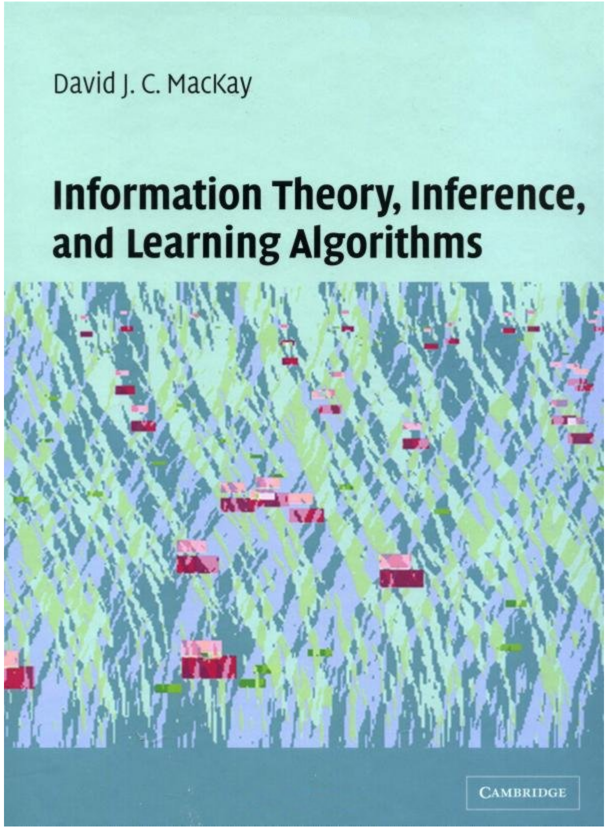
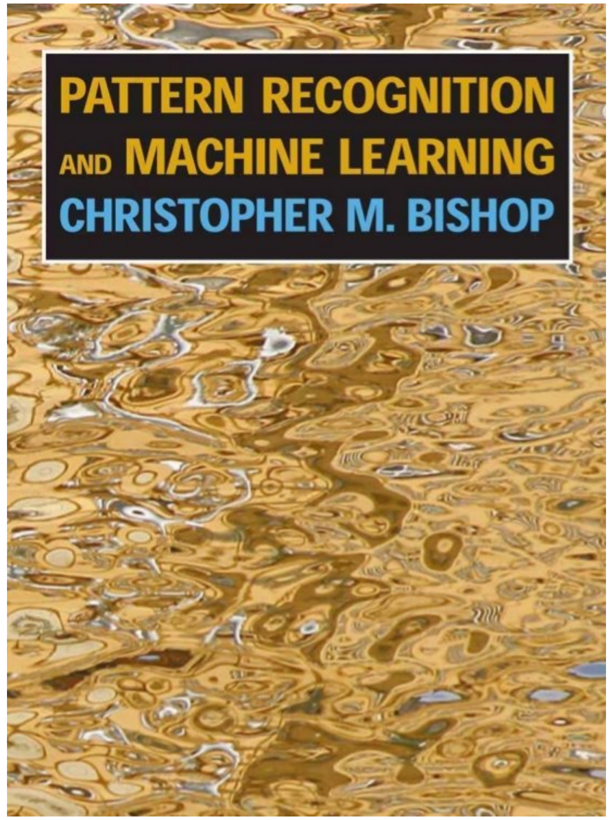
Probabilistic Modelling

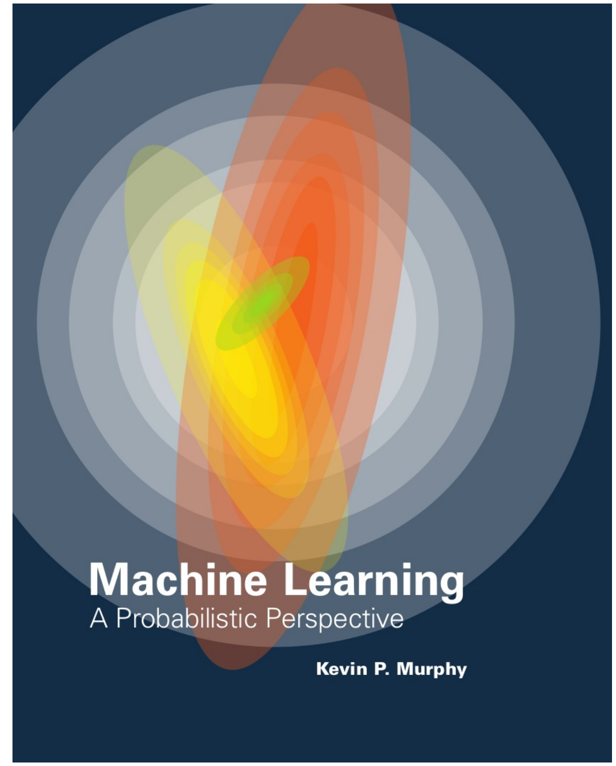
Image adapted from Sabbut from es,
commons.wikimedia.org/wiki/File:Dados_4_a_20_caras.jpg

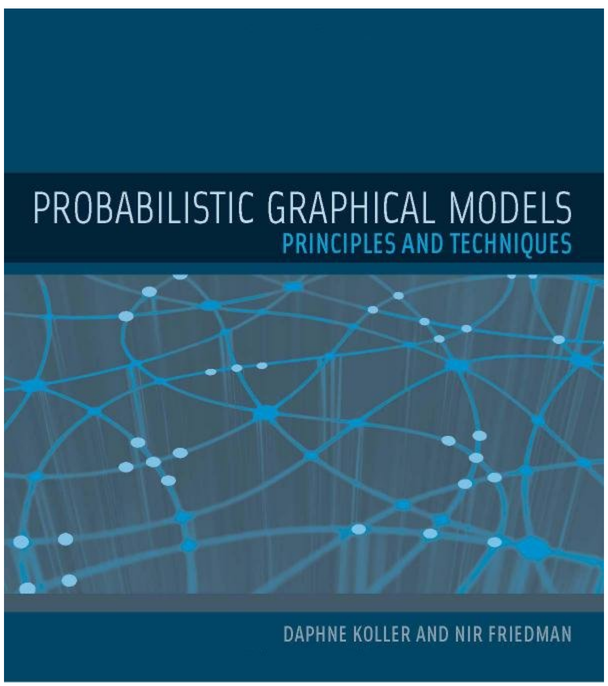
Iain Murray

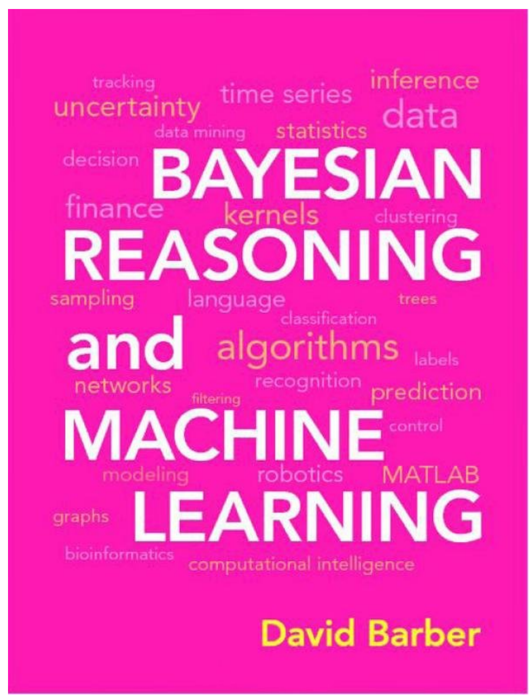
School of Informatics, University of Edinburgh

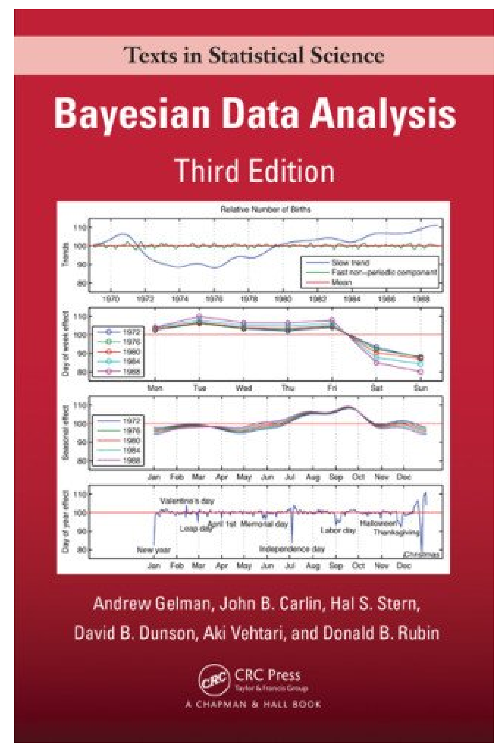

David J. C. MacKay
Information Theory, Inference, and Learning Algorithms
CAMBRIDGE


PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP


Machine Learning
A Probabilistic Perspective
Kevin P. Murphy


PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND NIR FRIEDMAN


tracking
uncertainty
time series
inference
data
data mining
statistics
decision
BAYESIAN
finance
kernels
clustering
REASONING
sampling
language
trees
classification
and
algorithms
labels
networks
recognition
filtering
prediction
MACHINE
control
modeling
robotics
MATLAB
graphs
LEARNING
bioinformatics
computational intelligence
David Barber


Texts in Statistical Science
Bayesian Data Analysis
Third Edition
Relative Number of Births
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin
CRC Press
A CHAPMAN & HALL BOOK

Card prediction

3 cards with coloured faces:

1. one white and one black face
2. two black faces
3. two white faces

I shuffle cards and turn them over randomly. I select a card and way-up uniformly at random and place it on a table.

I've taken this demo from David MacKay, although don't know its origin.

Card prediction

3 cards with coloured faces:

1. one white and one black face
2. two black faces
3. two white faces

I shuffle cards and turn them over randomly. I select a card and way-up uniformly at random and place it on a table.

Question: You see a black face. What is the probability that the other side of the same card is white?

$$P(x_2 = \mathtt{W} \,|\, x_1 = \mathtt{B}) = \quad {}^1\!/_3, \quad {}^1\!/_2, \quad {}^2\!/_3, \quad \text{other}, \quad \text{don't know?}$$

Roadmap

— **Probability fundamentals**

— Inferring a physical parameter

— Probabilistic models and machine learning

— Graphical models

— Monte Carlo basics, probabilistic inference in practice

Probability fundamentals

The sum rule:

$$P(A = a) = \sum_{b \in \mathcal{A}_B} P(A = a, B = b)$$

Probability fundamentals

The sum rule:

$$P(A\!=\!a) \;=\; \sum_{b \in \mathcal{A}_B} P(A\!=\!a, B\!=\!b)$$

Compressed:

$$P(a) \;=\; \sum_{b} P(a, b)$$

Situation made explicit:

$$P(a \,|\, c) \;=\; \sum_{b} P(a, b \,|\, c)$$

Probability fundamentals

The product rule:

$$P(a, b) \;=\; P(a \,|\, b)\, P(b) \;=\; P(b \,|\, a)\, P(a)$$

Probability fundamentals

The product rule:

$$P(a, b) \;=\; P(a \,|\, b)\, P(b) \;=\; P(b \,|\, a)\, P(a)$$

$$P(a, b \,|\, c) \;=\; P(a \,|\, b, c)\, P(b \,|\, c) \;=\; P(b \,|\, a, c)\, P(a \,|\, c)$$

Applied recursively, “the chain rule”:

$$P(a, b, c, d) \;=\; P(a)\, P(b \,|\, a)\, P(c \,|\, a, b)\, P(d \,|\, a, b, c)$$

$$P(\mathbf{x}) \;=\; P(x_1) \prod_{d=2}^{D} P(x_d \,|\, \mathbf{x}_{<d})$$

Probability fundamentals

Sum rule: $P(a) = \sum_b P(a, b)$

Product rule: $P(a, b) = P(a \,|\, b)\, P(b) = P(b \,|\, a)\, P(a)$

Probability fundamentals

Sum rule: $P(a) = \sum_b P(a, b)$

Product rule: $P(a, b) = P(a \,|\, b)\, P(b) = P(b \,|\, a)\, P(a)$

Bayes rule:

$$P(a \,|\, b) = \frac{P(b \,|\, a)\, P(a)}{P(b)}$$

Probability fundamentals

Sum rule: $P(a) = \sum_b P(a,b)$

Product rule: $P(a,b) = P(a \,|\, b)\, P(b) = P(b \,|\, a)\, P(a)$

Bayes rule:

$$P(a \,|\, b) = \frac{P(b \,|\, a)\, P(a)}{P(b)}$$

Probability of everything:

$$\begin{aligned} P(a \,|\, b) &\propto P(a,b) \\ &\propto \sum_c P(a,b,c) \end{aligned}$$

Card prediction

3 cards with coloured faces:

1. one white and one black face
2. two black faces
3. two white faces

I shuffle cards and turn them over randomly. I select a card and way-up uniformly at random and place it on a table.

