Structure learning and the growth of knowledge

Sam Gershman
Department of Psychology and Center for Brain Science
Harvard University

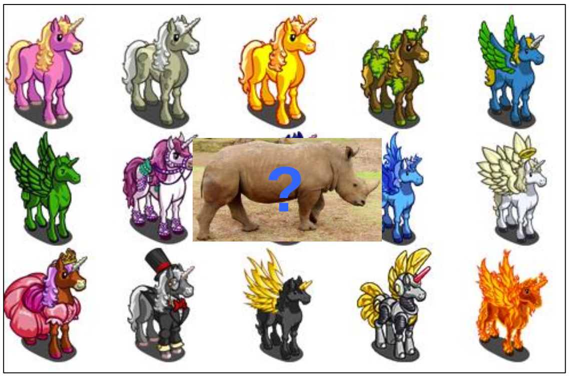
Marco Polo

Marco Polo

Is this a funny-looking unicorn?

Structure and parameters

Parameter learning: what do unicorns tend to look like?



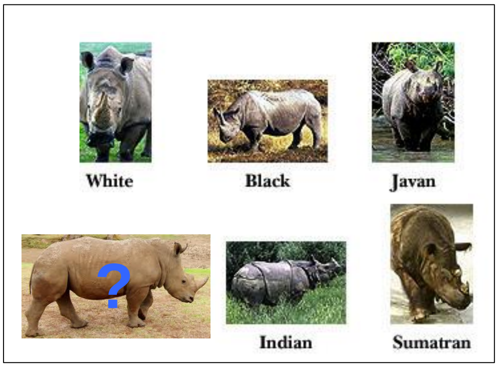
Unicorn space

Structure and parameters

Structure learning: what kinds of animals exist?

Unicorn space

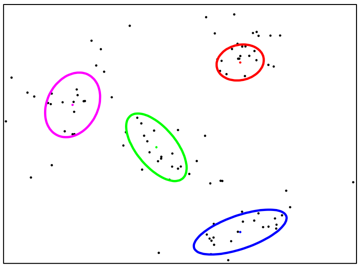
Rhino space



Central problem of structure learning

What's out there?

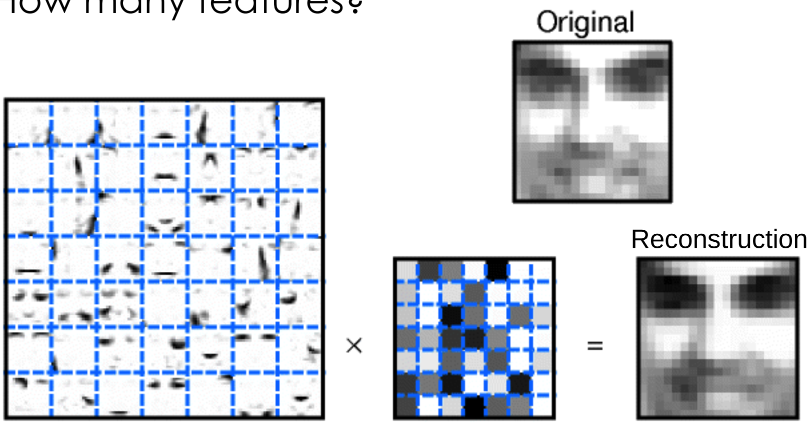
What is structure learning?



How many clusters?

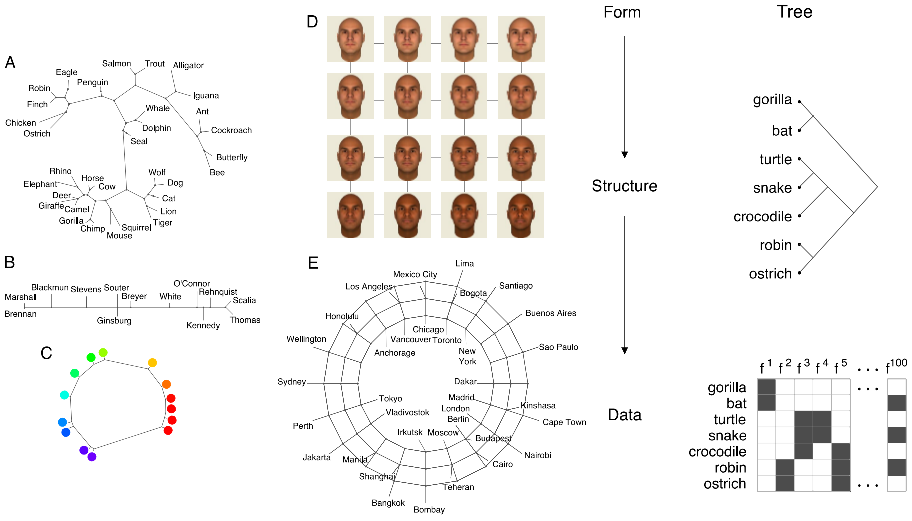
What is structure learning?

How many features?

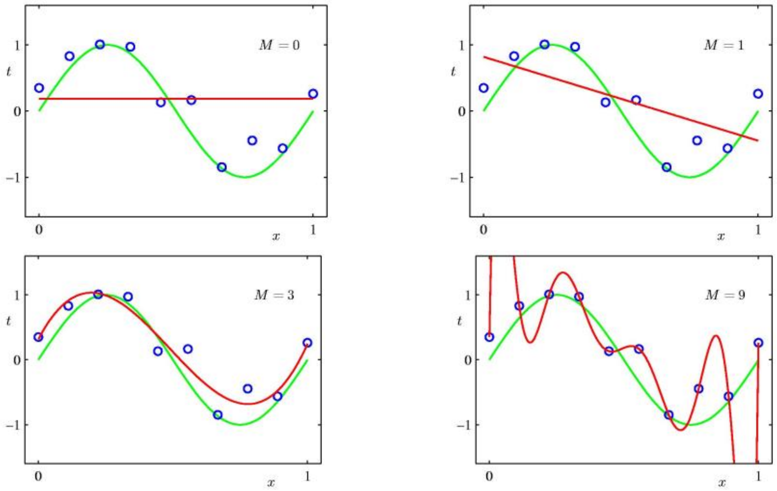


What is structure learning?

Which structural form?



What is structure learning?



Which functional form?

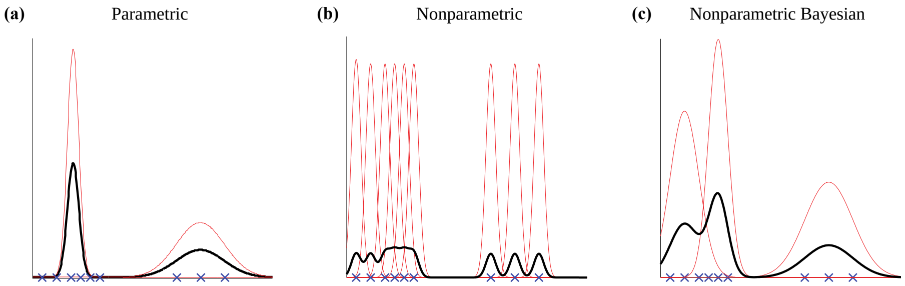
The big picture

- Bayes' rule tells us how to infer hypotheses given data.
- But where do hypotheses come from?
- We can apply the same Bayesian principles to the discovery of hypothesis spaces.

Nonparametric Bayes

- Priors on hypothesis spaces need to be sufficiently rich to accommodate complex data, but must also prefer simpler hypotheses (to avoid overfitting).
- Nonparametric Bayes: priors on "infinite" hypothesis spaces.

What's nonparametric about nonparametric Bayes?

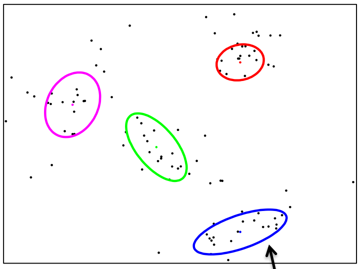


Building blocks

- Mixture models (clustering): Chinese restaurant process
- Latent feature models (factor analysis): Indian buffet process
- Function learning (regression): Gaussian process

PART 1: MIXTURE MODELS AND CLUSTERING

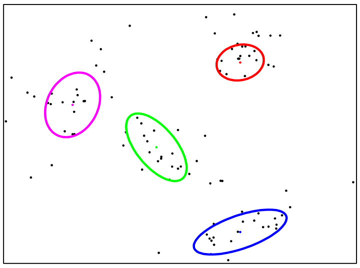
Mixture models



How many clusters?

Each cluster corresponds to a mixture component: a distribution over data

Mixture models



How many clusters?

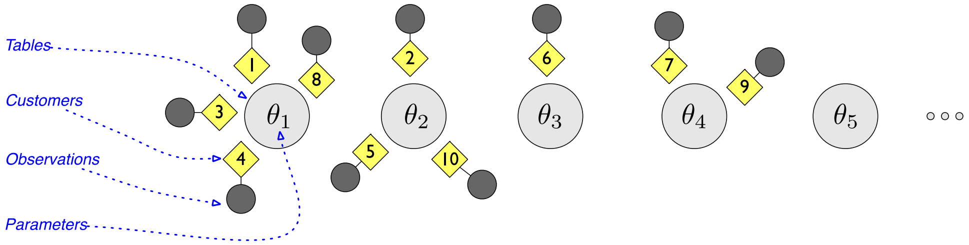
$$P(z|x) \propto P(x|z)P(z)$$

Likelihood of data x given cluster z

Prior probability of cluster z

Chinese restaurant process

Prior over clusters, where the number of clusters is unbounded (formally: distribution on partitions of the integers)



Chinese restaurant process

1. First customer (datapoint) enters and sits at the first table (cluster)
2. Subsequent customers enter and sit at table *k* with probability

$$P(z_t = k) \propto \begin{cases} N_k & \text{if } k \text{ is old} \\ \alpha & \text{if } k \text{ is new} \end{cases}$$

Concentration parameter: larger values produce more clusters

Social structure learning

- Individuals are organized into *latent groups*.
- Beliefs about latent groups determine social influence on decisions.
- Because latent groups are unobservable, people reason about them probabilistically.

