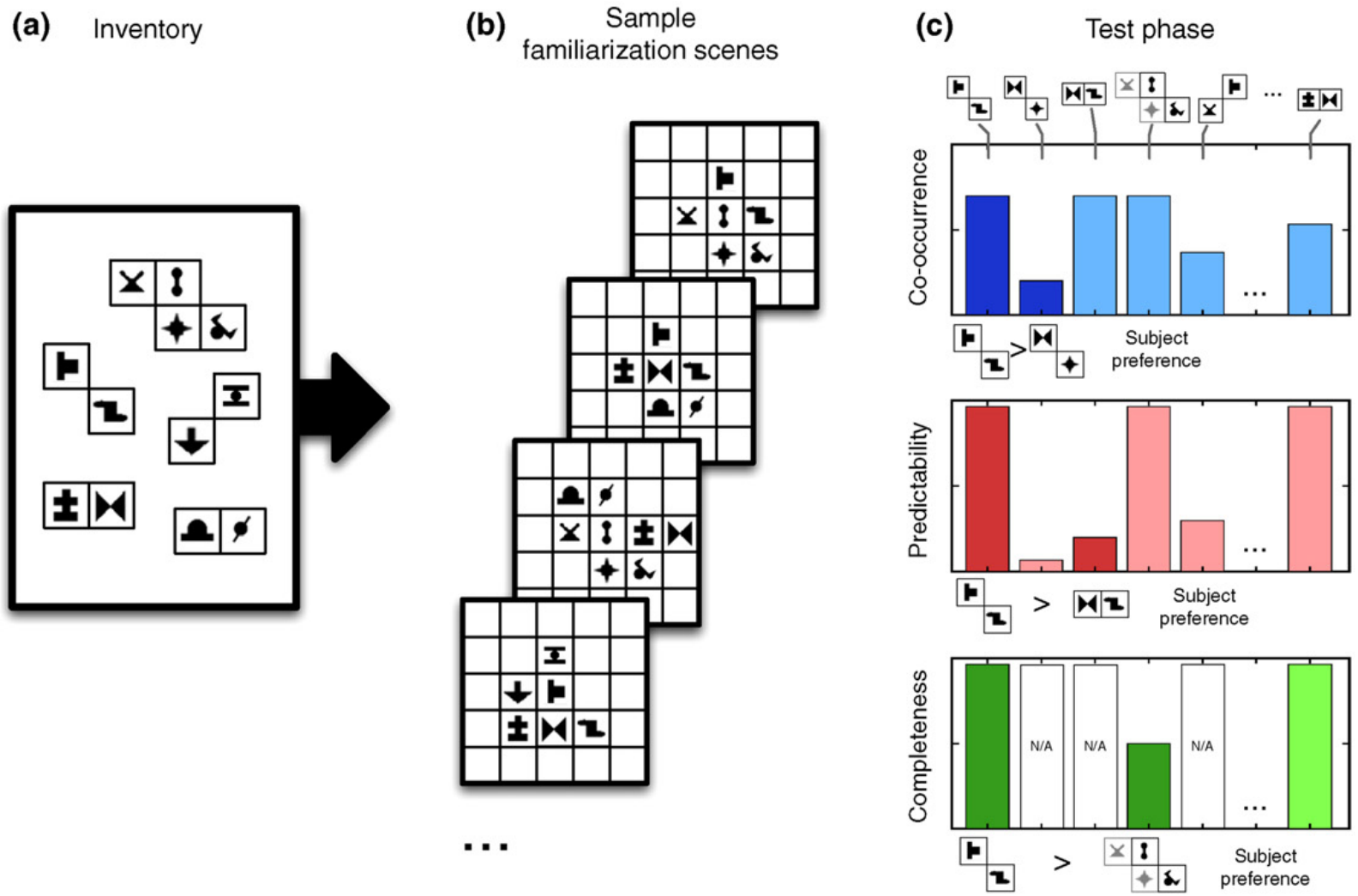
what probabilistic representations are good for