Question: You see a black face. What is the probability that the other side of the same card is white?

$$P(x_2 = \mathtt{W} \,|\, x_1 = \mathtt{B}) = \quad {}^1\!/_3, \quad {}^1\!/_2, \quad {}^2\!/_3, \quad \text{other?}$$

Notes on the card prediction problem:

This card problem is Ex. 8.10a), MacKay's textbook, p142.

It is *not* the same as the famous 'Monty Hall' puzzle: Ex. 3.8–9 and `http://en.wikipedia.org/wiki/Monty_Hall_problem`

The Monty Hall problem is also worth understanding. Although the card problem is (hopefully) less controversial and more straightforward. The process by which a card is selected should be clear: $P(c) = 1/3$ for $c = 1, 2, 3$, and the face you see first is chosen at random: e.g., $P(x_1 = \mathtt{B} \,|\, c = 1) = 0.5$.

Many people get this puzzle wrong on first viewing, including more than half of previous summer school audiences (it's easy to mess up given limited time). If you got the answer right immediately, maybe it will be an example to help in your own teaching.

How do we solve it formally?

Use Bayes rule?

$$P(x_2 = \mathtt{W} \,|\, x_1 = \mathtt{B}) = \frac{\boxed{P(x_1 = \mathtt{B} \,|\, x_2 = \mathtt{W})}\, P(x_2 = \mathtt{W})}{P(x_1 = \mathtt{B})}$$

The $\boxed{\text{boxed}}$ term is no more obvious than the answer!

Bayes rule is used to 'invert' forward generative processes that we understand.

The first step to solve inference problems is to write down a model of your data.

The card game model

Cards: 1) $\texttt{B}|\texttt{W}$, 2) $\texttt{B}|\texttt{B}$, 3) $\texttt{W}|\texttt{W}$

$$P(c) = \begin{cases} 1/3 & c = 1, 2, 3 \\ 0 & \text{otherwise.} \end{cases}$$

$$P(x_1 \!=\! \texttt{B} \,|\, c) = \begin{cases} 1/2 & c = 1 \\ 1 & c = 2 \\ 0 & c = 3 \end{cases}$$

Bayes rule can 'invert' this to tell us $P(c \,|\, x_1 \!=\! \texttt{B})$; infer the generative process for the data we have.

Inferring the card

Cards: 1) $\texttt{B}|\texttt{W}$, 2) $\texttt{B}|\texttt{B}$, 3) $\texttt{W}|\texttt{W}$

$$P(c \,|\, x_1{=}\texttt{B}) \;=\; \frac{P(x_1{=}\texttt{B} \,|\, c)\, P(c)}{P(x_1{=}\texttt{B})} \;\propto\; P(x_1{=}\texttt{B} \,|\, c)\, P(c)$$

$$\propto \begin{cases} {}^1\!/_2 \cdot {}^1\!/_3 = {}^1\!/_6 & c = 1 \\ 1 \cdot {}^1\!/_3 = {}^1\!/_3 & c = 2 \\ 0 & c = 3 \end{cases}$$

$$= \begin{cases} {}^1\!/_3 & c = 1 \\ {}^2\!/_3 & c = 2 \end{cases}$$

Q *"But aren't there two options given a black face, so it's 50–50?"*
A There are two options, but the likelihood for one of them is $2\times$ bigger

Predicting the next outcome

For this problem we can spot the answer, for more complex problems we want a formal means to proceed.

$P(x_2 \,|\, x_1 \!=\! \mathtt{B})$?

Need to introduce c to use expressions we know:

$$\begin{aligned} P(x_2 \,|\, x_1 \!=\! \mathtt{B}) &= \sum_{c \in 1,2,3} P(x_2, c \,|\, x_1 \!=\! \mathtt{B}) \\ &= \sum_{c \in 1,2,3} P(x_2 \,|\, x_1 \!=\! \mathtt{B}, c)\, P(c \,|\, x_1 \!=\! \mathtt{B}) \end{aligned}$$

Predictions we would make if we knew the card, weighted by the posterior probability of that card.

$P(x_2 \!=\! \mathtt{W} \,|\, x_1 \!=\! \mathtt{B}) = 1/3$

Strategy for solving inference and prediction problems:

When interested in predicting something y, we often find we can't immediately write down mathematical expressions for $P(y \,|\, \text{data})$.

So we introduce stuff, z, that is related to the data and/or y:

$$P(y \,|\, \text{data}) = \textstyle\sum_z P(y, z \,|\, \text{data})$$

by using the sum rule. And then split it up:

$$P(y \,|\, \text{data}) = \textstyle\sum_z P(y \,|\, z, \text{data})\, P(z \,|\, \text{data})$$

using the product rule. If knowing extra stuff z we can predict y, we are set: weight all such predictions by the posterior probability of the stuff ($P(z \,|\, \text{data})$, found with Bayes rule).

Sometimes the extra stuff summarizes everything we need to know to make a prediction:

$$P(y \,|\, z, \text{data}) = P(y \,|\, z)$$

although not in the card game above.

Not convinced?

Not everyone believes the answer to the card game question.

Sometimes probabilities are counter-intuitive. I'd encourage you to write simulations of these games if you are at all uncertain. Here is an Octave/Matlab simulator I wrote for the card game question:

```
cards = [1 1;
         0 0;
         1 0];
num_cards = size(cards, 1);

N = 0; % Number of times first face is black
kk = 0; % Out of those, how many times the other side is white

for trial = 1:1e6
    card = ceil(num_cards * rand());
    face = 1 + (rand < 0.5);
    other_face = (face==1) + 1;
    x1 = cards(card, face);
    x2 = cards(card, other_face);

    if x1 == 0
        N = N + 1;
        kk = kk + (x2 == 1);
    end
end

approx_probability = kk / N
```

The probability of everything

c	x_1	x_2	$P(c, x_1, x_2)$
1	B	B	0
1	B	W	1/6
1	W	B	1/6
1	W	W	0
2	B	B	1/3
2	B	W	0
2	W	B	0
2	W	W	0
3	B	B	0
3	B	W	0
3	W	B	0
3	W	W	1/3

Cards: 1) $\mathtt{B}|\mathtt{W}$, 2) $\mathtt{B}|\mathtt{B}$, 3) $\mathtt{W}|\mathtt{W}$

$$P(x_2 \,|\, x_1 = \mathtt{B})$$

$$\propto \sum_c P(c, x_1 = \mathtt{B}, x_2)$$

$$\propto \begin{cases} 0 + {}^1\!/\!_3 + 0 & x_2 = \mathtt{B} \\ {}^1\!/\!_6 + 0 + 0 & x_2 = \mathtt{W} \end{cases}$$

$$= \begin{cases} {}^2\!/\!_3 & x_2 = \mathtt{B} \\ {}^1\!/\!_3 & x_2 = \mathtt{W} \end{cases}$$

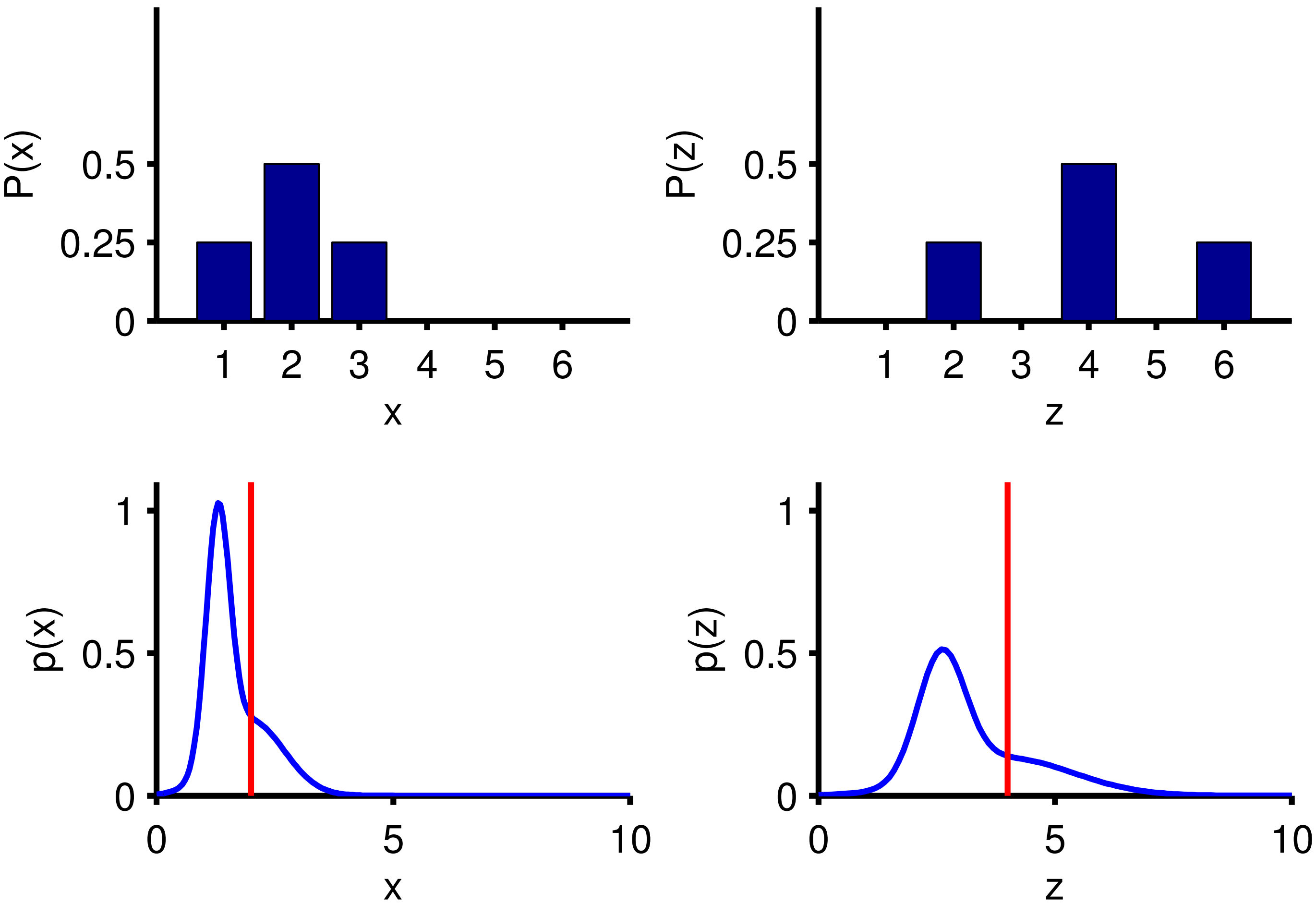
Final theory: real-valued variables

Probability densities:

$$P(a < X < b) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(x - {}^{\delta}\!/_{2} < X < x + {}^{\delta}\!/_{2}) \approx p(x)\delta$$