Social influence on choice

- Observing the choices of others is a rich source of information about one’s own preferences

Social influence on choice

- Observing the choices of others is a rich source of information about one’s own preferences
  - What movies to see

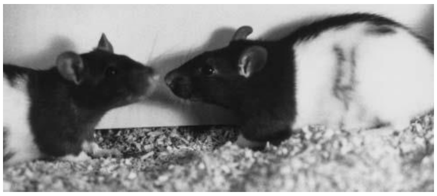
Social influence on choice

- Observing the choices of others is a rich source of information about one's own preferences
  - What movies to see
  - What music to listen to

Social influence on choice

- Observing the choices of others is a rich source of information about one's own preferences
  - What movies to see
  - What music to listen to
  - What food to eat

Social learning in animals



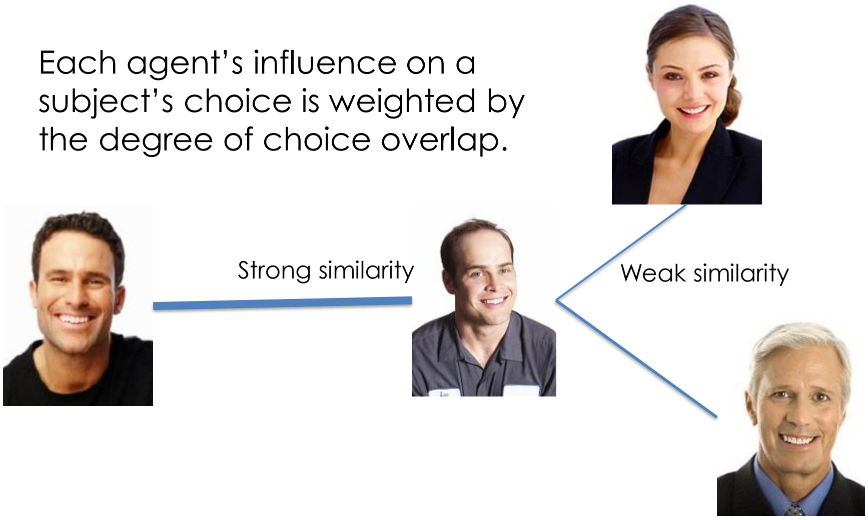
Norway rats sniffing one another's breath to determine what food a conspecific has recently eaten. The rats subsequently show an enhanced preference for that food that lasts for weeks.

The similarity principle

- Social influence is stronger from similar than from dissimilar others.
- Brock (1965) showed that a salesman who reported his own paint consumption to be similar to a customer's sold a larger quantity of paint.

A dyadic similarity model

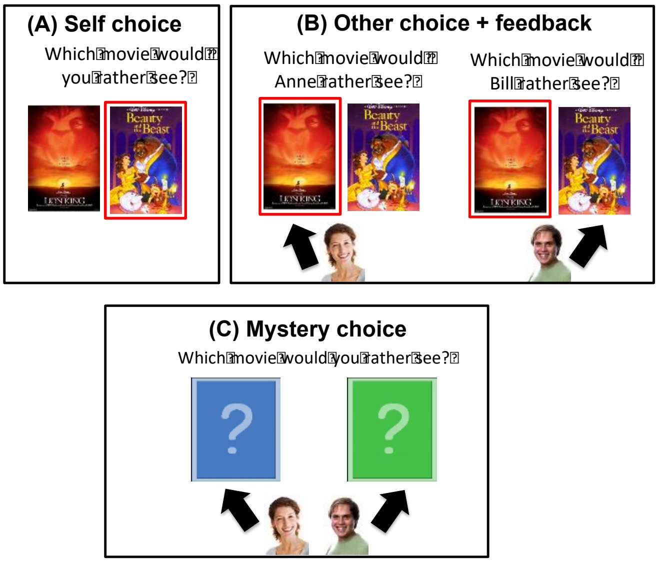
Each agent's influence on a subject's choice is weighted by the degree of choice overlap.



Dyadic similarity vs. latent groups

- I will demonstrate that the dyadic similarity model is too simple to explain social decision making.
- People seem to be guided by inferences about latent groups.

Experimental design



(A) Self choice

Which movie would you rather see?

(B) Other choice + feedback

Which movie would Anne rather see?

Which movie would Bill rather see?

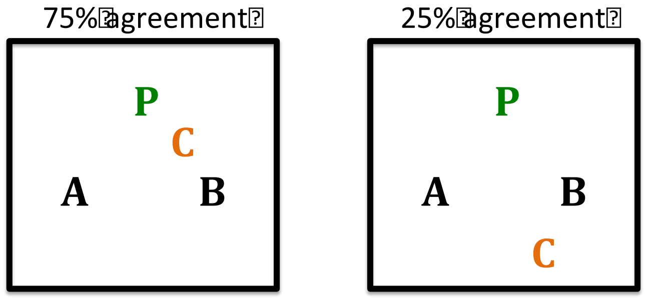
(C) Mystery choice

Which movie would you rather see?

Prediction

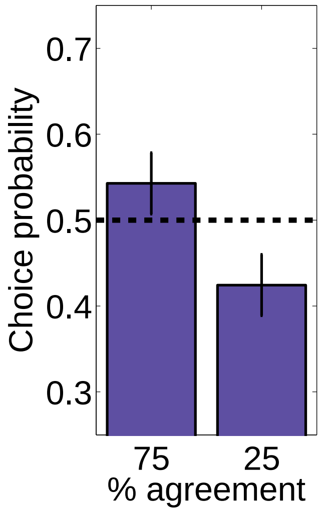
The dyadic similarity model predicts that agents with equal choice overlap should show no differential social influence.

Experimental design



Gershman, Pouncy & Gweon (2017)

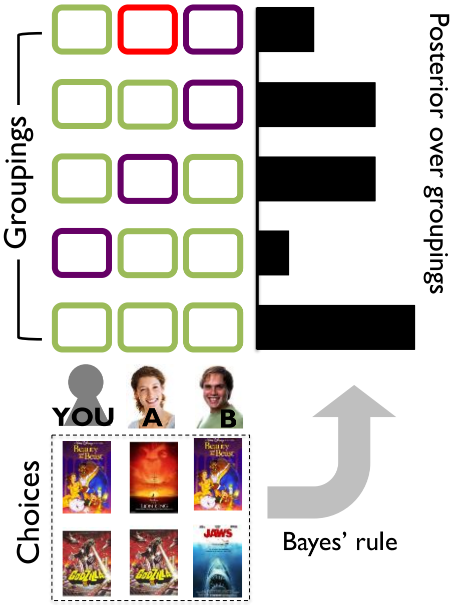
Results



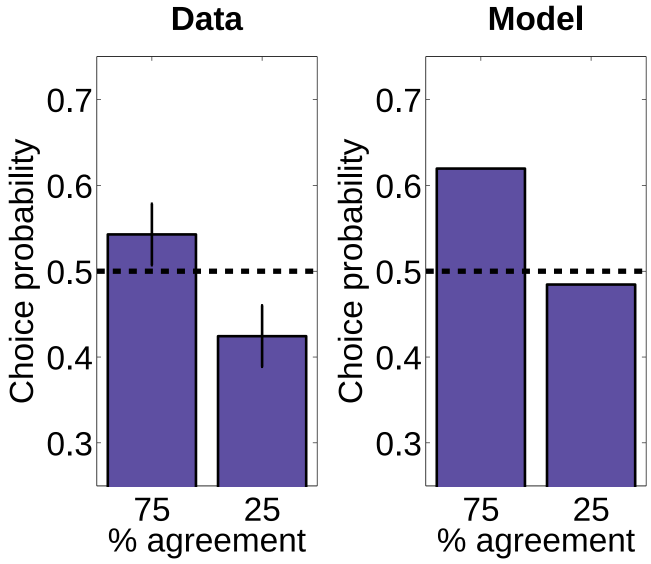
Model schematic

Likelihood: Groupings have high probability when individuals within the same group tend to make the same choices.

Prior: Simpler groupings are preferred over more complex ones.