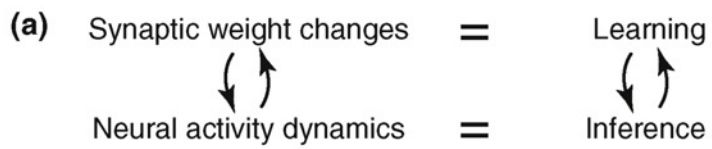
learning in a probabilistic model



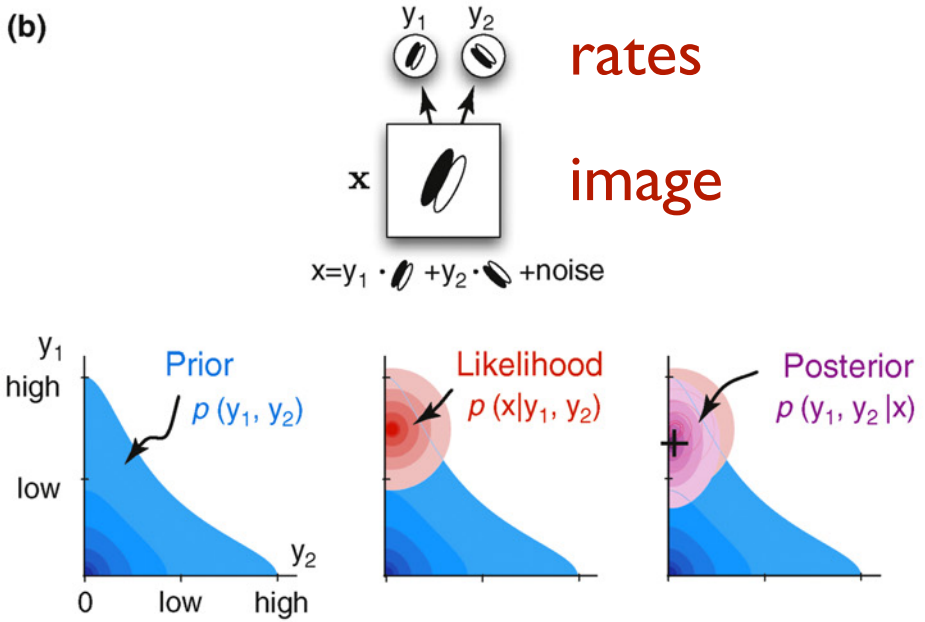
statistical learning in humans



probabilistic inference and learning with neurons



sparse coding model

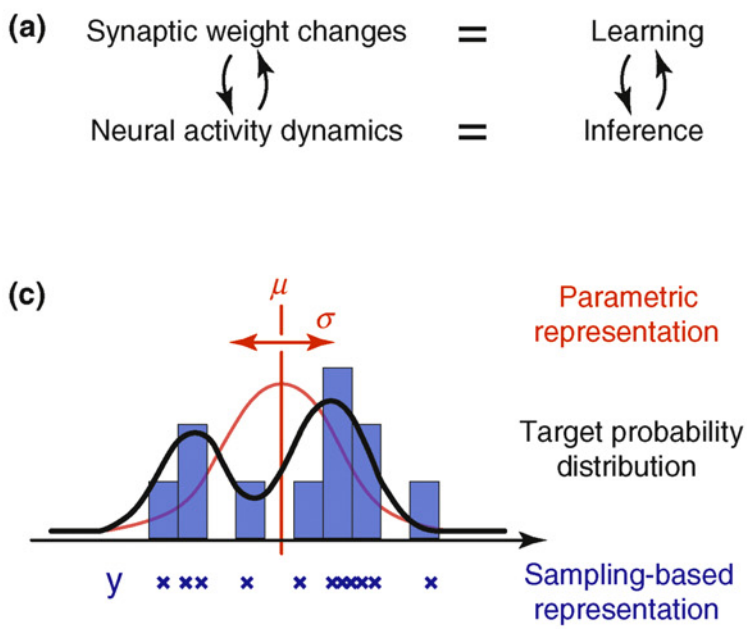


sparse prior $P(r) = \lambda^n \prod_i e^{-\lambda |r_i|}$

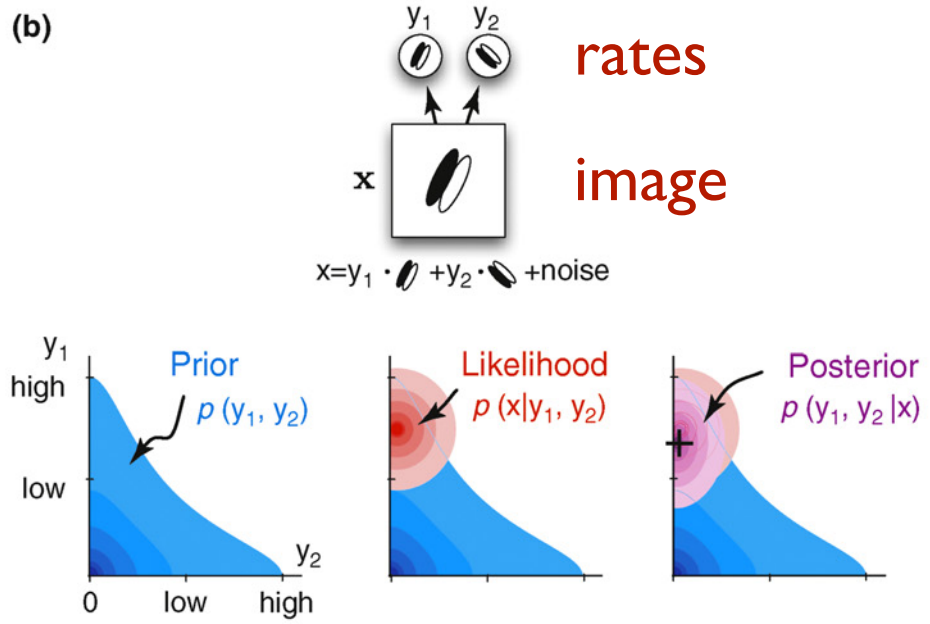
"receptive field"

posterior $P(S|r) = \frac{1}{Z} \exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i \cdot r_i\right)^2\right]$

probabilistic inference and learning with neurons



sparse coding model



sparse prior $P(r) = \lambda^n \prod_i e^{-\lambda |r_i|}$

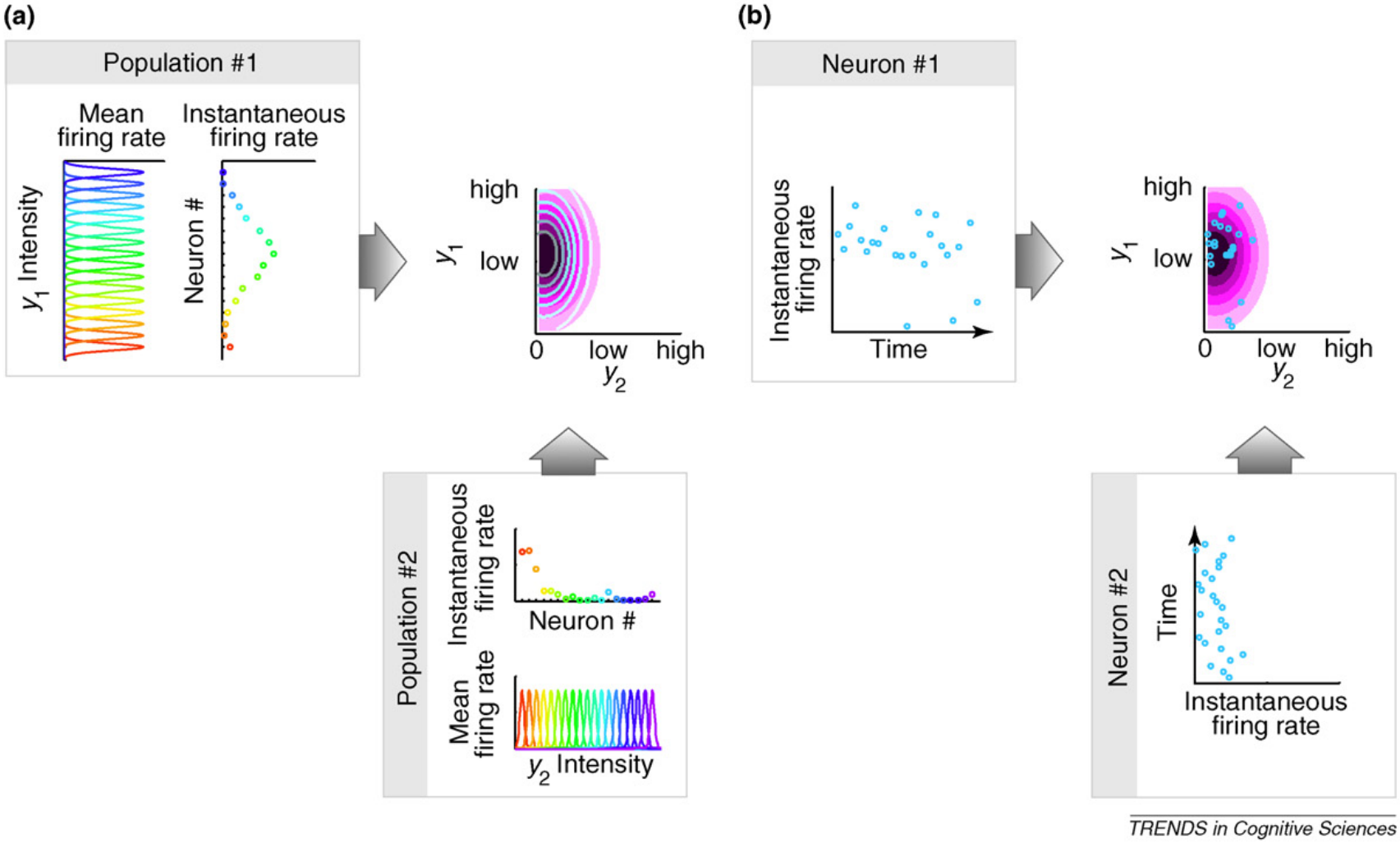
"receptive field"