Transformations



$$z = 2x$$

$$P(z\!=\!4) = P(x\!=\!2)$$

$$p(z\!=\!4) \neq p(x\!=\!2)$$

$$p(z\!=\!4) = \frac{p(x\!=\!2)}{2}$$

Probability densities: $\int p(x)\,\mathrm{d}x = 1$

Nonlinear transformations

For 1–1 mappings between small elements δx and δz:

$$p(x)\,\delta x = p(z)\,\delta z$$

Taking limits:

$$p(z) \;=\; p(x(z))\left|\frac{\mathrm{d}x}{\mathrm{d}z}\right| \;=\; p(x(z))\Big/\left|\frac{\mathrm{d}z}{\mathrm{d}x}\right|$$

Example:

$$p(\sigma^2) \;=\; \frac{p(\log\sigma^2)}{\sigma^2}$$

Multivariate version with Jacobian:

$$p(\mathbf{z}) \;=\; p(\mathbf{x})\begin{vmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} & \cdots & \frac{\partial x_1}{\partial z_D} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_D}{\partial z_1} & \frac{\partial x_D}{\partial z_2} & \cdots & \frac{\partial x_D}{\partial z_D} \end{vmatrix}$$

delta functions

Let $z = 2x$ again

Discrete:

$$P(z \,|\, x) \;=\; \mathbb{I}(z\!=\!2x) \;=\; \delta_{z,2x} \;=\; \begin{cases} 1 & z = 2x \\ 0 & \text{otherwise} \end{cases}$$

(Kronecker delta)

Continuous:

$$p(z \,|\, x) \;=\; \delta(z - 2x) \;=\; \lim_{\sigma^2 \to 0} \mathcal{N}(z;\, 2x,\, \sigma^2)$$

(Dirac delta)

$p(z\!=\!2x \,|\, x) = \infty$, *not* 1!

deltas and change of variables

Let $z = 2x$, or $p(z \,|\, x) = \delta(z - 2x)$

$$p(z) = \int p(x, z)\, \mathrm{d}x$$
$$= \int p(x)\, \delta(z - 2x)\, \mathrm{d}x$$

$\delta \Rightarrow$ "$z = 2x$" $\Rightarrow$ "$x = {}^{z}\!/_{2}$"

but $p(z) \neq p(x = {}^{z}\!/_{2})$

deltas and change of variables

Let $z = 2x$, or $p(z \,|\, x) = \delta(z - 2x)$

$$p(z) = \int p(x, z)\, \mathrm{d}x$$
$$= \int p(x)\, \delta(z - 2x)\, \mathrm{d}x$$

$\delta \Rightarrow$ "$z = 2x$" $\Rightarrow$ "$x = {}^{z}/_{2}$"
but $p(z) \neq p(x = {}^{z}/_{2})$

Change of variables, $u = 2x$, $x = {}^{u}/_{2}$, $\mathrm{d}x = \mathrm{d}u/2$

$$p(z) = \int p(x = {}^{u}/_{2})\, \delta(z - u)\, \mathrm{d}u/2$$
$$= \tfrac{1}{2}\, p(x = {}^{z}/_{2})\,,$$ as before

Summary: real-valued variables

Be careful with determinism, however expressed:

— changes of variables

— distributions constrained to a manifold

— Gaussians with low-rank covariance matrices

— MCMC updates within a subspace

Roadmap

- Probability fundamentals
- **Inferring a physical parameter**
- Probabilistic models and machine learning
- Graphical models
- Monte Carlo basics, probabilistic inference in practice

Infer motion from a snapshot

In 1D, stars do *simple harmonic motion* (SHM)

Common orbital frequency ω

$\Rightarrow$ mass black hole

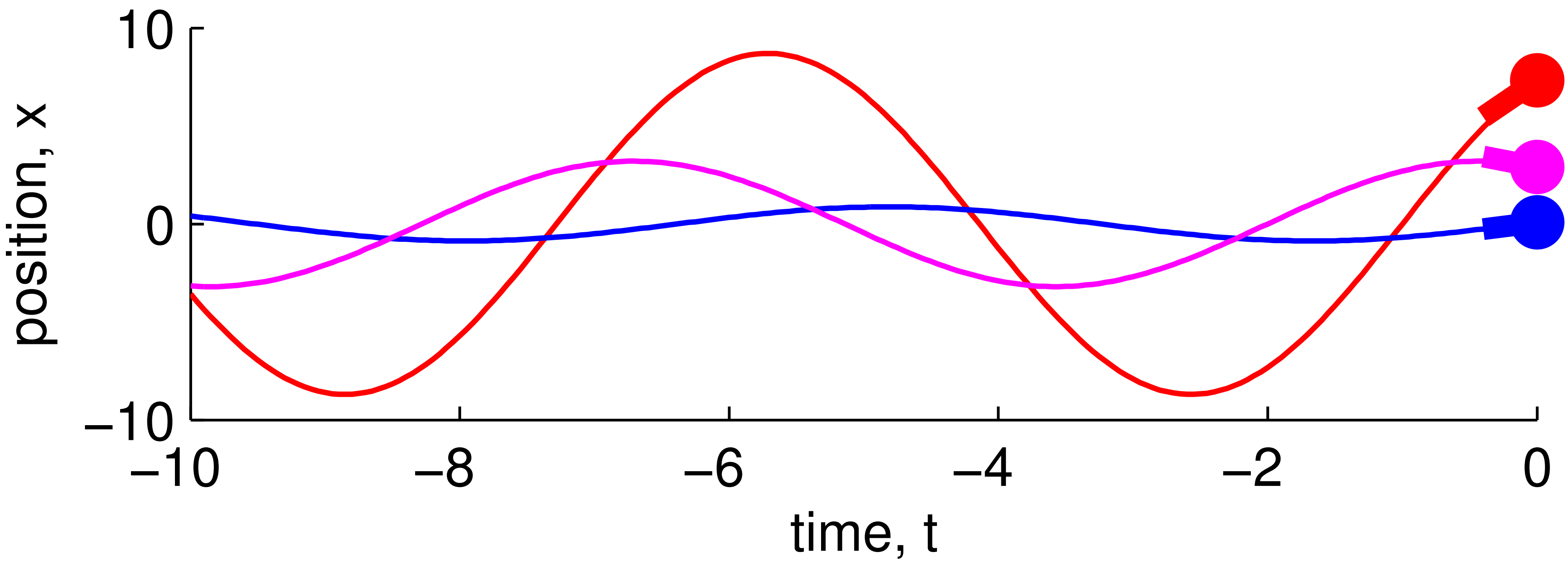
Thanks: David W. Hogg (NYU) first told me this problem

SHM equations and model

Positions and velocities:

$$x_n = A_n \sin(\omega t + \phi_n)$$

$$v_n = \frac{\mathrm{d}x_n}{\mathrm{d}t} = A_n \,\omega \cos(\omega t + \phi_n)$$



SHM equations and model

Positions and velocities:

$$x_n = A_n \sin(\omega t + \phi_n)$$

$$v_n = \frac{\mathrm{d}x_n}{\mathrm{d}t} = A_n \omega \cos(\omega t + \phi_n)$$

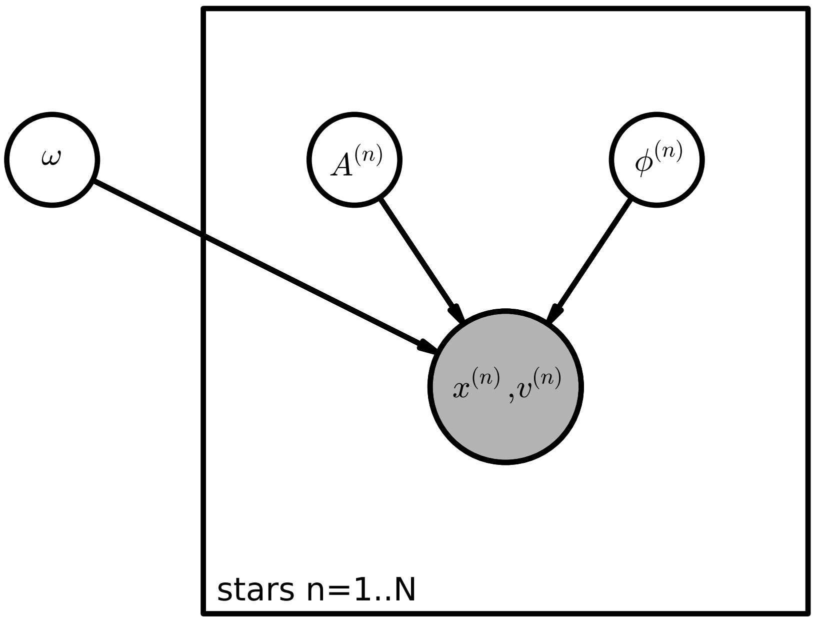
Evaluate at $t = 0$ (wlog)

$$x_n = A_n \sin \phi_n$$

$$v_n = A_n \omega \cos \phi_n$$

SHM equations and model

$$x_n = A_n \sin\phi_n\,, \qquad v_n = A_n\,\omega\cos\phi_n$$



Priors:

$$\log\omega \sim \mathrm{Uniform}[\log\omega_{\min}, \log\omega_{\max}]$$

$$\phi_n \sim \mathrm{Uniform}[0, 2\pi]$$

$$\log A_n \sim \mathrm{Uniform}[\log A_{\min}, \log A_{\max}]$$

$$p(\omega, \{A_n, \phi_n, x_n, v_n\}) \\ = p(\omega)\prod_n p(A_n)\,p(\phi_n)\,p(x_n, v_n\,|\,\omega, A_n, \phi_n)$$

Inferring the frequency

$$p(\omega \,|\, \{x_n, v_n\}) \;\propto\; \int \mathrm{d}A \int \mathrm{d}\phi \;\; p(\omega, \{A_n, \phi_n, x_n, v_n\})$$