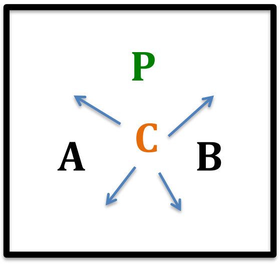


Model simulation

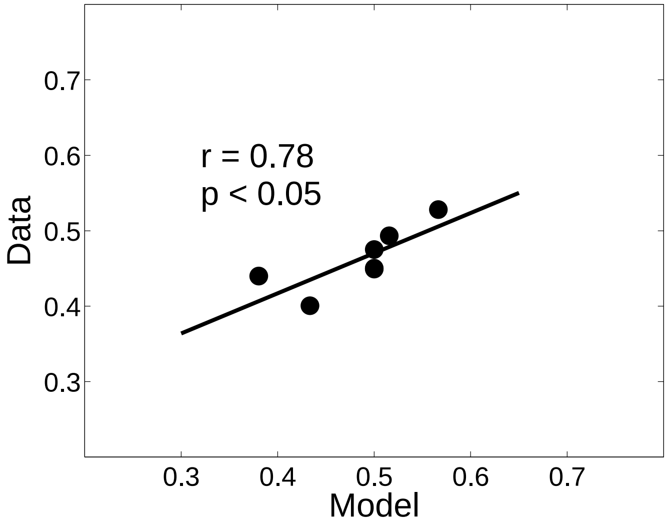


Experimental design

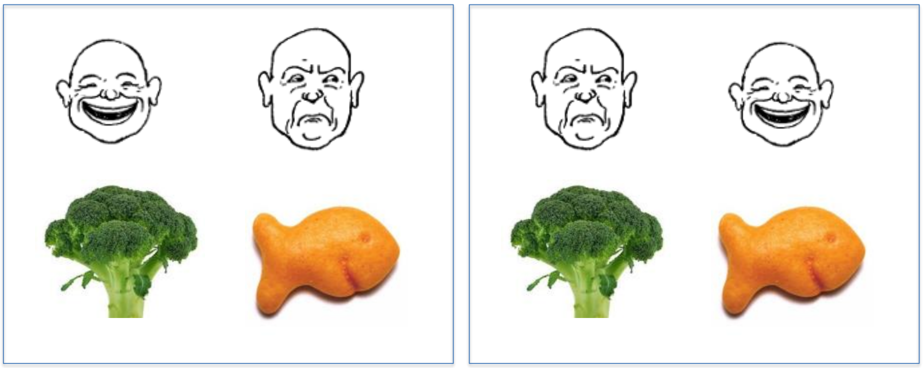
By changing the choice patterns of agent C, we can quantitatively test the model predictions.



Experimental results



The development of social influence

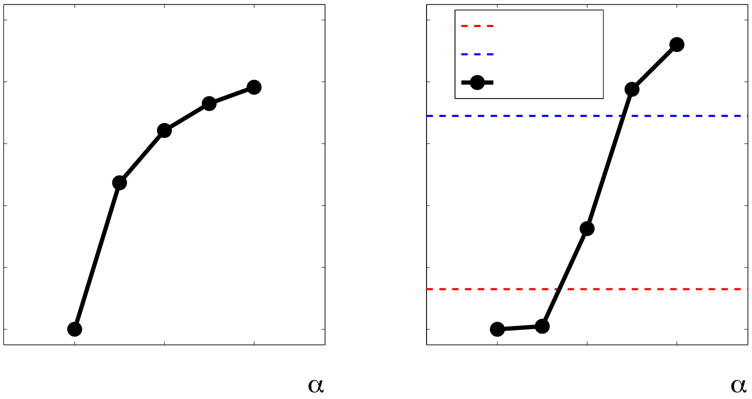


Mismatched

Matched

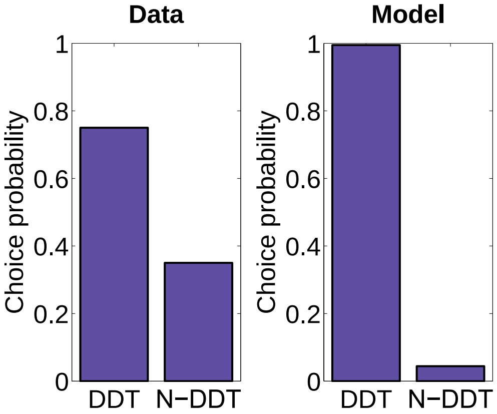
Design of Repacholi & Gopnik (1997)

Simulation



Diverse desires training

Exposing 14-month-olds to individuals with diverse desires gives them social structure model of 18-month-olds



Extensions

- Learning cross-cutting categories: CrossCat
- Clustering relations: the infinite relational model
- Multi-level category learning: the nested CRP

PART 2: LATENT FEATURE MODELS

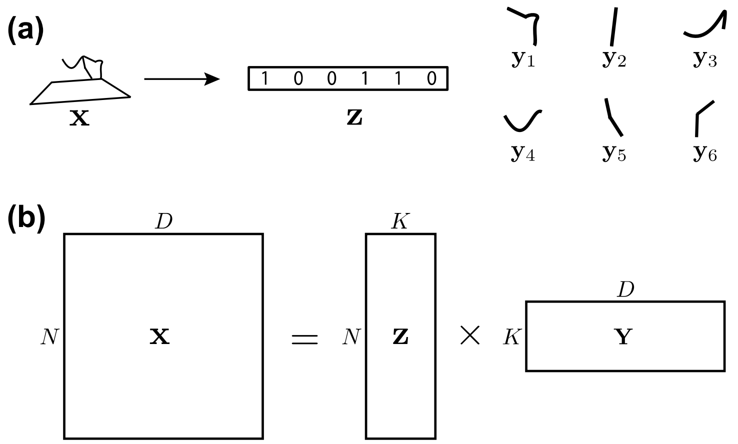
What is a feature?

Many models of human cognition assume objects are represented in terms of abstract features.

What are the features of this object?

What determines which features we identify?

Latent feature models



(Austerweil & Griffiths 2011)

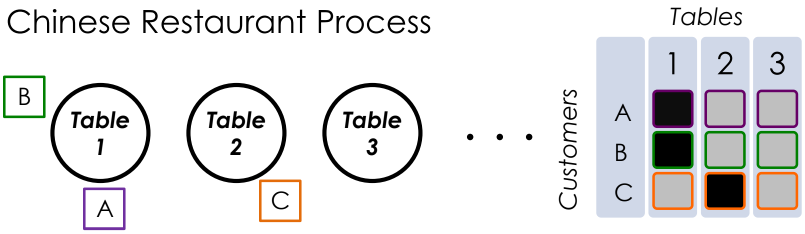
The Indian buffet process

Prior on feature ownership matrices with an unbounded number of features:

1. First customer (datapoint) enters and samples Poisson(a) number of dishes (features)
2. Customer *n* samples dish *k* with probability m_k/n and samples a new dish with probability a/*n*

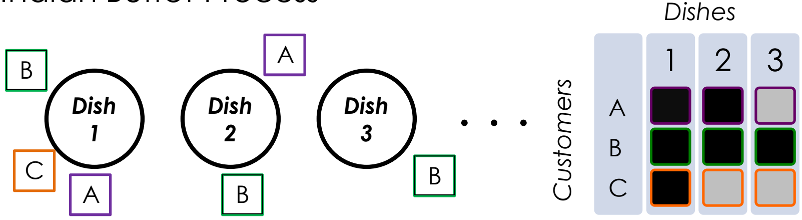
Comparison of CRP and IBP

Chinese Restaurant Process



Customers	1	2	3
A	■	□	□
B	■	□	□
C	□	■	□

Indian Buffet Process



Customers	1	2	3
A	■	■	□
B	■	■	■
C	■	□	□

The problem of summation in classical conditioning

- Elemental theories of conditioning assume that elemental predictions summate: if you like bananas and ice cream separately, you'll like them even more together.
- Example: overexpectation

A+
B+
AB+
B?

Even though AB is reinforced, *less* reinforcement is received than expected, and hence the elemental predictions will be weakened

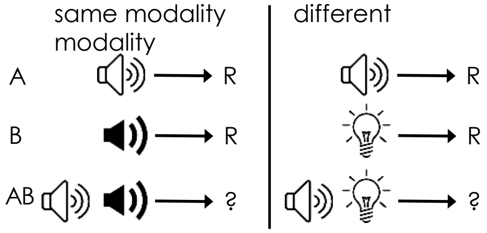
The problem of summation in classical conditioning

- However, sometimes summation does not occur. Instead, the stimulus compound acts *configurally*.
- Example: negative patterning

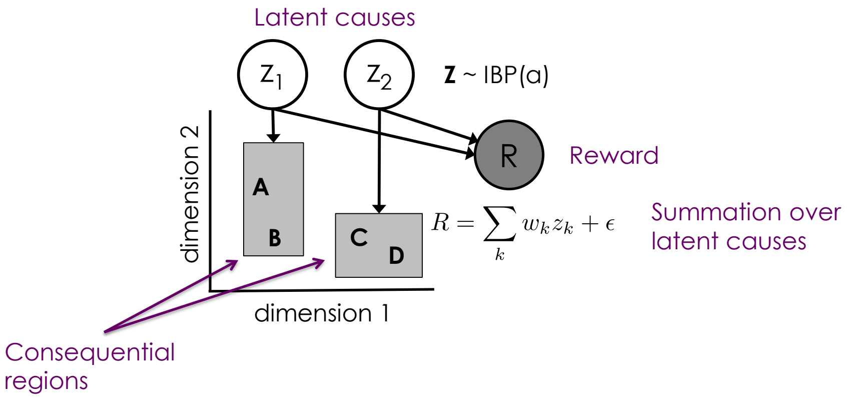
A+
B+
AB-
B?

Animals can learn that the compound predicts no reinforcement even when both elements predict reinforcement.

When does summation occur?



A model with multiple simultaneous latent causes



Latent causes

Z ~ IBP(a)

Reward

$$R = \sum_k w_k z_k + \epsilon$$

Summation over latent causes

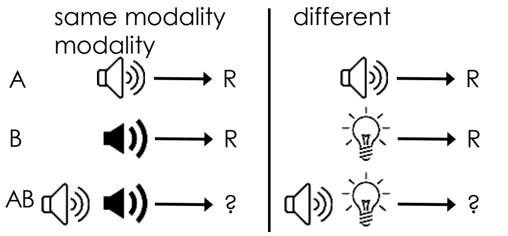
Consequential regions

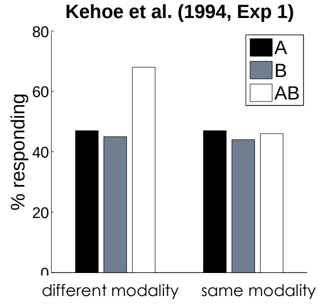
Soto, Gershman, & Niv (2014), *Psych Review*

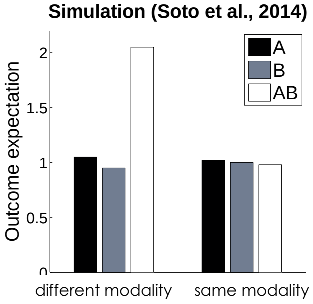
The size principle

- If data are sampled uniformly from a concept’s extension (strong sampling), then concepts with large extensions will receive less evidential support from data.
- This is a form of Bayesian Occam’s razor: concepts that are more “complex” (can predict more patterns of data) place less probability mass on any particular pattern and hence are disfavored relative to concepts that are “simpler” (predict *only* the observed patterns).