posterior $P(S|r) = \frac{1}{Z} \exp\left[-\frac{1}{2\sigma^2}\left(S - \sum B_i \cdot r_i\right)^2\right]$

How to represent a 2D distribution

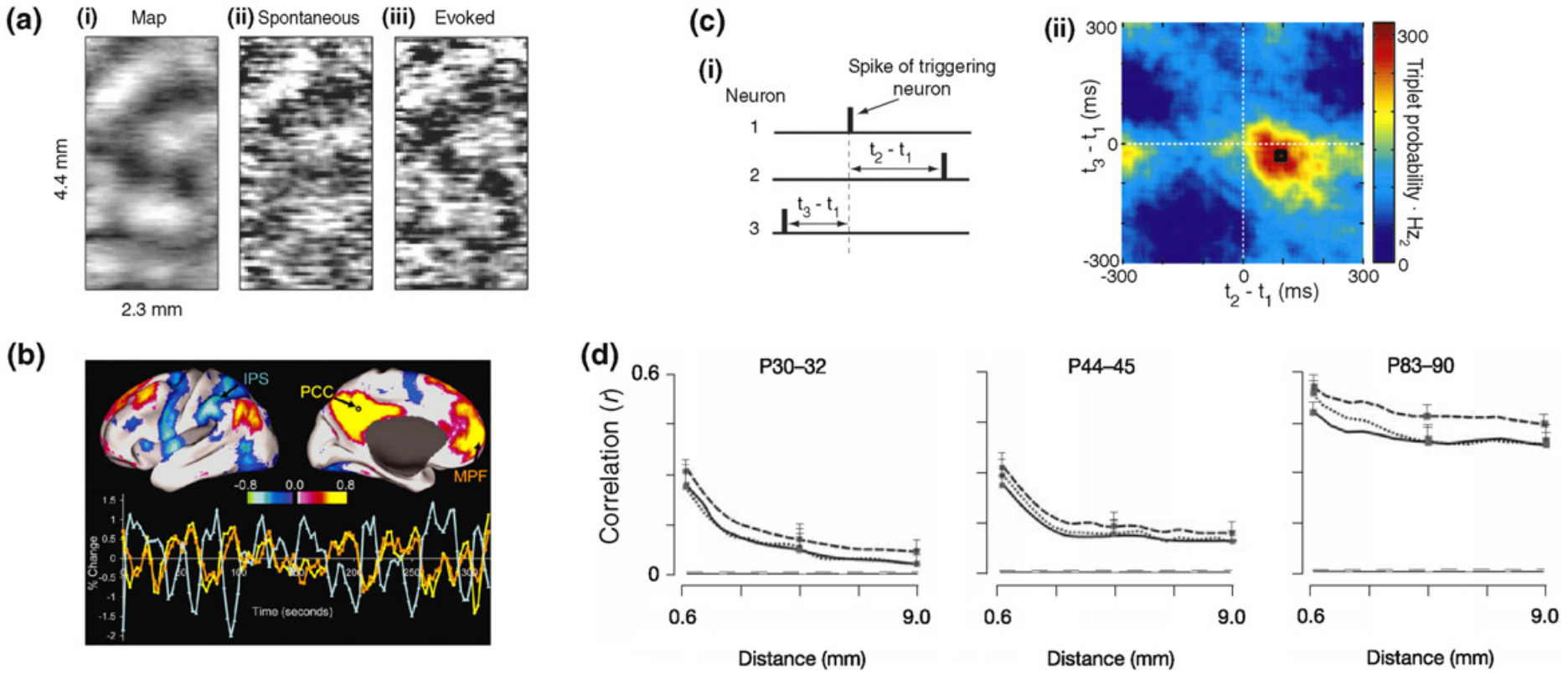
PPC
(2 populations)

Sampling
(2 neurons!)