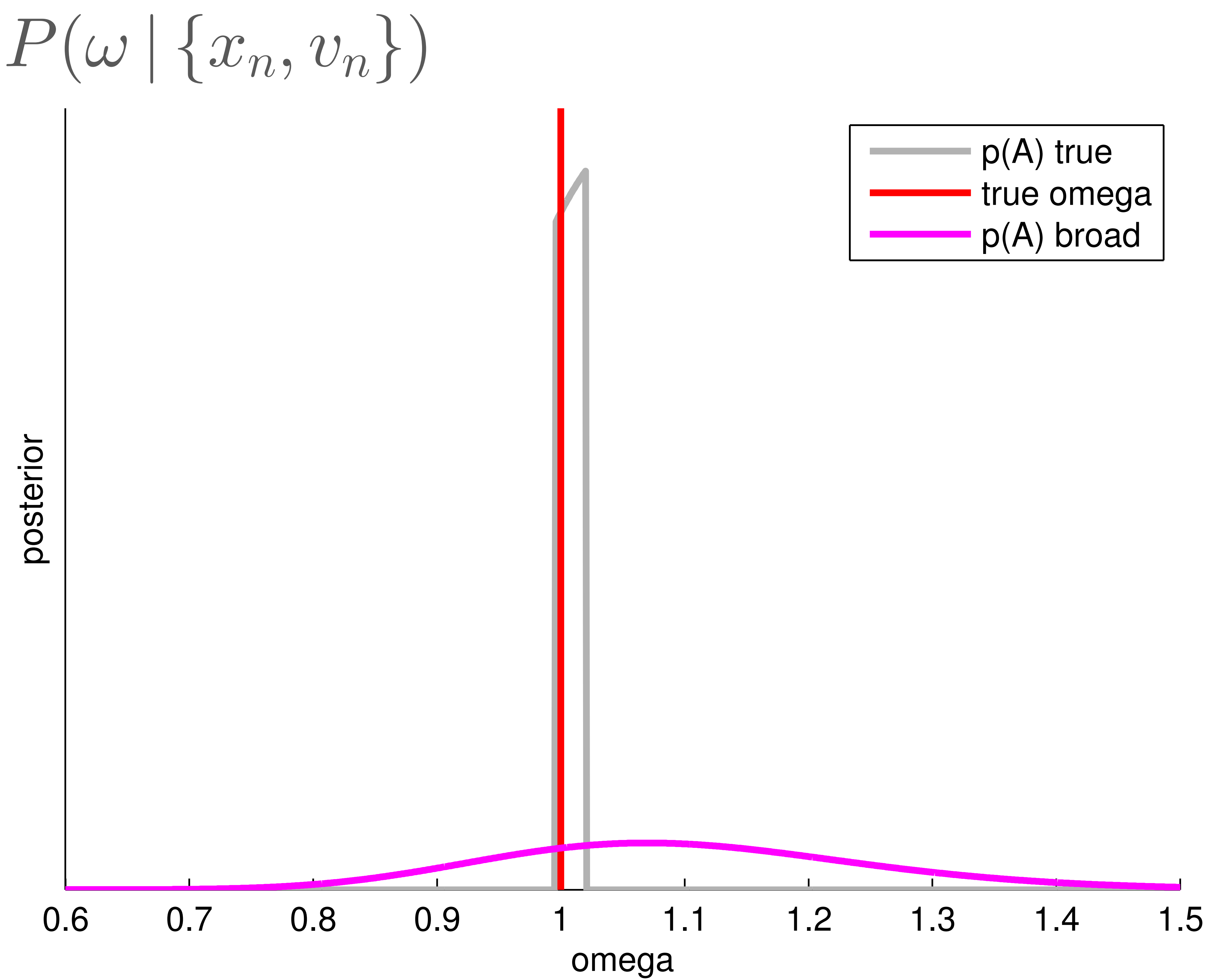
Substitute and integrate delta functions carefully. . .

or. . .

$$p(\omega \,|\, \{x_n, v_n\}) \;\propto\; p(\omega) \int \mathrm{d}A \int \mathrm{d}\phi \;\; p(\{A_n, \phi_n, x_n, v_n\} \,|\, \omega)$$

$$\propto\; p(\omega) \prod_n p(x_n, v_n \,|\, \omega)$$

where $p(x_n, v_n \,|\, \omega)$ is $p(A_n, \phi_n)$ divided by a simple Jacobian of a transformation


$P(\omega \mid \{x_n, v_n\})$
p(A) true
true omega
p(A) broad
posterior
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
omega

The mistake

Reasonable prior for one amplitude (fine):

$$p(\log A_n) = \frac{1}{\log A_{\max} - \log A_{\min}}$$

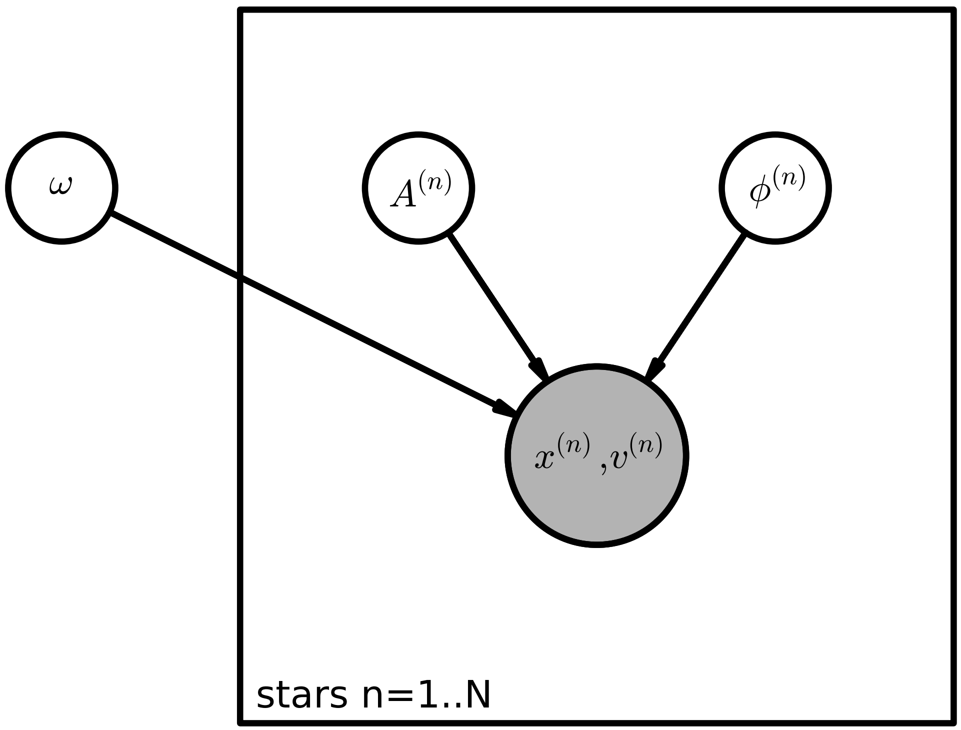
$$A_{\min} < A_n < A_{\max}$$

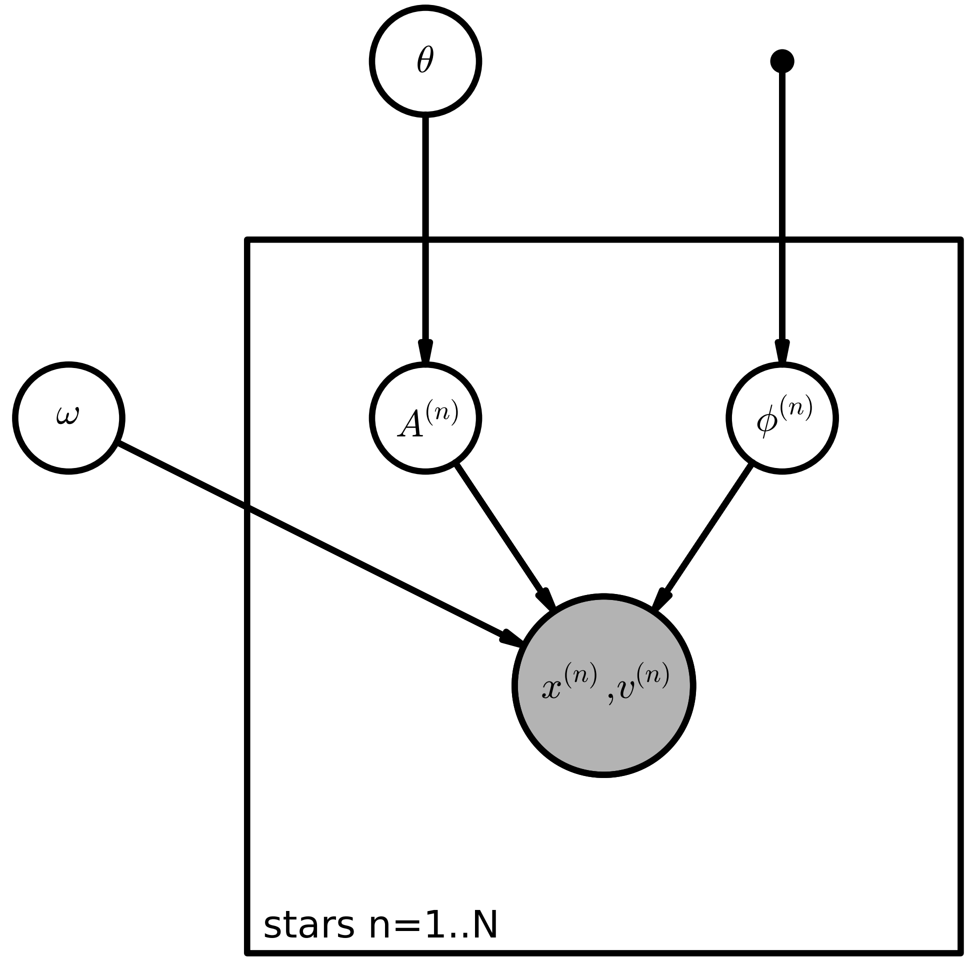
Does *not* extend to:

$$p(\{\log A_n\}_{n=1}^{N}) = \prod_n \frac{1}{\log A_{\max} - \log A_{\min}}$$

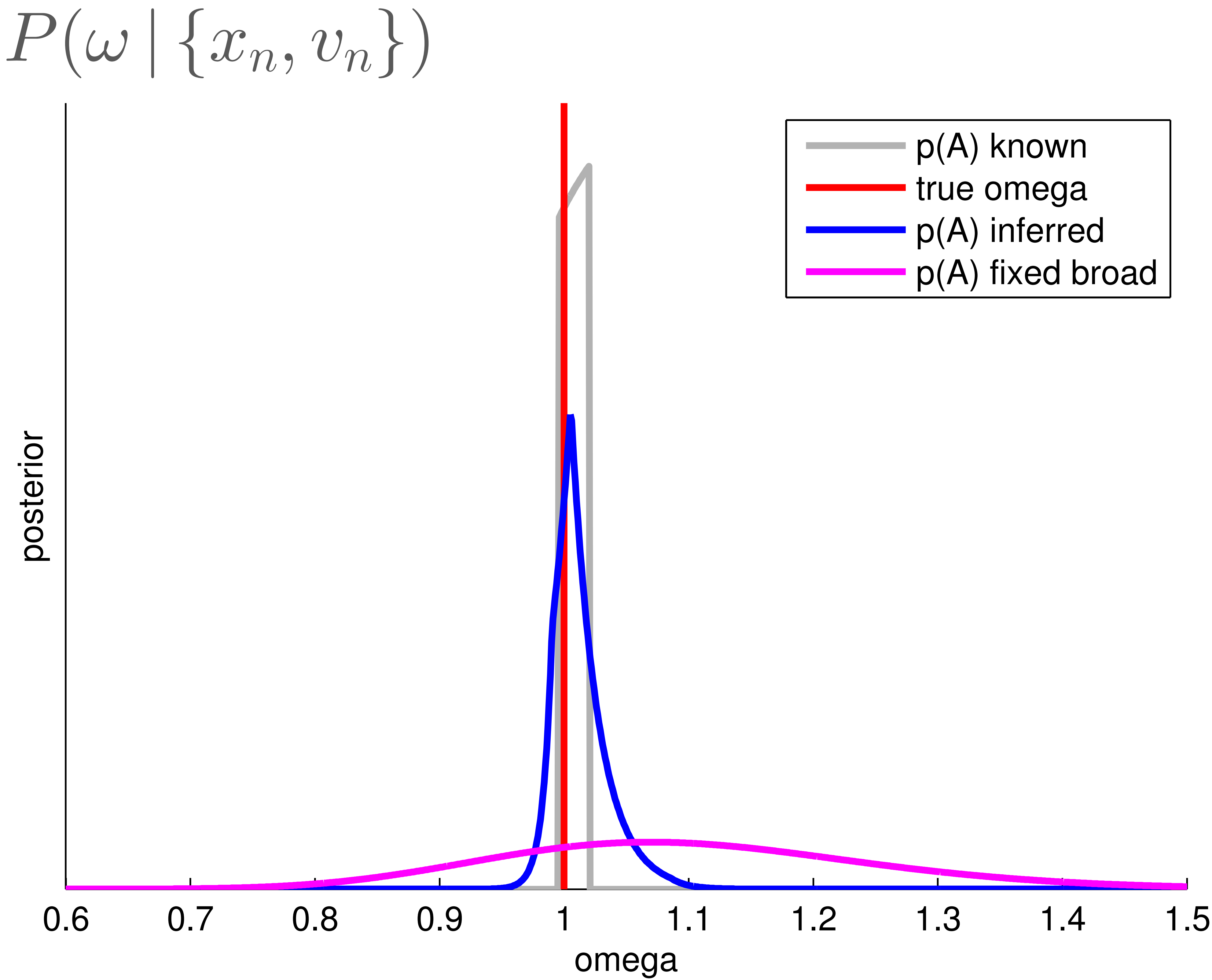
$$A_{\min} < A_n < A_{\max}, \quad \forall n$$