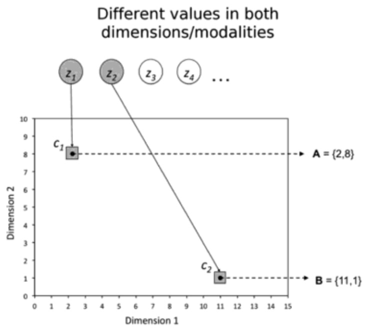
When does summation occur?

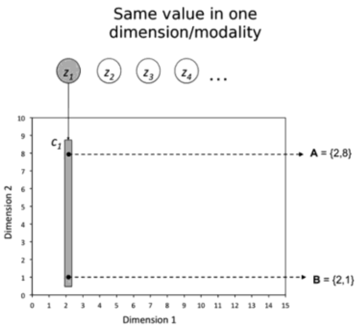






Explaining summation

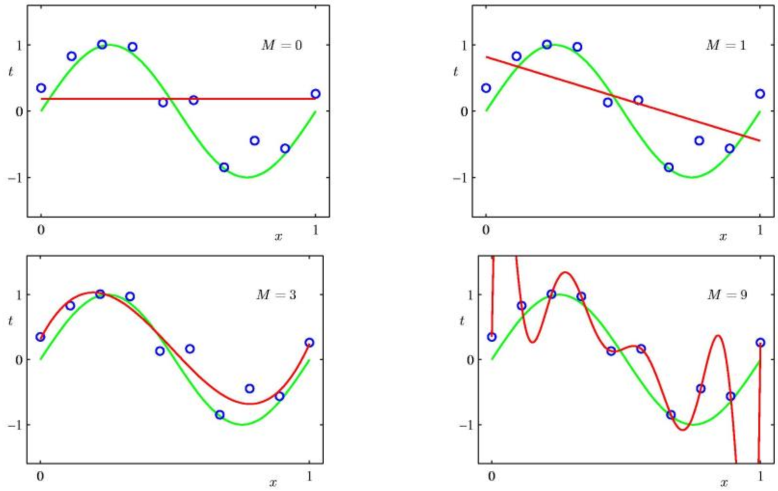




Soto, Gershman, & Niv (2014), *Psych Review*

PART 3: FUNCTION LEARNING

What function generated these data?



Gaussian processes

$$y = f(x) + \epsilon, \quad f \sim \mathrm{GP}(m, k)$$

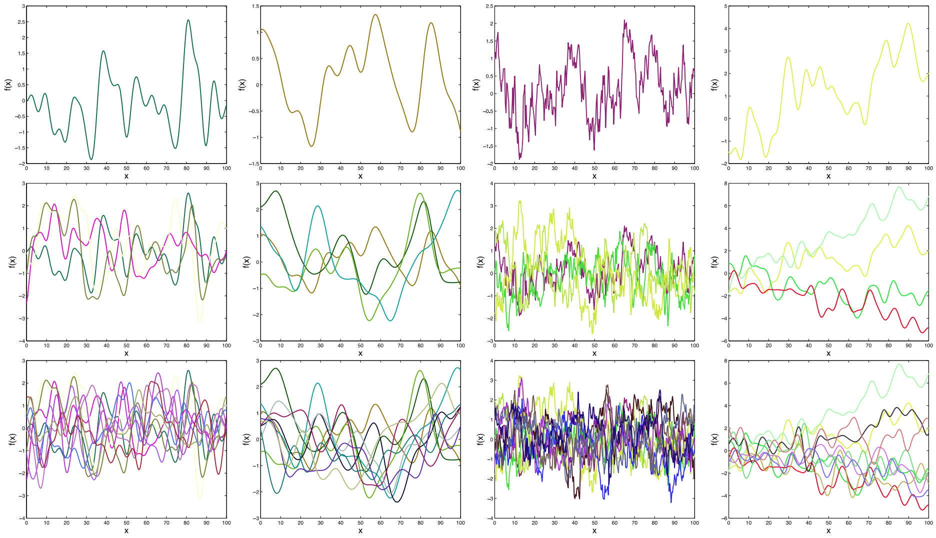
GPs can be thought of a distributions over functions

- m(x) is the mean function
- k(x,x’) is the covariance function (kernel)

The kernel specifies the smoothness of the function

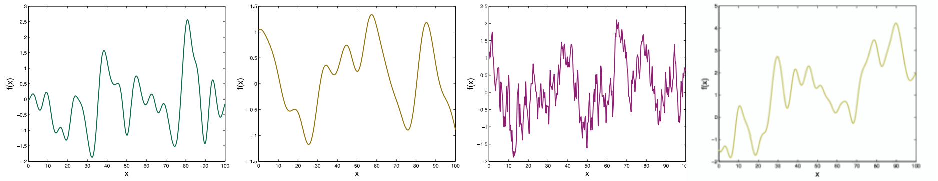
Given data, posterior predictions of function values at arbitrary inputs are computable in closed-form

Samples from GPs with different kernels

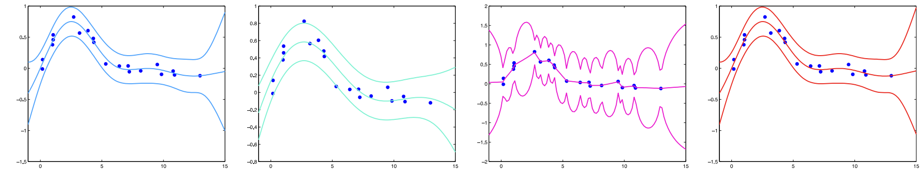


Modeling functions with GPs

A sample from the prior for each covariance function:

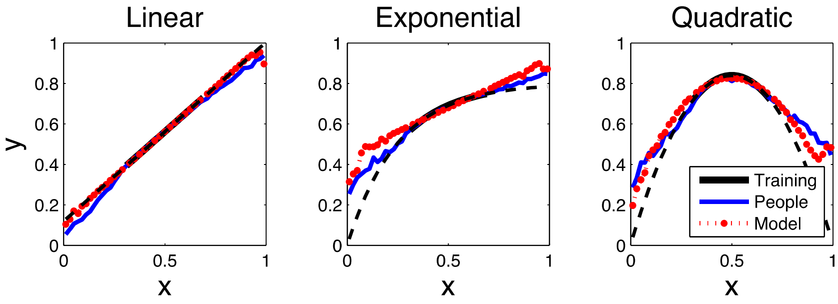


Corresponding predictions, mean with two standard deviations:



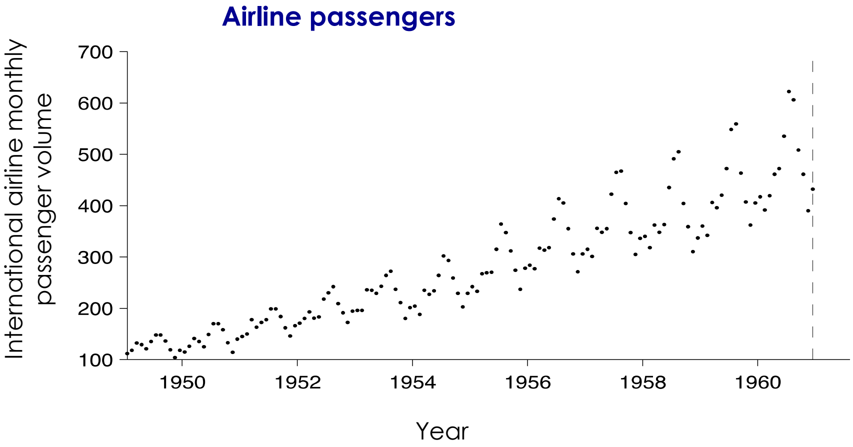
We can use Bayesian model selection to choose the optimal covariance function (and its parameters)

Human function learning



Lucas et al. (2015)

Structure and compositionality

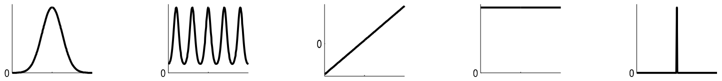


Compositional functions

- To capture compositionality of functions, we need a grammar consisting of:
  - Functional atoms (base kernels)
  - Compositional operators (maps from sets of functions to new functions)
- Note that we don’t specify the functions themselves—only priors on functions (GPs).