TRENDS in Cognitive Sciences

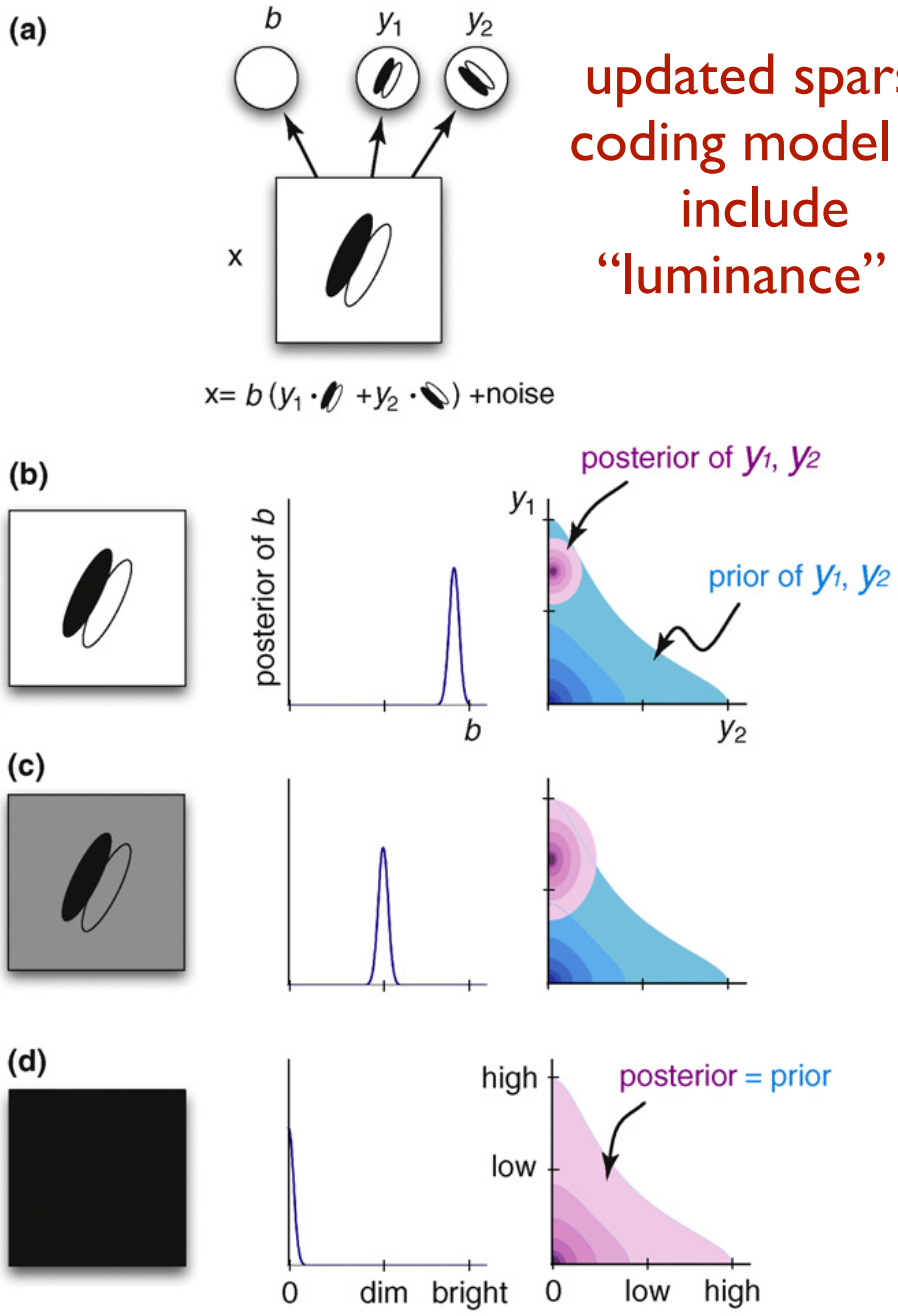
spontaneous activity resembles evoked activity



is structured, correlated

In darkness,
posterior = prior

therefore,
spontaneous
activity = sampling
from the prior



updated sparse coding model to include "luminance" b

comparison of 2 schemes for encoding distributions

Table I. Comparing characteristics of the two main modeling approaches to probabilistic neural representations

	PPCs	Sampling-based
Neurons correspond to	Parameters	Variables
Network dynamics required (beyond the first layer)	Deterministic	Stochastic (self-consistent)
Representable distributions	Must correspond to a particular parametric form	Can be arbitrary
Critical factor in accuracy of encoding a distribution	Number of neurons	Time allowed for sampling
Instantaneous representation of uncertainty	Complete, the whole distribution is represented at any time	Partial, a sequence of samples is required
Number of neurons needed for representing multimodal distributions	Scales exponentially with the number of dimensions	Scales linearly with the number of dimensions
Implementation of learning	Unknown	Well-suited

Berkes *et al*, Science 2011

- empirical test of the sampling hypothesis
- **main result**: spontaneous activity and the average stimulus-evoked activity have the same distribution

Berkes *et al*, Science 2011

- empirical test of the sampling hypothesis
- **main result**: spontaneous activity and the average stimulus-evoked activity have the same distribution

$$P(r) = \int P(r|S)P(S)dS$$

prior (spontaneous activity)

evoked activity ("sampling")

distribution of natural images

average evoked activity

Berkes *et al*, Science 2011

- empirical test of the sampling hypothesis
- **main result**: spontaneous activity and the average stimulus-evoked activity have the same distribution