Fixing the graphical model





$$p(\omega, \theta, \{A_n, \phi_n, x_n, v_n\}) = p(\omega)\, p(\theta) \prod_n p(\phi_n)\, p(A_n \,|\, \theta)\, p(x_n, v_n \,|\, \omega, A_n, \phi_n)$$



$P(\omega \mid \{x_n, v_n\})$
p(A) known
true omega
p(A) inferred
p(A) fixed broad
posterior
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
omega

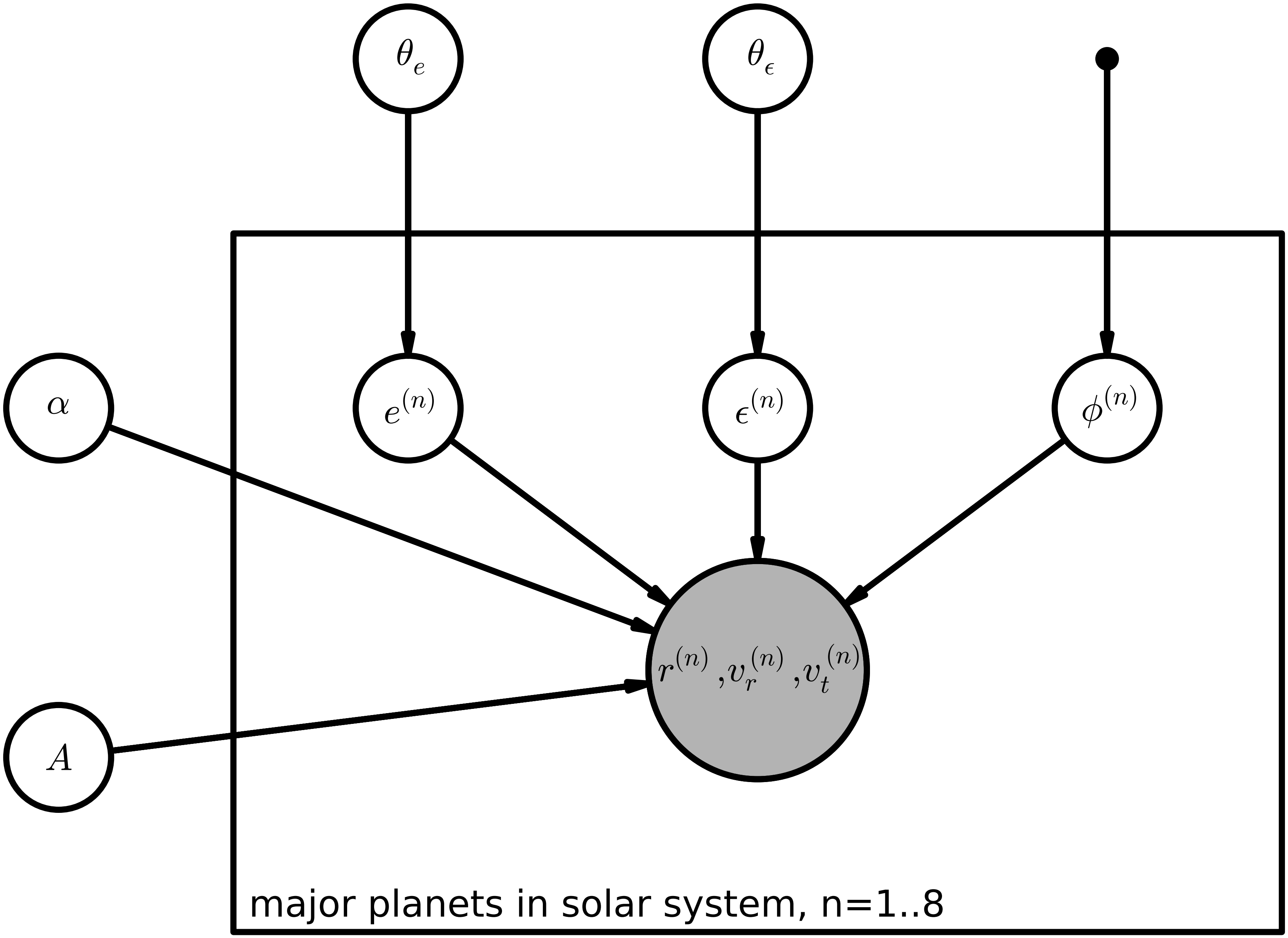
Acceleration law around the sun

$$a(r) = -A\left(\frac{r}{r_0}\right)^{-\alpha}$$

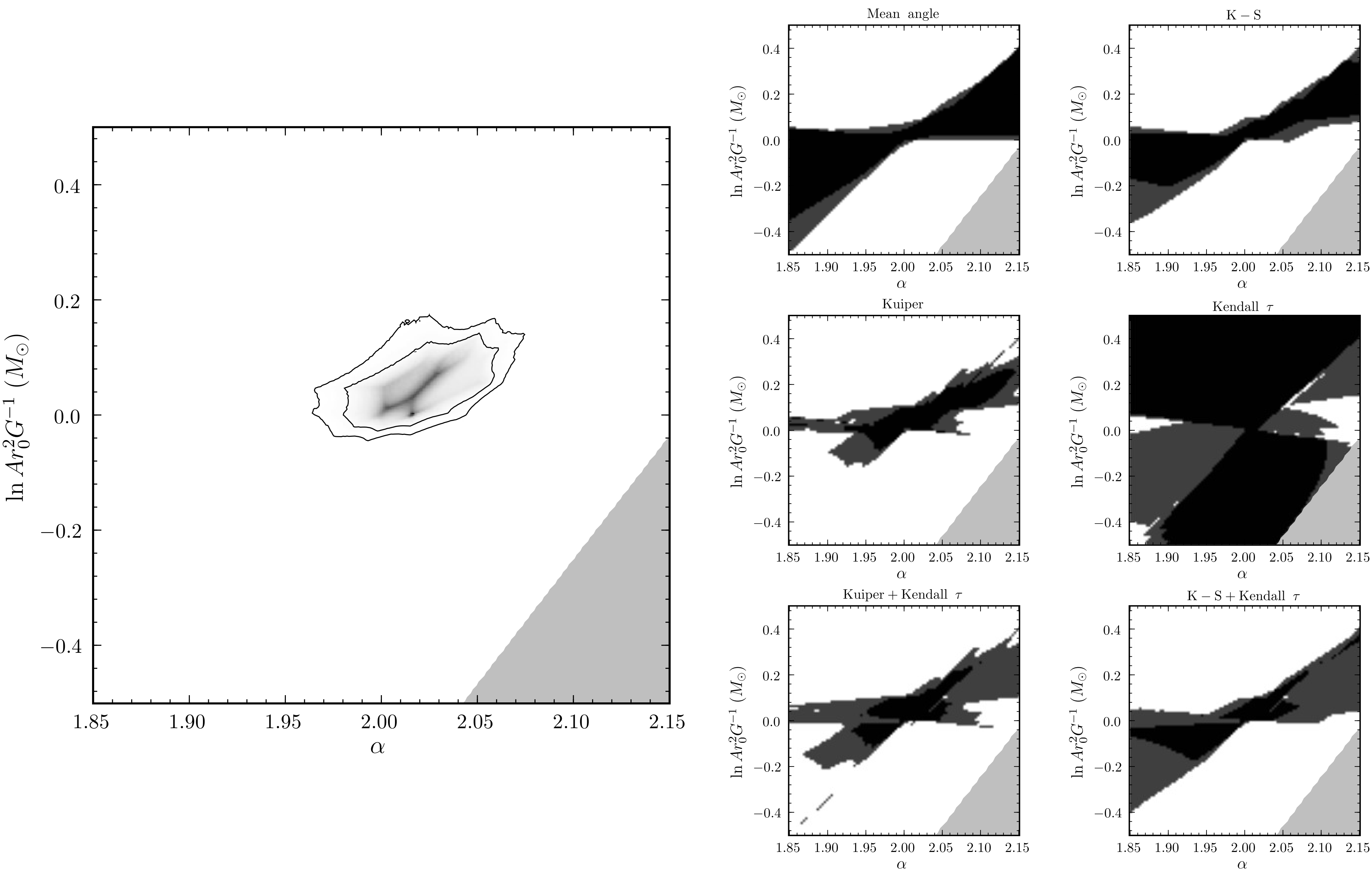
From a snapshot:

8 planet positions and velocities

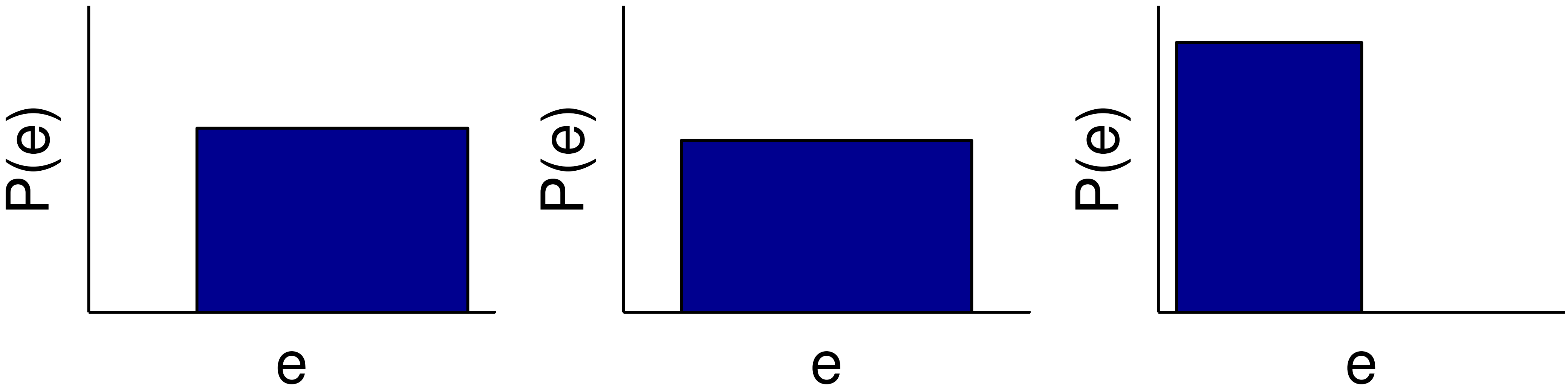
Solarsystem snapshot model



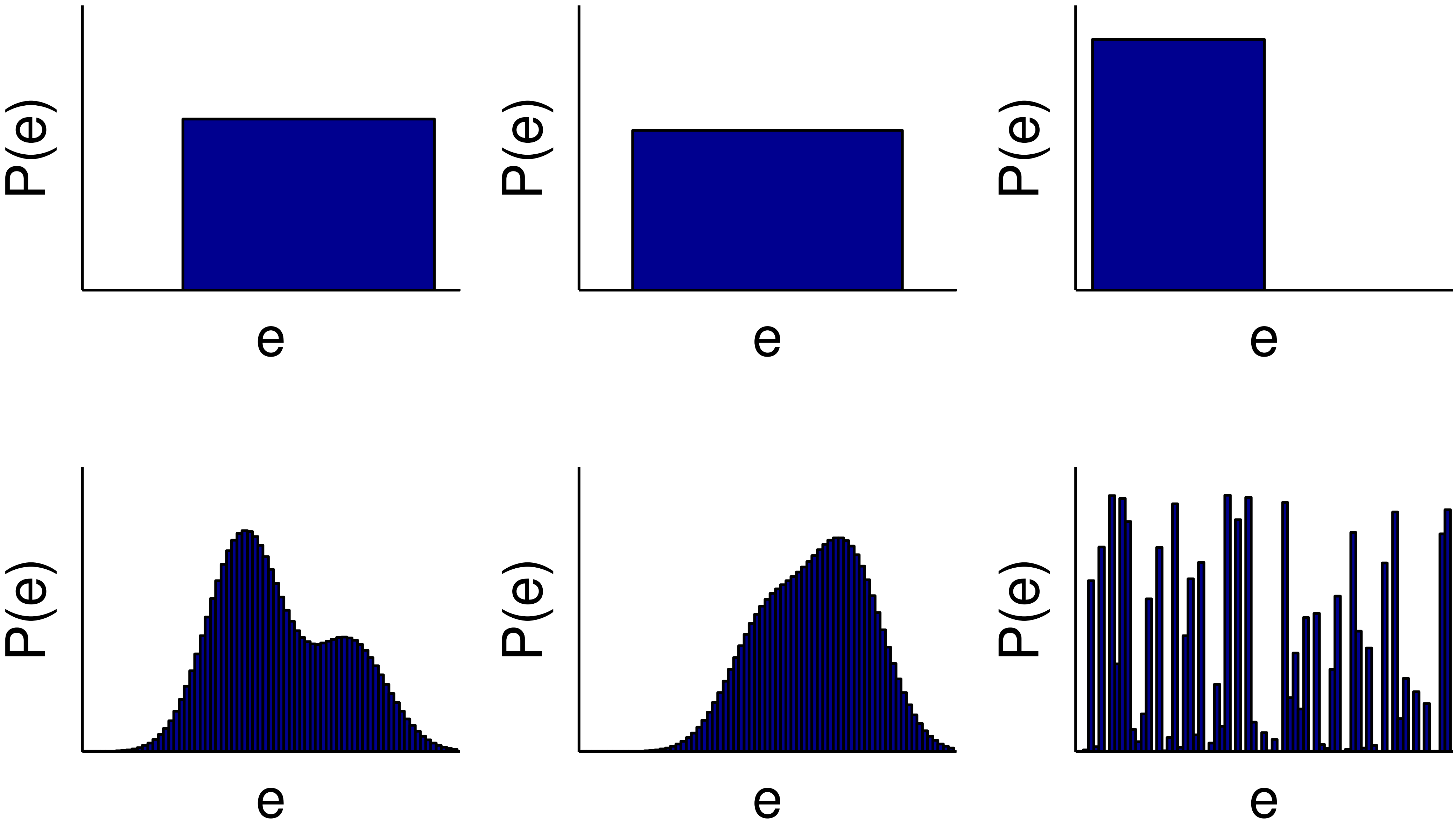
Inferences about the Sun



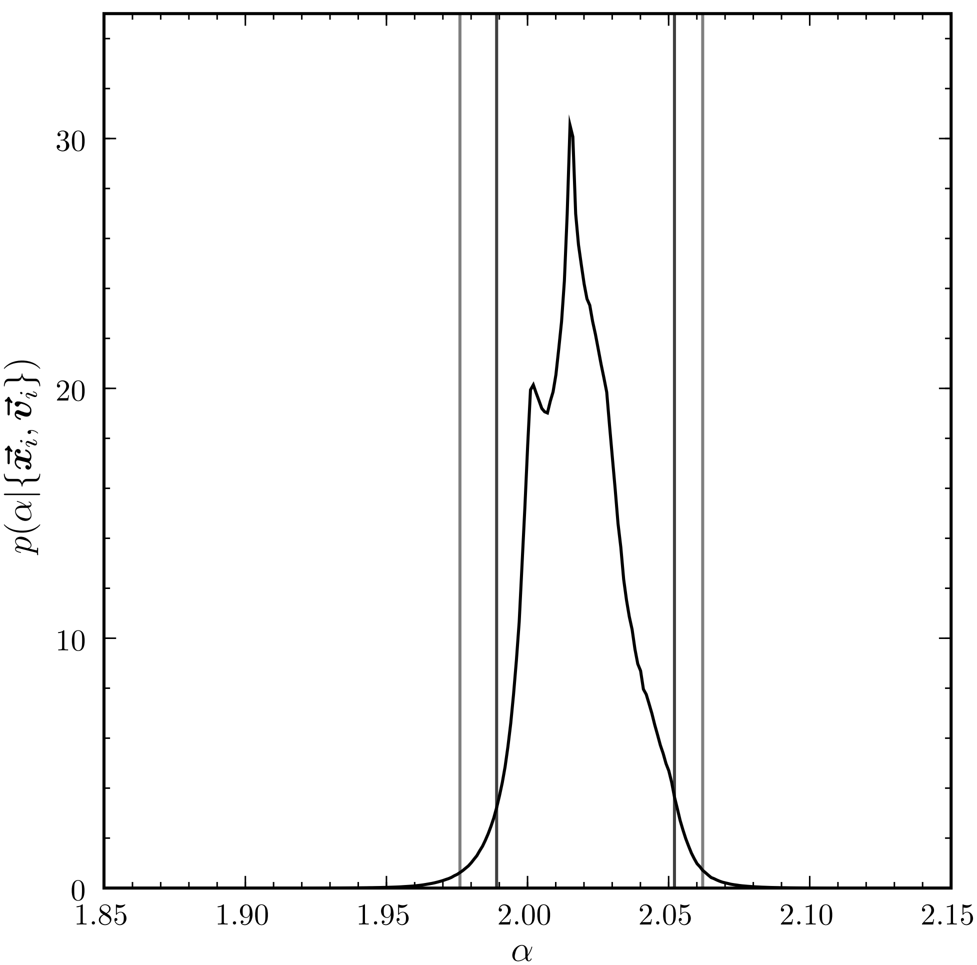
Priors on nusiance distributions

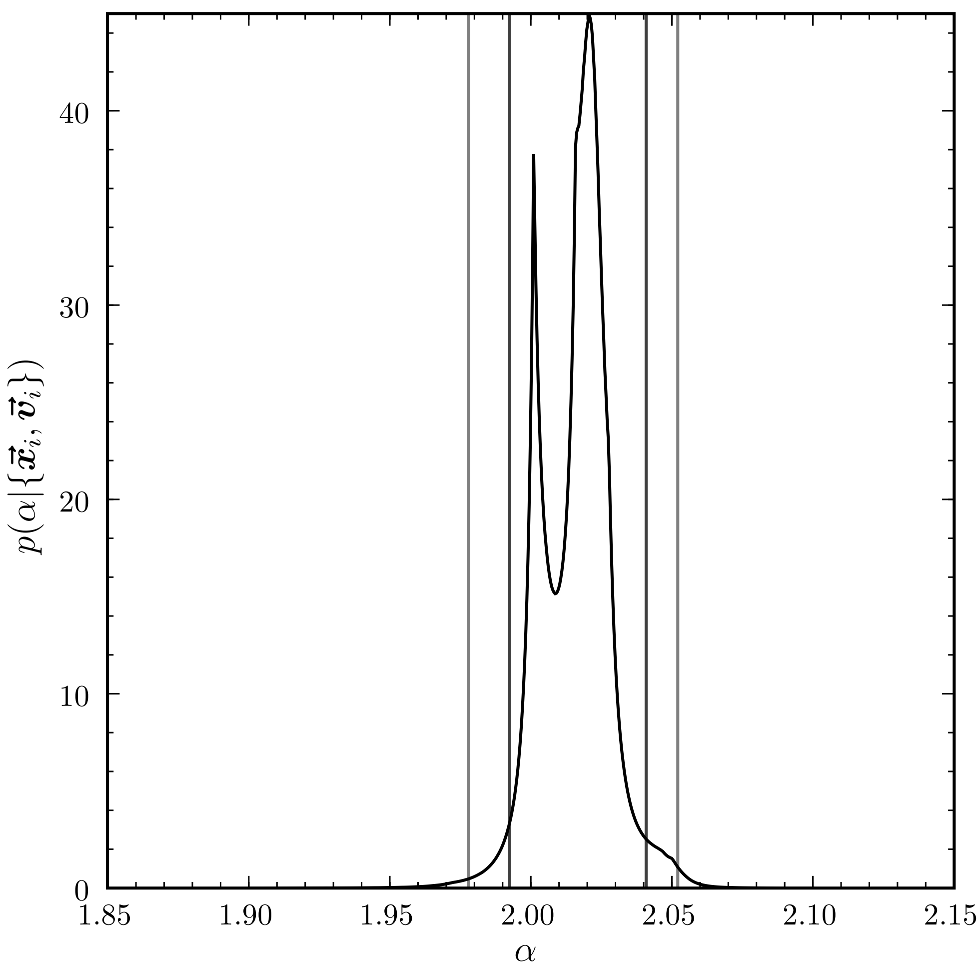


Priors on nusiance distributions



Gravitational exponent





Try it for yourself

Practical exercise:

`http://iainmurray.net/teaching/09mlss/`

Roadmap

- Probability fundamentals
- Inferring a physical parameter
- **Probabilistic models and machine learning**
- Graphical models
- Monte Carlo basics, probabilistic inference in practice

Probabilistic modellers claim. . .

Easy to include rich structure / knowledge

Can handle missing/unlabelled/noisy data

Should be Bayesian when have really limited data:

individual users/entities of a large system

limited trials to set neural net learning rates / hyperparameters

Automatic complexity control ("Occam's razor")

Polyhedral dice

One die chosen uniformly at random:
$D \in \{d_4, d_6, d_8, d_{10}, d_{12}, d_{20}\}$
(subscript gives number of sides)

Rolled 5 times, giving rolls:
$R = [2, 7, 6, 1, 5]$

Q1) What's the most probable die given the data?