Functional atoms

Five base kernels



Squared exp. (SE)

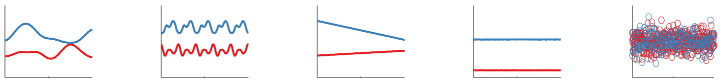
Periodic (PER)

Linear (LIN)

Constant (C)

White noise (WN)

Encoding for the following types of functions



Smooth functions

Periodic functions

Linear functions

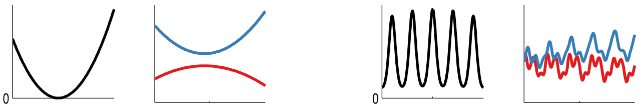
Constant functions

Gaussian noise

(Lloyd, Duvenaud, Tenenbaum, Ghahramani)

Compositional operators

Multiplication



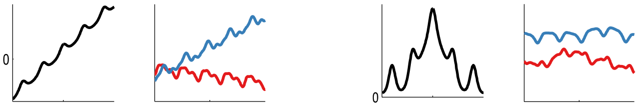
LIN × LIN

quadratic functions

SE × PER

locally periodic

Addition



LIN + PER

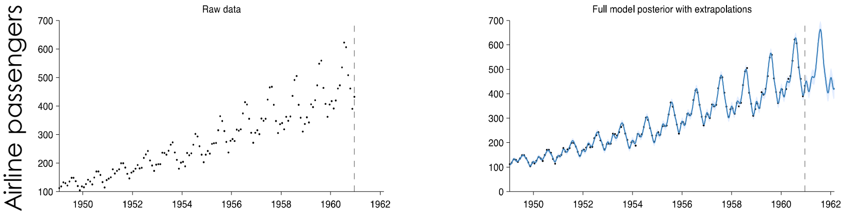
periodic plus linear trend

SE + PER

periodic plus smooth trend

(Lloyd, Duvenaud, Tenenbaum, Ghahramani)

Illustration



Four additive components have been identified in the data

- A linearly increasing function.
- An approximately periodic function with a period of 1.0 years and with linearly increasing amplitude.
- A smooth function.
- Uncorrelated noise with linearly increasing standard deviation.

(Lloyd, Duvenaud, Tenenbaum, Ghahramani)

An alternative: spectral mixture

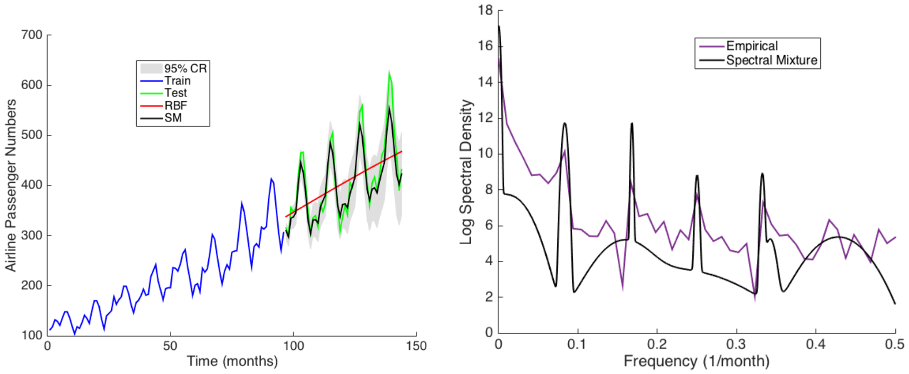
Fourier transform of a stationary kernel (only depends on x-x') yields a spectral representation:

$$k(x, x') = \int_s S(s) e^{2\pi i s^\top (x - x')} ds$$

Roughly speaking, the spectral density S(s) specifies the contribution of the eigenfunction with frequency *s*.

We can define flexible kernels by directly parameterizing the spectral density.

An alternative: spectral mixture



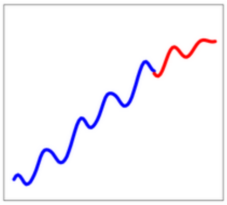
Derive kernels by approximating the spectral density with a mixture of Gaussians.
This function is smooth and flexible but non-compositional.

(Wilson & Adams, 2013)

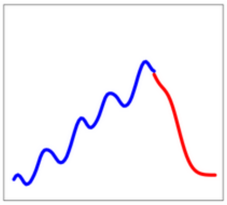
Extrapolation experiment

Choose a pattern completion

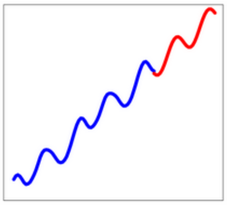
Number of trials left: 20



Spectral mixture



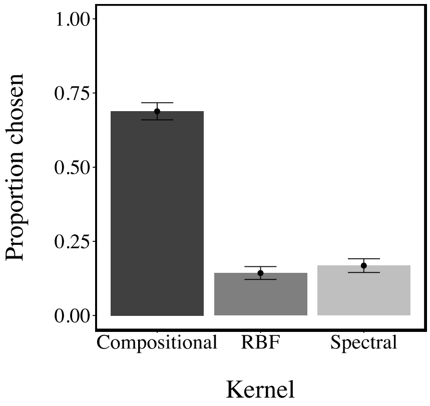
RBF



Compositional

Functions were drawn from the compositional grammar

Results



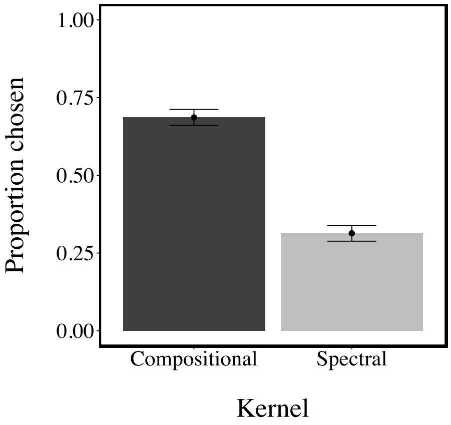
Compositional extrapolations are preferred to non-compositional extrapolations.

Schulz et al. (2017)

Pattern completion (2)

- Same as first experiment, but now functions are sampled from the spectral mixture kernel.

Results



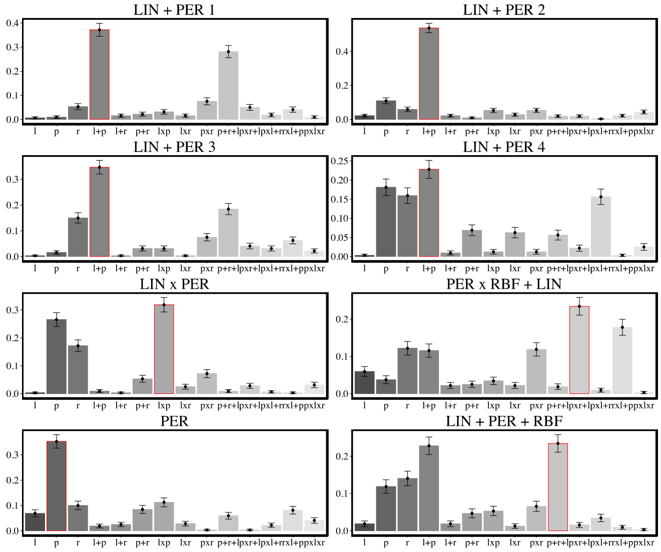
Compositional functions are favored even when the ground truth is non-compositional

Markov chain Monte Carlo with people

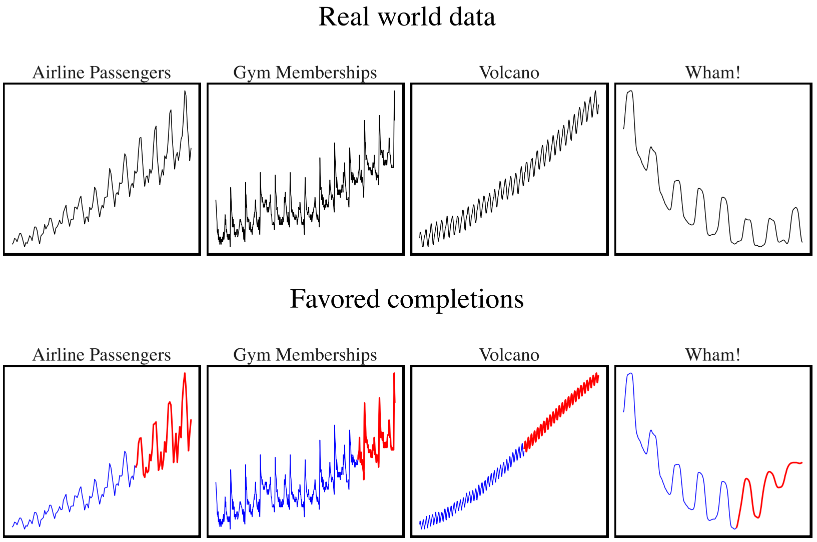
- Generate samples from subjects' posterior by having them simulate a Markov chain.
- Provides a richer picture of their inductive bias.

Results

Posterior distributions over functions favor compositional structures.



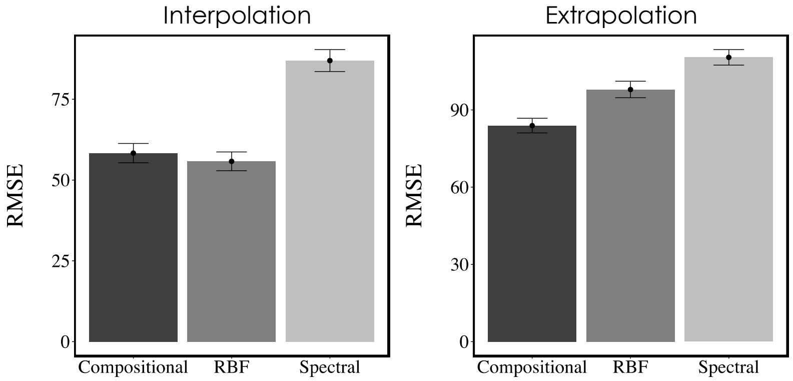
Real-world functions



Manual pattern completion

- Instead of discrete choices, subjects completed the function manually.
- We used the root mean squared error (RMSE) from each kernel's predictions as an index of that kernel's fit.