$$P(r) = \int P(r|S)P(S)dS$$

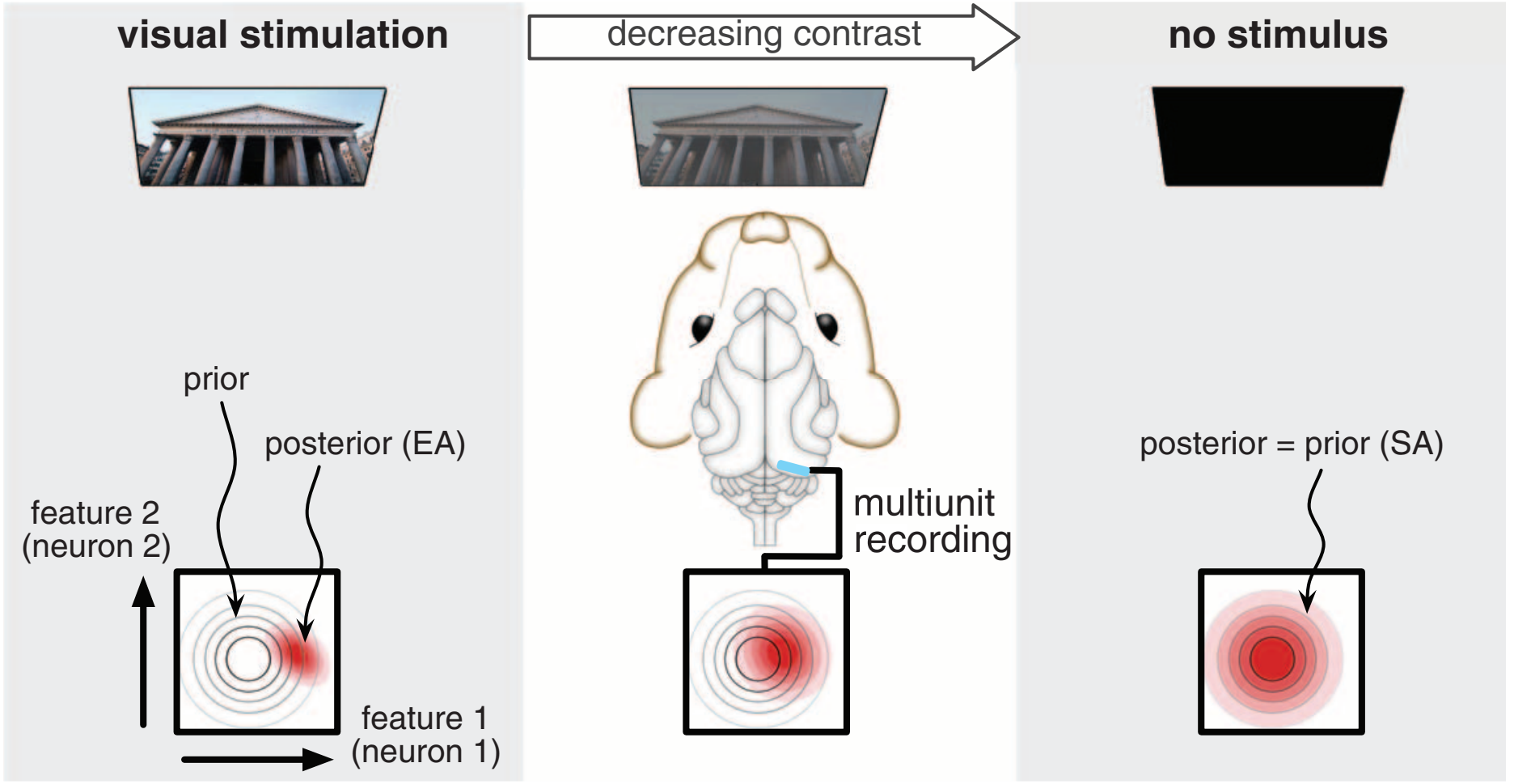
prior (spontaneous activity)

evoked activity ("sampling")

distribution of natural images

average evoked activity

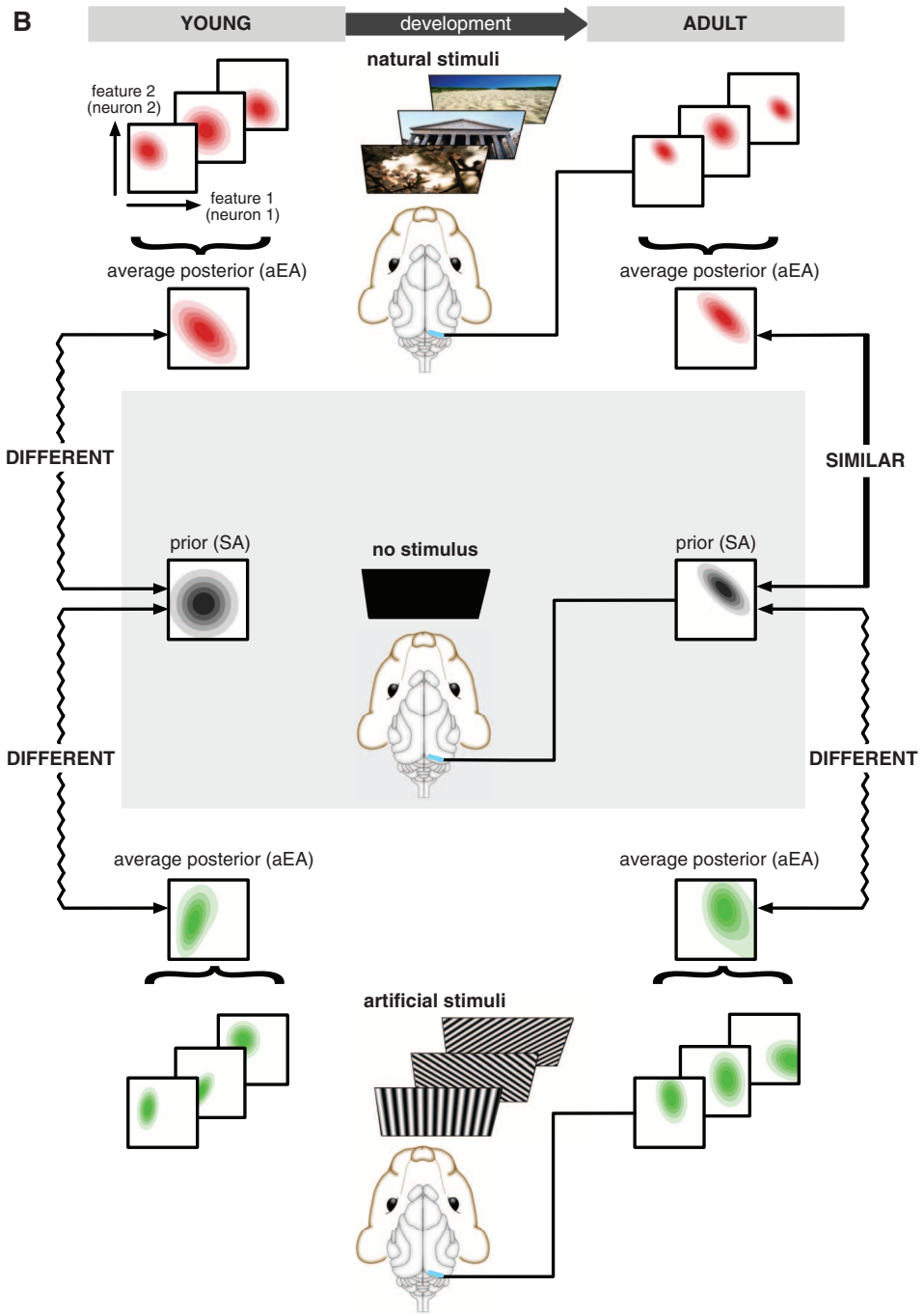
if 2 features co-occur in natural scenes, the spontaneous activity of those neurons should become correlated


visual stimulation
decreasing contrast
no stimulus
prior
posterior (EA)
feature 2
(neuron 2)
feature 1
(neuron 1)
multiunit
recording
posterior = prior (SA)

basic logic:

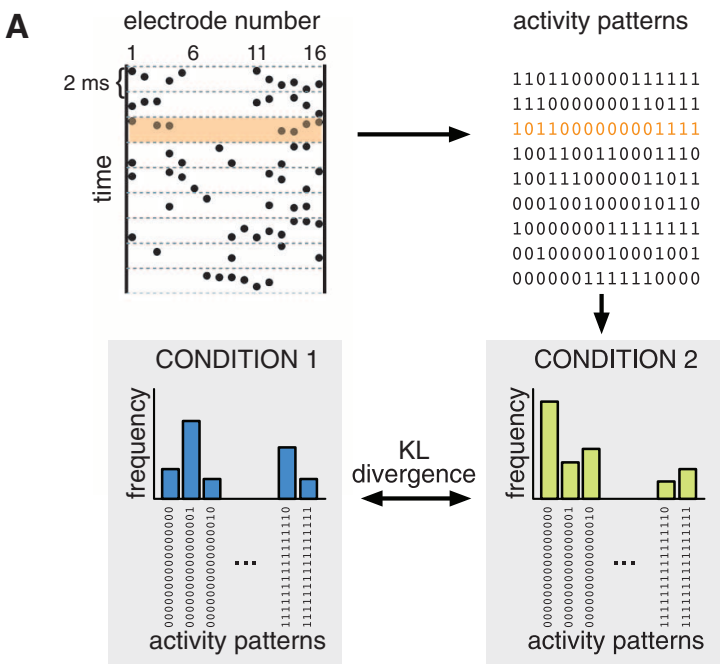
over course of development, prior (spontaneous activity) comes to match average posterior (sampled activity) in response to natural images

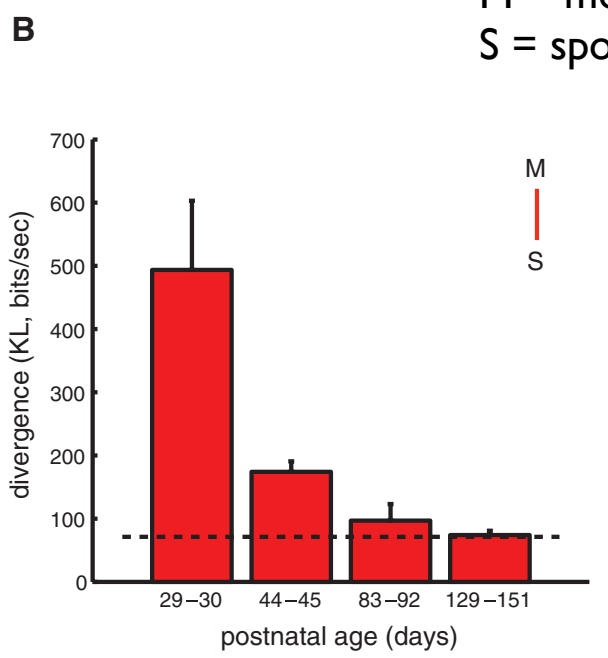
(that is, vis. experience required to learn correct prior)