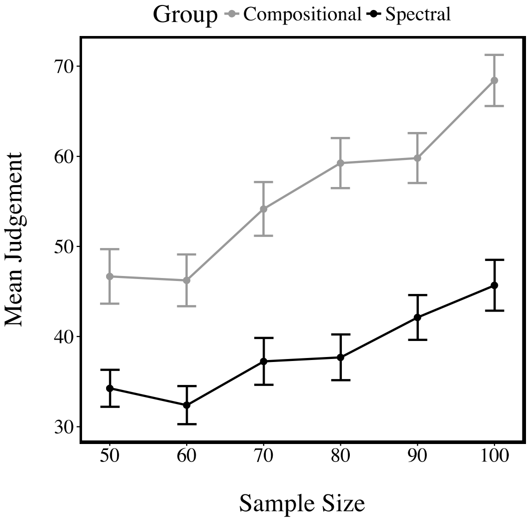
Manual pattern completion



Predictability

- Do people find compositional functions more predictable?

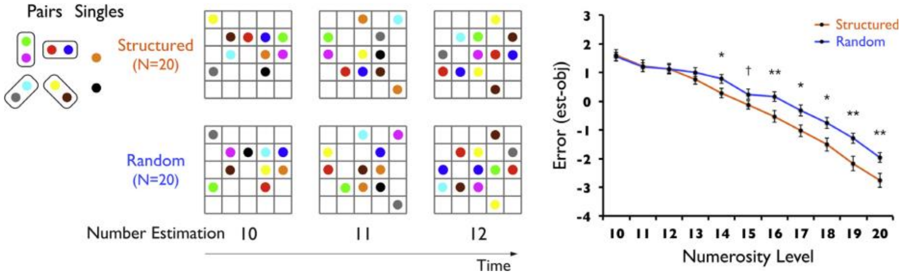
Predictability results



Beyond function learning

- We next explored the implications of compositional functions for several other domains:
  - Numerosity perception
  - Change detection
  - Short-term memory

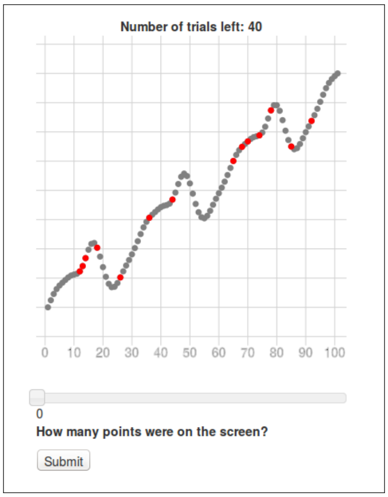
Statistical regularities reduce numerosity estimates



In structured displays, certain color pairs co-occurred, whereas in random displays the co-occurrence statistics were uniform.

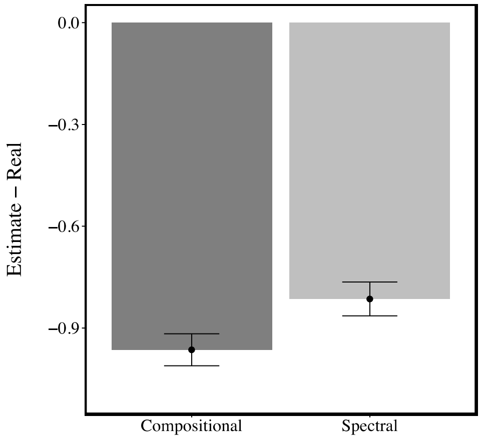
(Zhao & Yu, 2016)

Numerosity paradigm


Number of trials left: 40

How many points were on the screen?

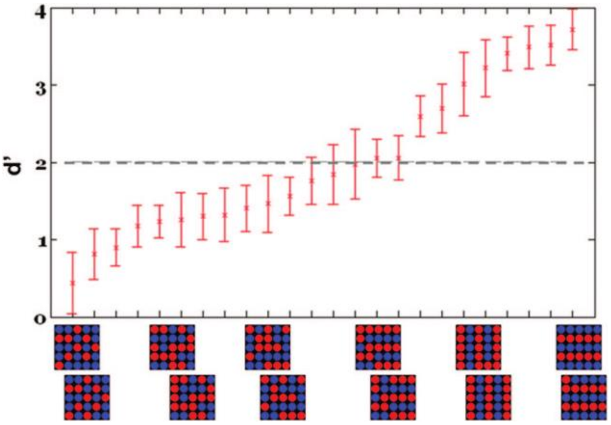
Results



Displays sampled from compositional functions are perceived as less numerous than displays sampled from spectral mixture functions.

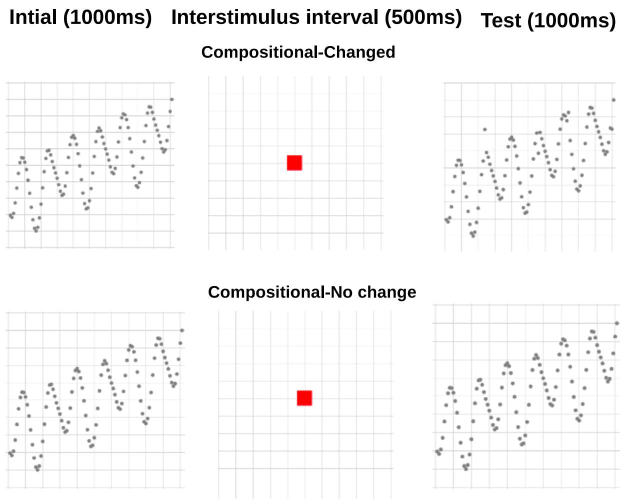
Change detection

Statistical regularities also aid change detection.



(Brady & Tenenbaum, 2013)

Functional change detection

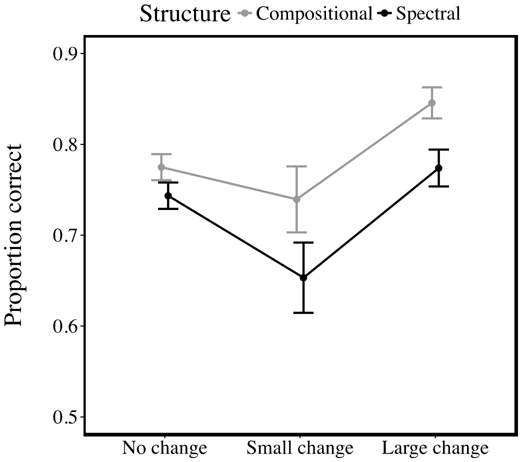


Intial (1000ms) Interstimulus interval (500ms) Test (1000ms)

Compositional-Changed

Compositional-No change

Results



Easier to detect changes in displays sampled from compositional functions.

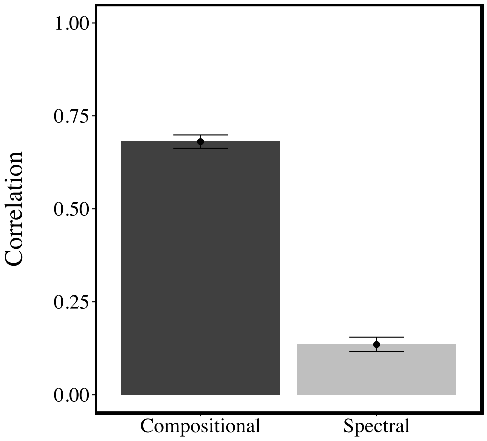
Computational modeling

Posterior probability that two displays were generated by different functions:

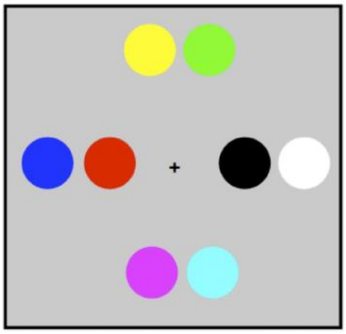
$$P(f_1 \neq f_2|\mathcal{D}_1, \mathcal{D}_2) = \frac{P(\mathcal{D}_1, \mathcal{D}_2|f_1 \neq f_2)}{P(\mathcal{D}_1, \mathcal{D}_2|f_1 \neq f_2) + P(\mathcal{D}_1, \mathcal{D}_2|f_1 = f_2)}$$

We can use the GP model to compute this probability in closed-form for any two displays.