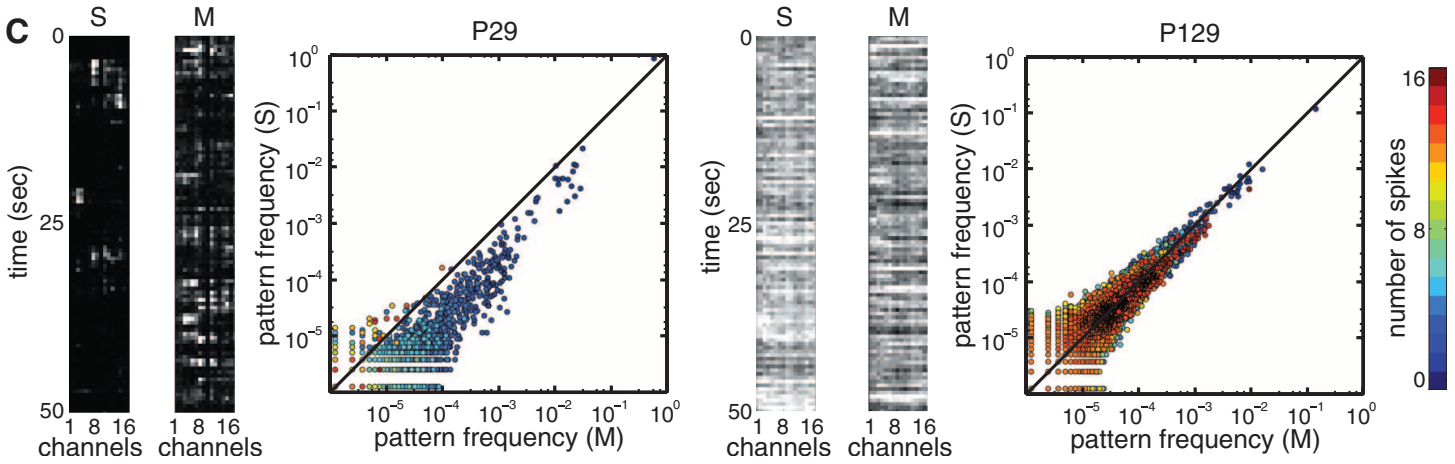


data analysis methods for basic conclusion

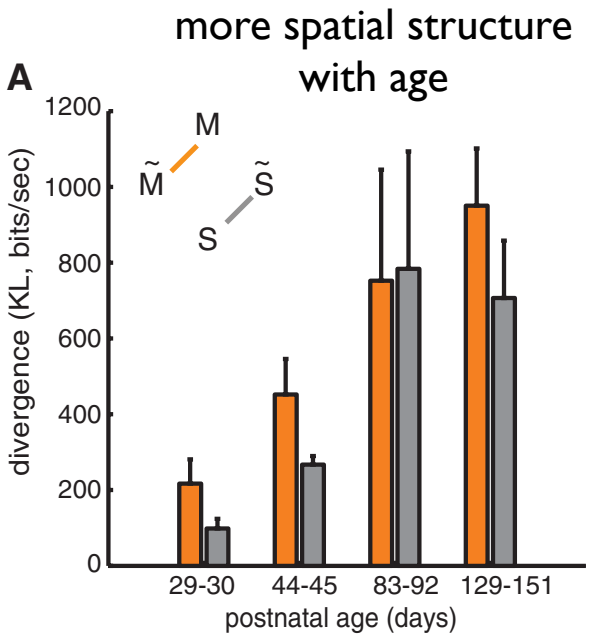
M = movie-evoked
S = spontaneous

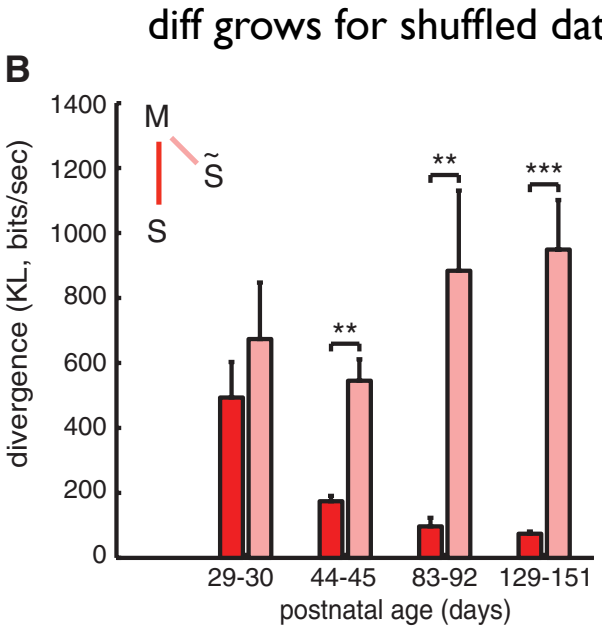




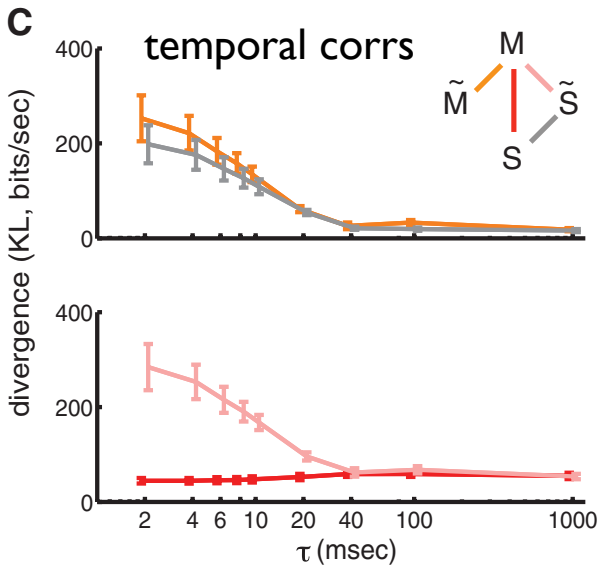


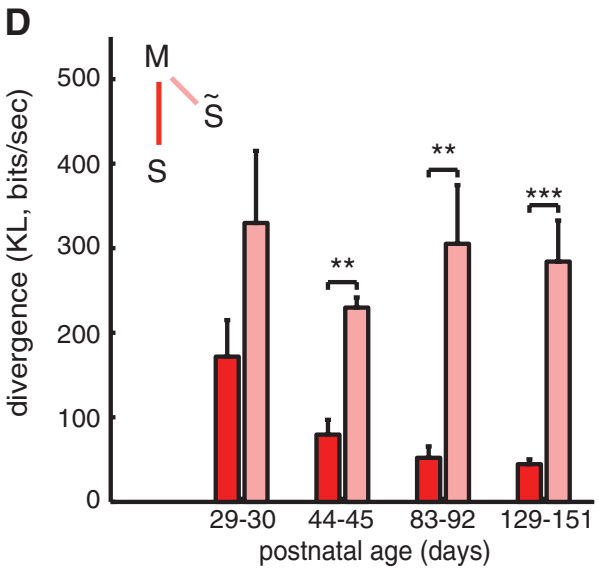
contributions of spatial and temporal correlations



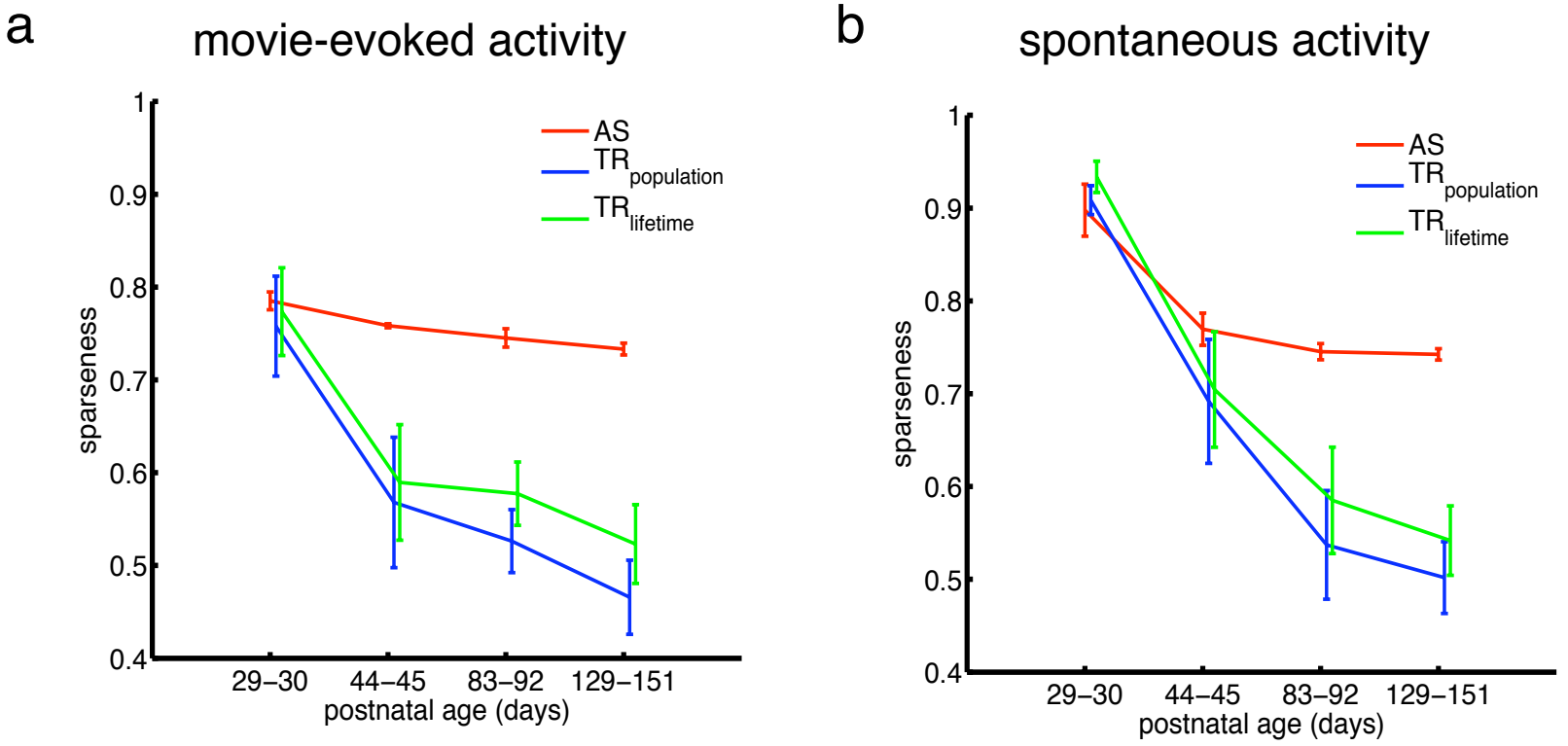


M = movie-evoked
S = spontaneous
~ = shuffled



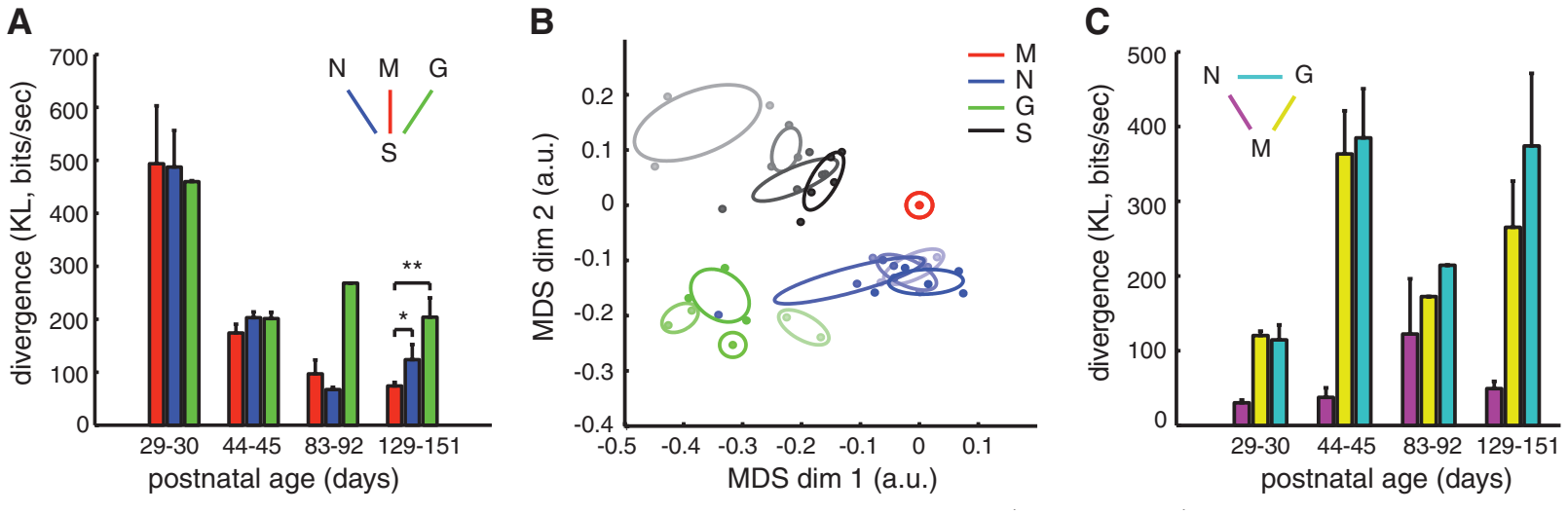


Also, activity becomes *less* sparse during development, arguing against the "redundancy reduction hypothesis" (Barlow)



Finding is specific to Natural Stimuli

M = movie-evoked
S = spontaneous
N = noise-evoked
G = gratings



Conclusions

spontaneous activity	evoked activity: stimulus-dependent $P(\mathbf{r}\|\mathbf{y})$	evoked activity: stimulus-averaged $\int P(\mathbf{r}\|\mathbf{y})P^*(\mathbf{y})\,d\mathbf{y}$
ignored	*"classical" neural coding* (e.g., *Rieke et al., 1997*) information transmission	*Schneidman et al., 2006* error correction
analyzed $P(\mathbf{r})$	*Luczak et al., 2009* coding robustness	*this paper* probabilistic inference

Conclusions

- sampling is great
- interesting, surprising account of "what spontaneous activity means"
- key prediction: spontaneous activity matches average evoked activity
- match gets better during development (learning story: "prior is being tuned to average posterior")

Conclusions

- sampling is great
- interesting, surprising account of "what spontaneous activity means"
- key prediction: spontaneous activity matches average evoked activity
- match gets better during development (learning story: "prior is being tuned to average posterior")

open question

- are there other stories consistent with the fact that the average evoked activity has the same distribution as spontaneous activity?