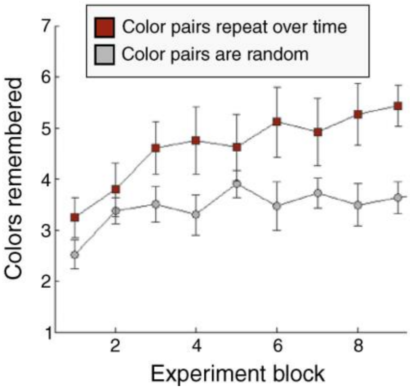
Model fit



Short-term memory



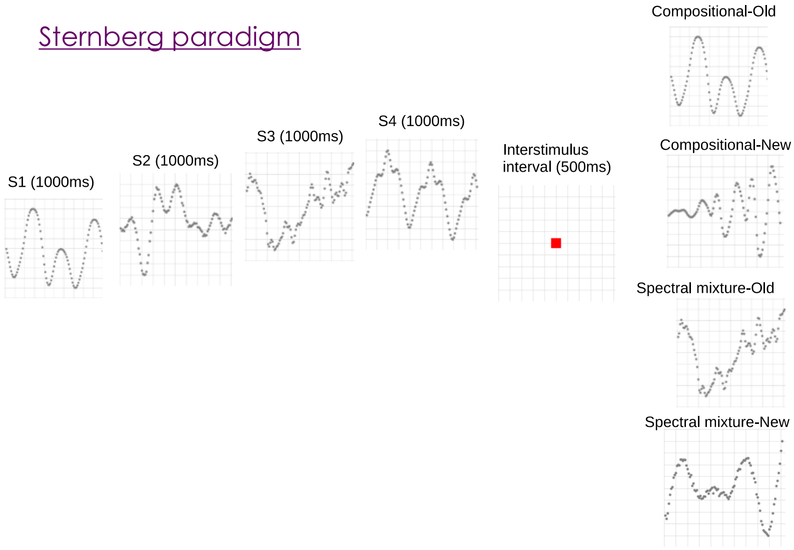
(Brady, Konkle & Alvarez, 2009)



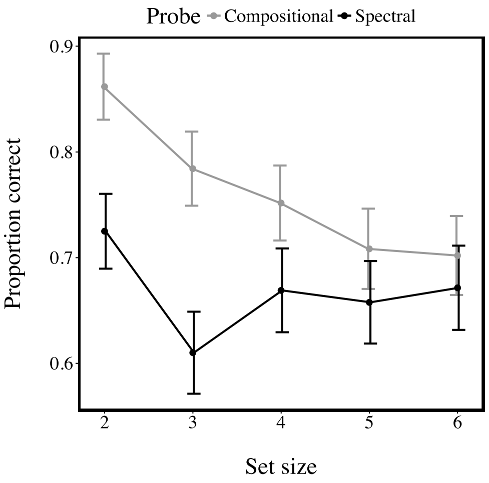
Statistical regularities aid visual short-term memory.

Functional short-term memory

Sternberg paradigm



Results



Compositional functions are more memorable/compressible.

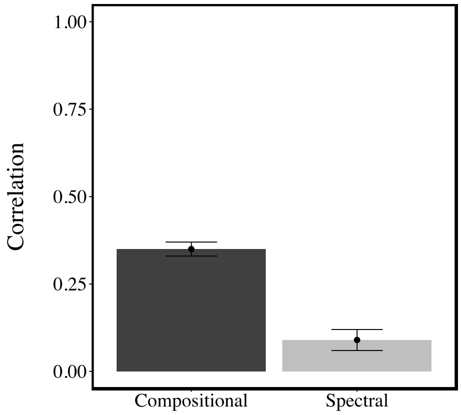
Computational modeling

Posterior probability that a probe display belongs to the study list:

$$P(f' \in f_{1:N}|Y) \propto \sum_{n=1}^{N} P(f' = f_n) P(\mathcal{D}_n, \mathcal{D}'|f' = f_n)$$

GP model can be used to compute this in closed-form.

Model fit



PART 4: PUTTING IT ALL TOGETHER

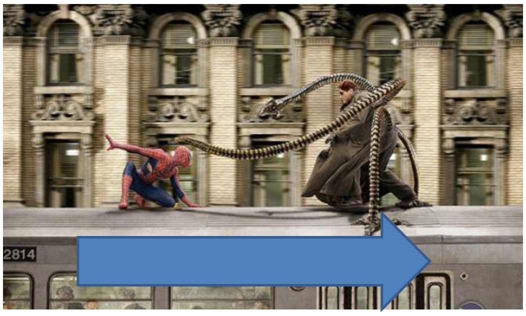
Composing the building blocks

- Mixture models, latent feature models, and function learning models can all be combined in interesting ways to capture more complexity
- Case study: motion perception

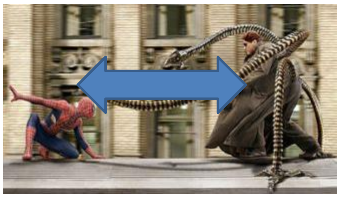
Case study: motion perception

How do we parse a moving scene?

Complex motions are composed of simpler motions



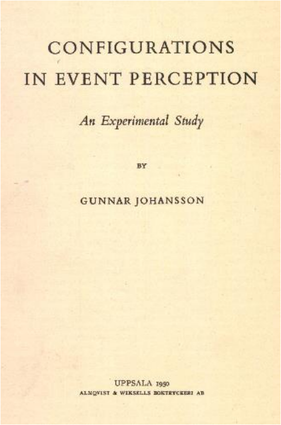
Motion relative to the background



Motion relative to the train

Motion relative to Dr. Octopus

Johansson's seminal contribution



CONFIGURATIONS
IN EVENT PERCEPTION

An Experimental Study

BY

GUNNAR JOHANSSON

UPPSALA 1950
ALMQVIST & WIKSELLS BOKTRYCKERI AB

1950

●

●

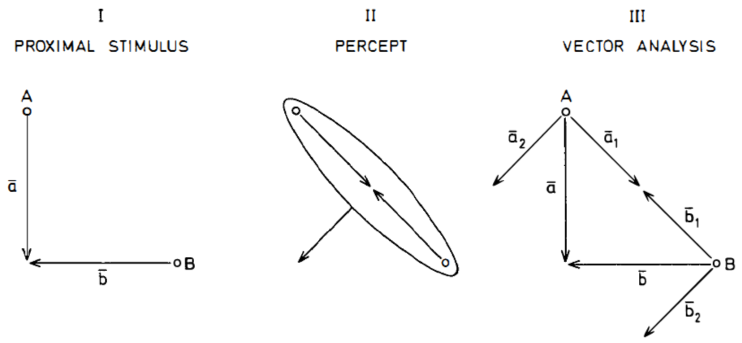
●

●

●

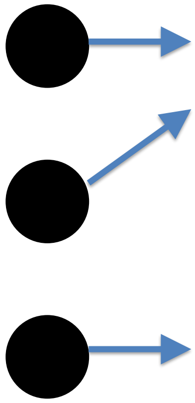
●

Vector analysis

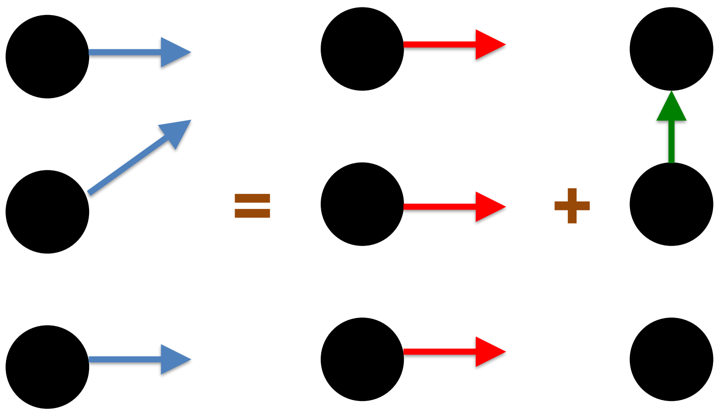


Johansson (1950)

Vector analysis



Vector analysis



Vector analysis

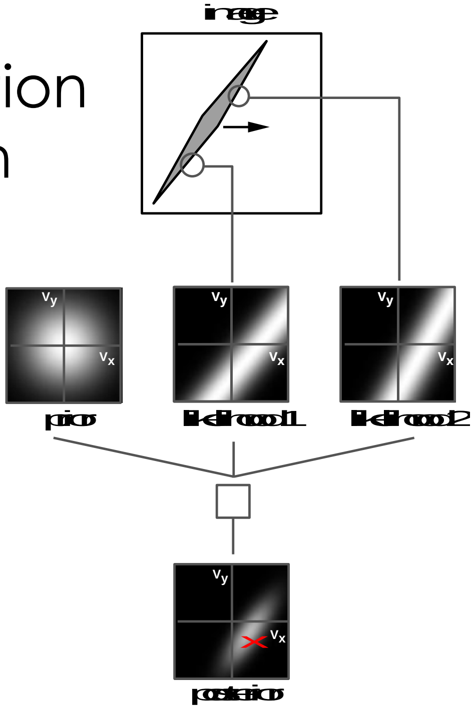
- Many different vector interpretations of a given motion pattern. How does our visual system choose one?

Vector analysis

- Many different vector interpretations of a given motion pattern. How does our visual system choose one?
- Appeal to "principles"
  - Minimum principle (Restle, 1979): simple motions preferred
  - Adjacency principle (Gogel, 1974): assign dots to nearest reference frame
- Need for a unifying computational theory

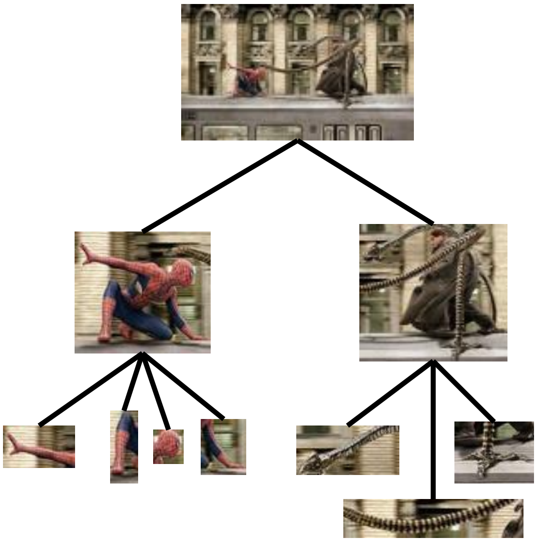
Bayesian motion perception

"slow and smooth"
Weiss & Adelson (1998)



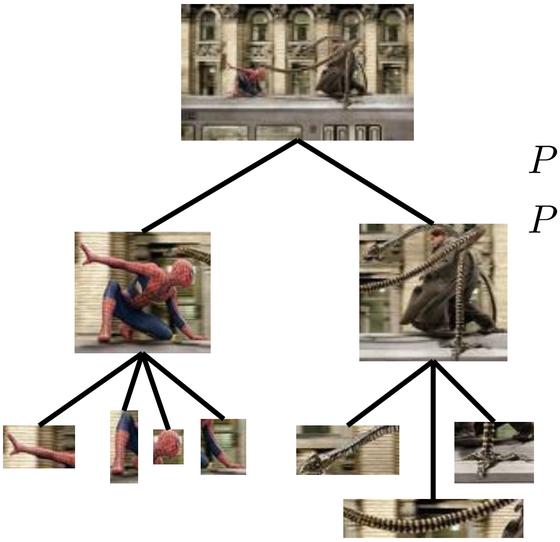
Bayesian vector analysis

Bayesian vector analysis



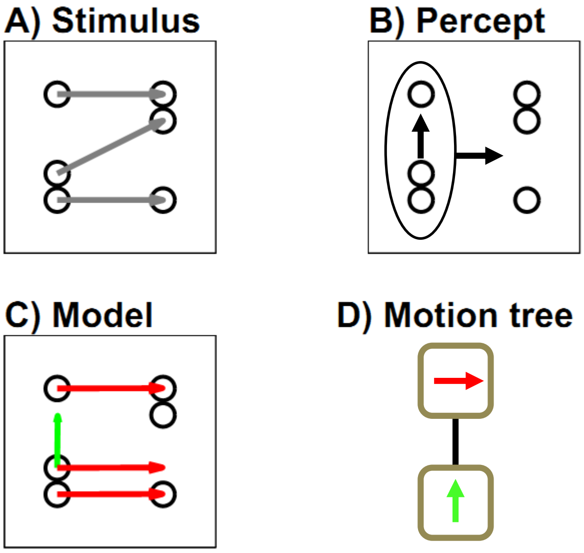
Motion tree

Bayesian vector analysis



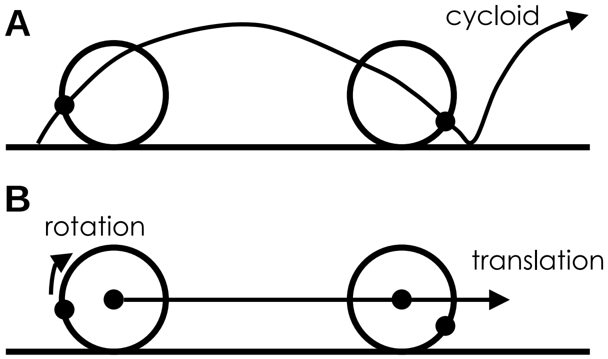
Motion tree

$$P(\text{tree}|\text{images}) \propto P(\text{images}|\text{tree})P(\text{tree})$$

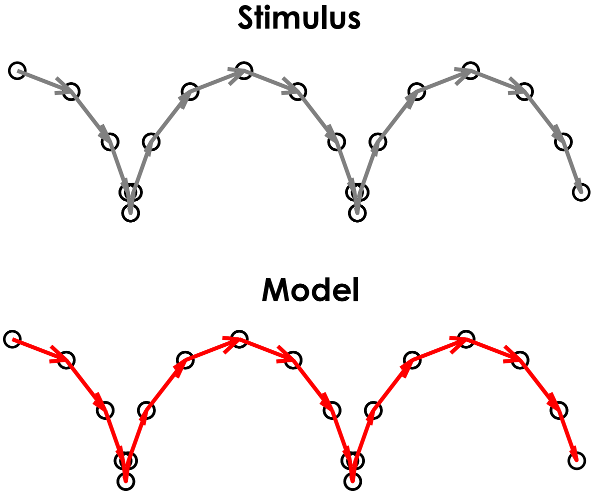


Gershman, Tenenbaum & Jaekel (2016)

Duncker wheel

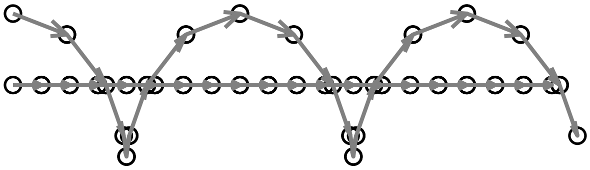


Simulations of the Duncker wheel

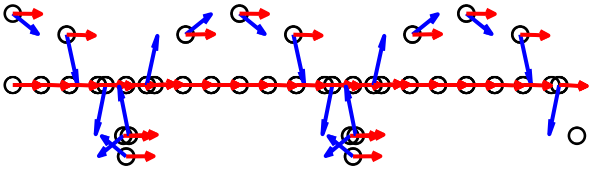


Simulations of the Duncker wheel

Stimulus

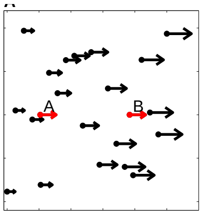


Model

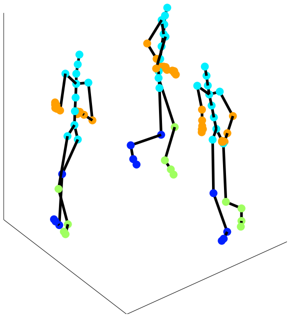


Other phenomena

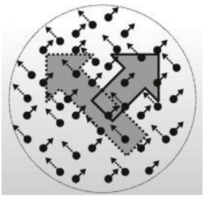
Motion contrast



Biological motion



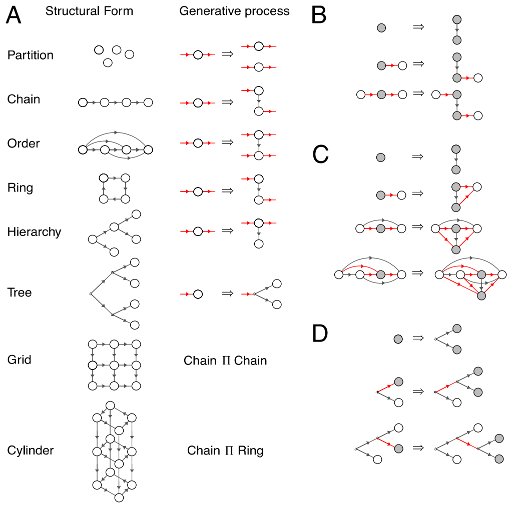
Transparent motion



Gershman, Tenenbaum & Jaekel (2016)

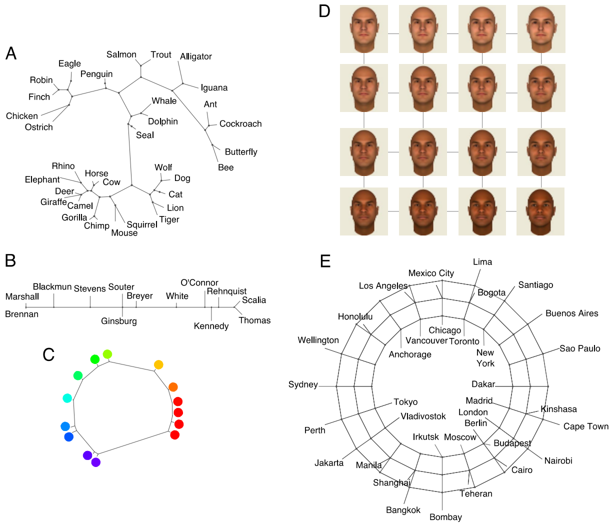
Can we discover structure automatically?

Can we discover structure automatically?



Graph grammars
(Kemp & Tenenbaum, 2008)

Can we discover structure automatically?



Automatic composition of modeling motifs

low-rank approximation $G\ !\ \ GG + G$

clustering $G\ !\ \ MG + G \mid GM^T + G$

$M\ !\ \ MG + G$

linear dynamics $G\ !\ \ CG + G \mid GC^T + G$

sparsity $G\ !\ \ \exp(G) \circ G$

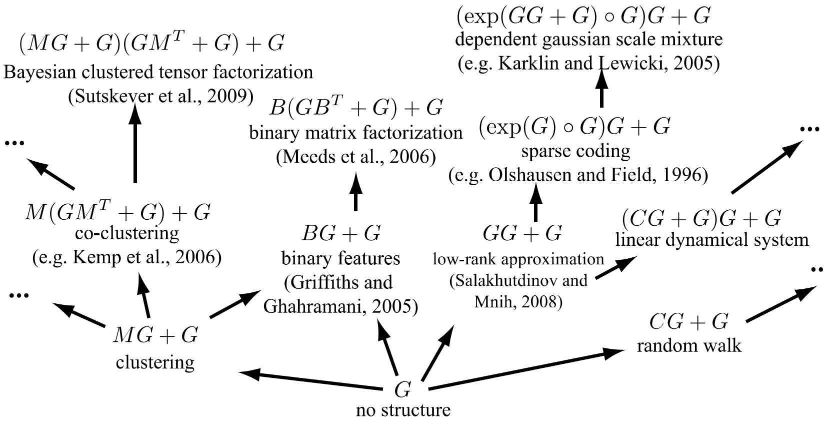
binary factors $G\ !\ \ BG + G \mid GB^T + G$

$B\ !\ \ BG + G$

$M\ !\ \ B$

Model grammars
(Grosse et al, 2012)

Automatic composition of modeling motifs



Summary

- Nonparametric Bayesian models can be used to flexibly capture structure that is “just right” (not too simple or complex)
- Growing experimental literature suggesting the brain implements these computational principles
- Basic building blocks (clusters, features, and functions) can be composed to capture a wider range of structures

Further reading

- Austerweil, Gershman, Tenenbaum, & Griffiths (2015). Structure and flexibility in Bayesian models of cognition. *Oxford Handbook of Computational and Mathematical Psychology*.
- Gershman & Blei (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*.
- Gershman & Niv (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*.