Q2) What's $\dfrac{P(d_{10} \,|\, R)}{P(d_{20} \,|\, R)}$?

Image credit: Sabbut from es, `http://commons.wikimedia.org/wiki/File:Dados_4_a_20_caras.jpg`

Discrete model choice

Automatic complexity control means not having to cross-validate lots of choices at all levels of a model. It's great! However, many people are (with reason!) suspicious of using the 'correct' probability theory way to choose whole models.

Marginal likelihood:

$$P(D \,|\, \mathcal{M}) = \int P(D \,|\, \theta, \mathcal{M}) \, p(\theta \,|\, \mathcal{M}) \; \mathrm{d}\theta$$

Cross-validation:

Safer? Look at performance on held-out data.

That's the way to make people believe your model is better

(if you can do it)

Communicating with probabilities

Probability theory tells us how to combine information

Speech recognition

— acoustics combined across time via HMM

— acoustics and language model probabilities combined

However, in practice there a bunch of hacks.

— Hidden Markov Model emitting 'deltas' is hard to justify model

— acoustic model's probabilities not trusted:
probabilities raised to power <1 (log-probs scaled/fudged)

Roadmap

— Probability fundamentals

— Inferring a physical parameter

— Probabilistic models and machine learning

— **Graphical models**

— Monte Carlo basics, probabilistic inference in practice

Directed graphical models

Useful for whiteboard discussions

Split up assumptions: check them

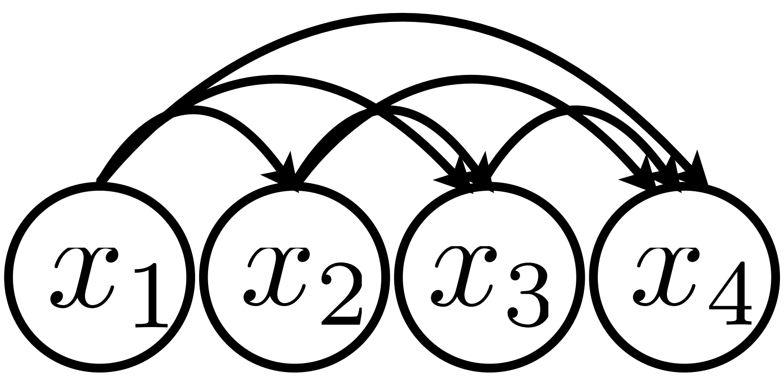
Encode dependencies / conditional independences

Directed graphical models

All distributions follow product rule:

$$P(a, b, c, d) \;=\; P(a)\, P(b \,|\, a)\, P(c \,|\, a, b)\, P(d \,|\, a, b, c)$$

$$P(\mathbf{x}) \;=\; P(x_1) \prod_{d=2}^{D} P(x_d \,|\, \mathbf{x}_{<d})$$

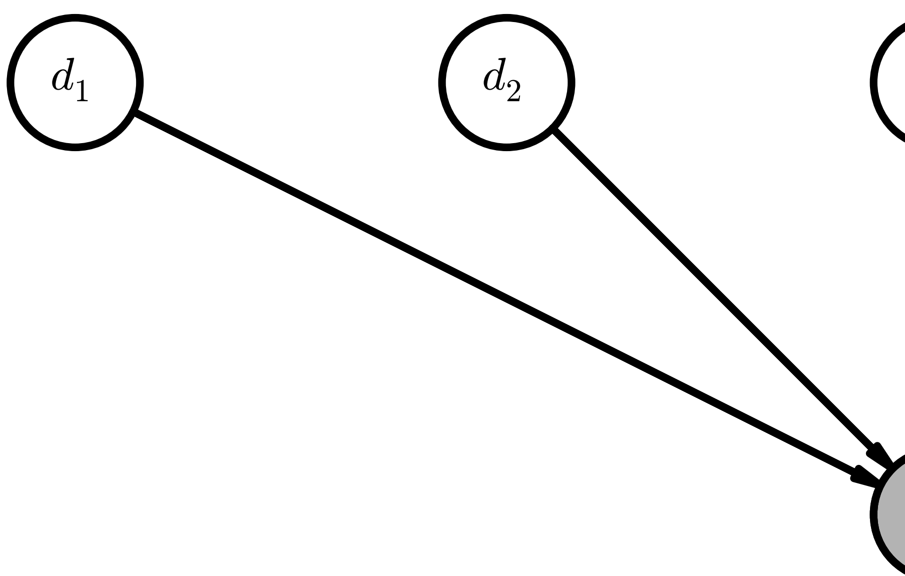


This model is always true!

Removing an edge implies independence structure

“Explaining away”

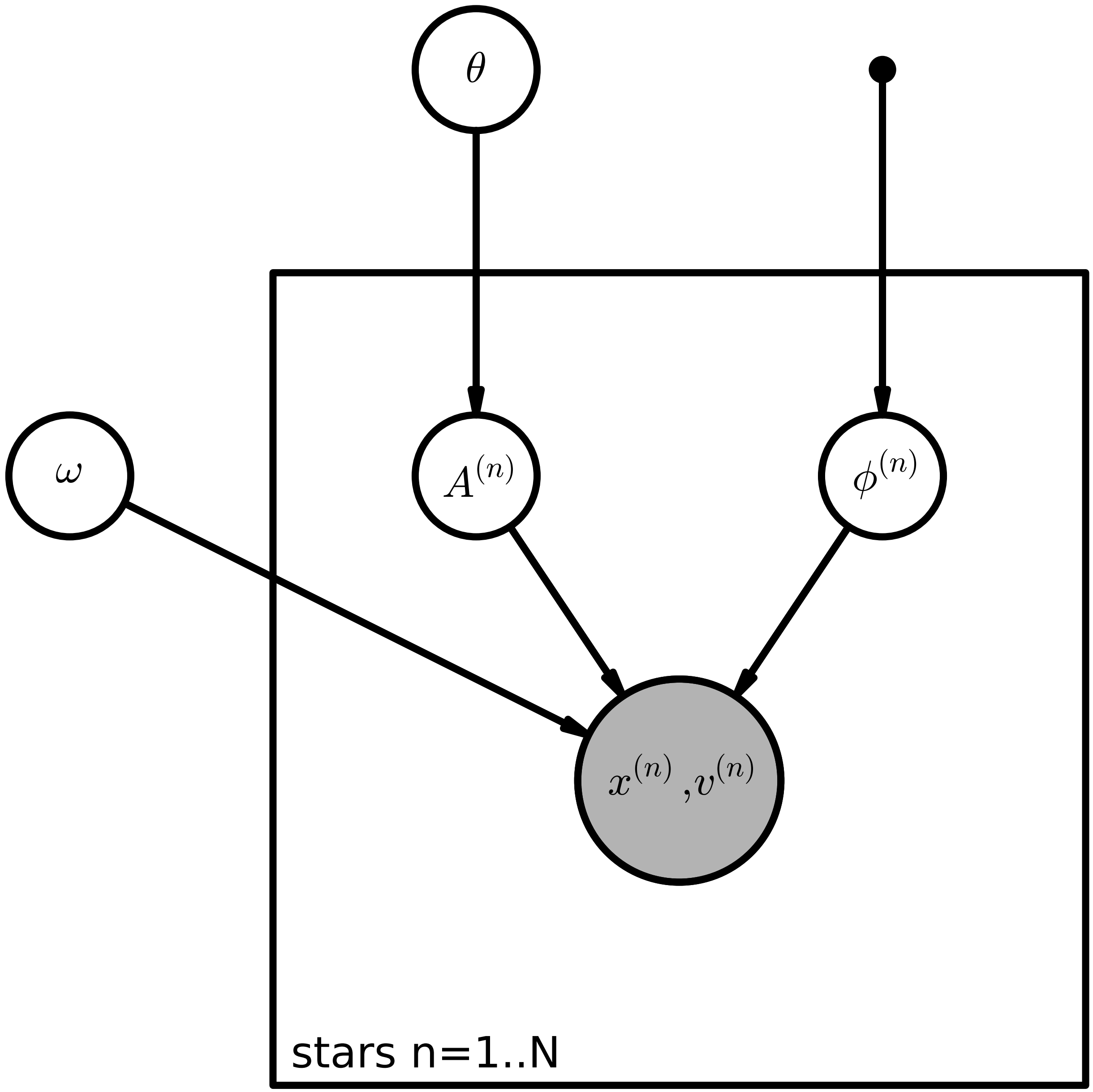
Classic example:



Many diseases d_k
may cause symptom s

Beliefs about parents of observed node become dependent

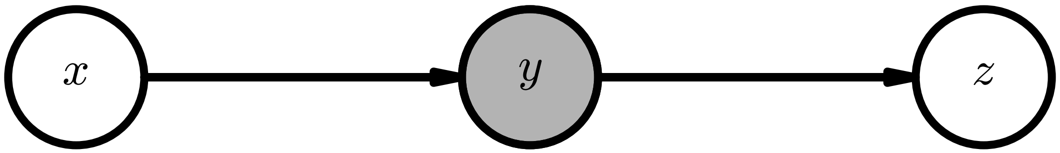
More induced dependencies



Learning about "irrelevant" stuff, helps pin down ω

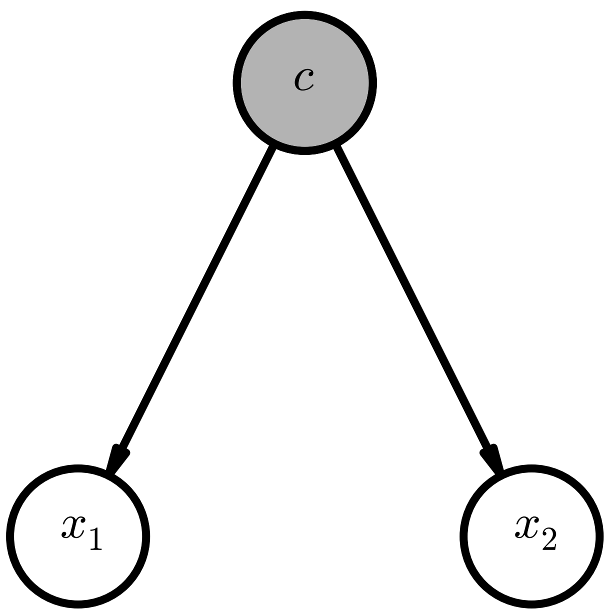
Losing dependencies

Separation from past:



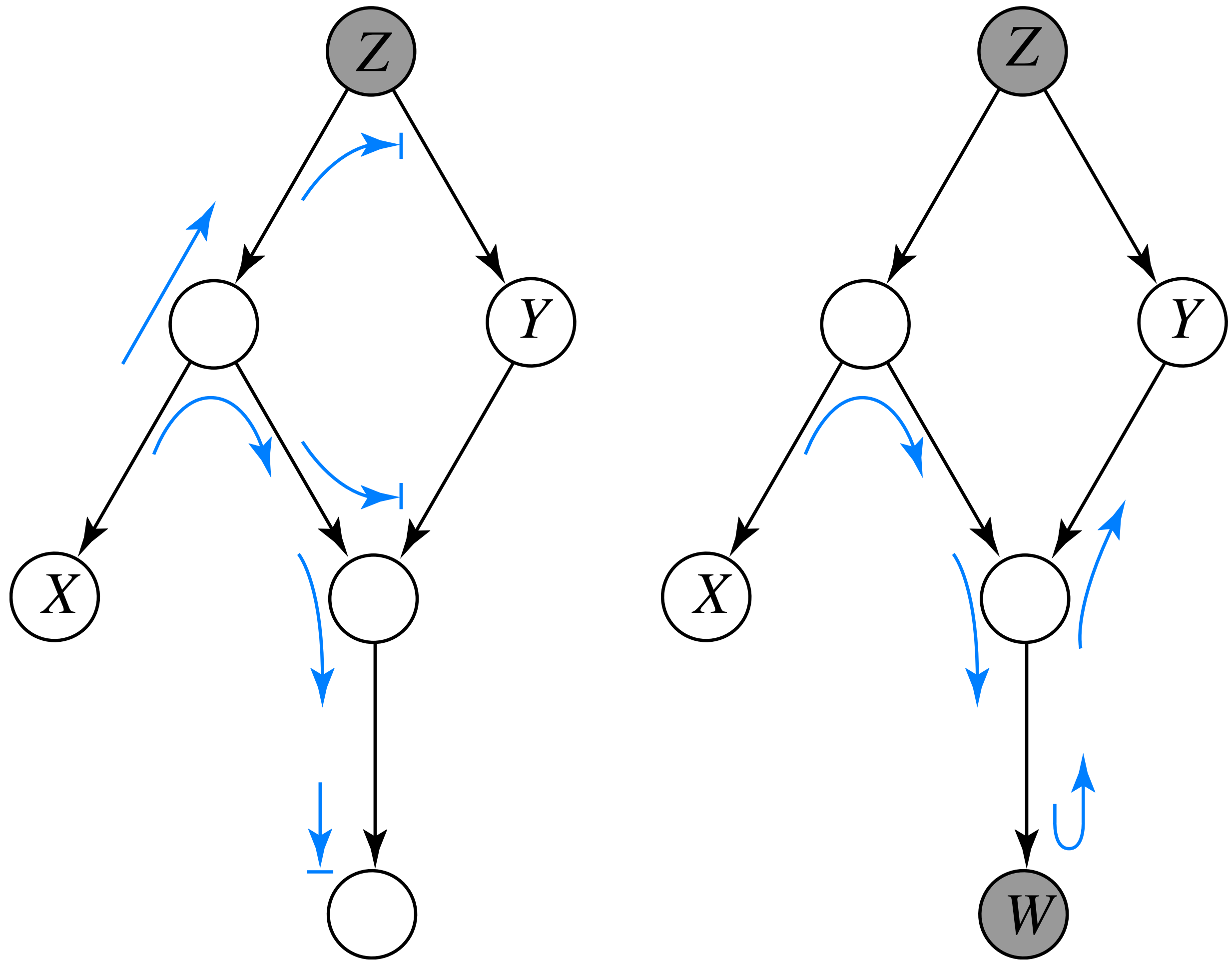
$$x \perp\!\!\!\perp z \mid y$$

The Naive Bayes model:



$$x_1 \perp\!\!\!\perp x_2 \mid c$$

"Bayes Ball" examples

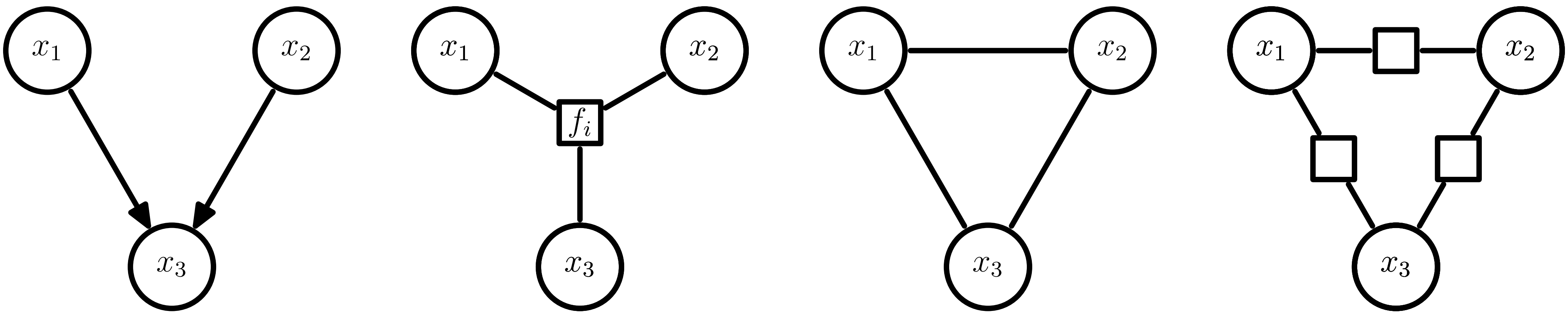


no active paths $\qquad\qquad$ one active path

$X \perp\!\!\!\perp Y \mid Z \qquad\qquad X \not\!\perp\!\!\!\perp Y \mid \{W, Z\}$

Slide 23 lec2, of Mark Paskin's graphical models course
http://ai.stanford.edu/~paskin/gm-short-course/

Undirected graphical models



Different factorizations of the probabilities of everything:

$$P(x_1)\,P(x_2)\,P(x_3\,|\,x_1, x_2)$$

$$\tfrac{1}{Z} f(x_1, x_2, x_3)$$

$$\tfrac{1}{Z} f_a(x_1, x_2)\, f_b(x_1, x_3)\, f_c(x_2, x_3)$$

Undirected is often easier

No need to choose ordering

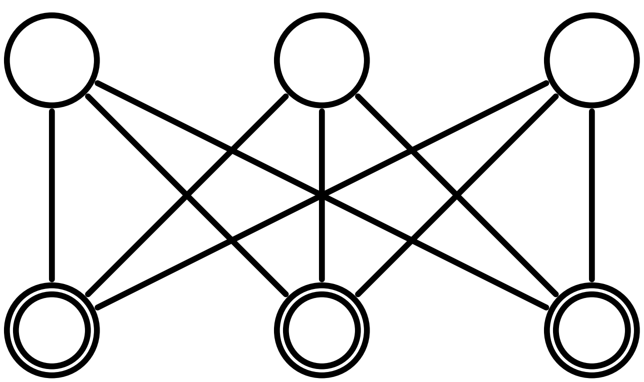
Message-passing-based inference is simpler

Independence rules simpler:

— remove observed vars and edges to them

— conditionally dependent iff path between vars

No "explaining away"



Exponential family models

$$p(\mathbf{x} \,|\, \theta) = \frac{1}{Z(\theta)} g(\mathbf{x}) \exp\left(\sum_k \theta_k \phi_k(\mathbf{x})\right)$$

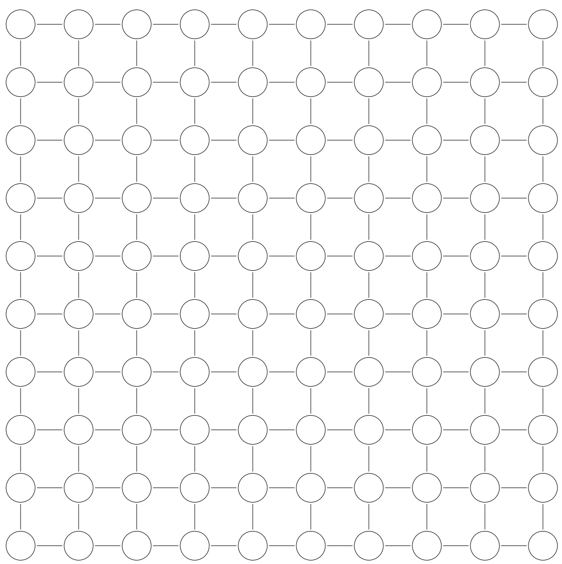
Learning signal:

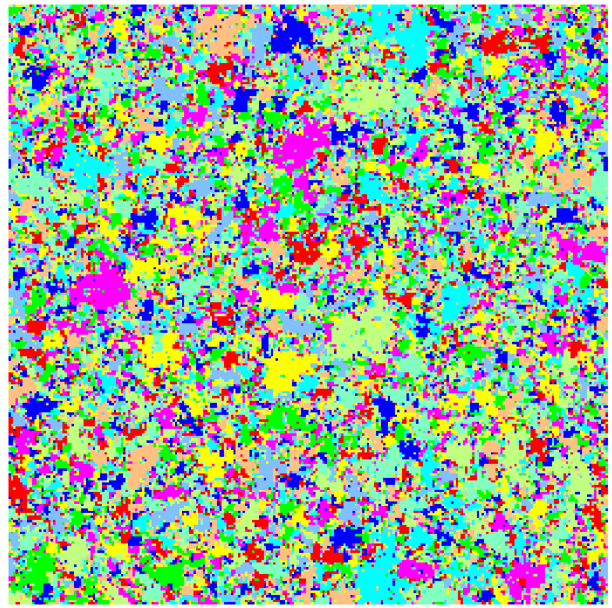
$$\frac{\partial \frac{1}{N} \sum_n \log P(\mathbf{x}^{(n)} \,|\, \theta)}{\partial \theta_k} = \mathbb{E}_{\mathrm{data}}[\phi_k] - \mathbb{E}_{P(\mathbf{x} \,|\, \theta)}[\phi_k]$$

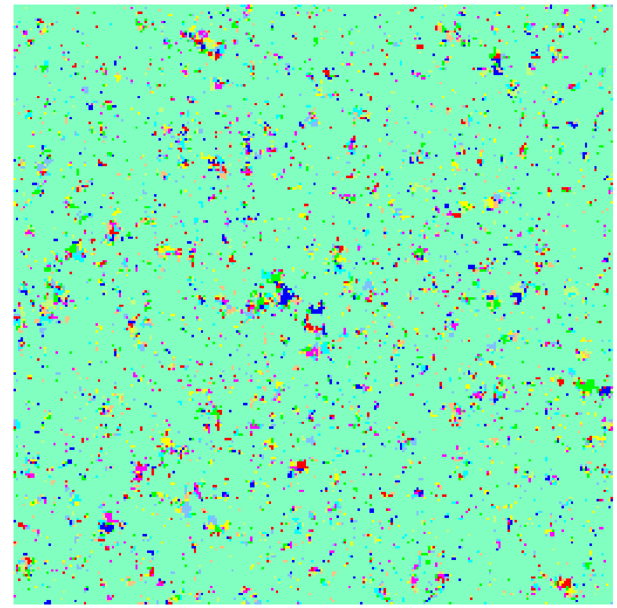
Maximum likelihood matches statistics ϕ

Finds maximum entropy distribution that does so

Undirected downsides







Potts models with 10 colors at the critical coupling

$$P(\mathbf{x} \mid J, h) = \frac{1}{Z(J,h)} \prod_{(i,j)} \phi(x_i, x_j; J) \prod_i \phi(x_i; h)$$

Gaussians are undirected models

$$
\begin{aligned}
P(\mathbf{x} \mid \Sigma, \mu = 0) &\propto \exp\left(-\,{}^{1}\!/\!{}_{2} \sum_{i,j} \Sigma^{-1}_{i,j} x_i x_j \right) \\
&\propto \prod_{i,j} \exp\left(-{}^{1}\!/\!{}_{2} \Sigma^{-1}_{i,j} x_i x_j \right) \\
&= \frac{1}{Z} \prod_{i,j} \phi_{i,j}(x_i, x_j)
\end{aligned}
$$

Latent Gaussian Models:

— tractable Gaussian as undirected backbone

— observation model matches data (e.g., discrete)

PMF FOR NBA BASKETBALL

	BOS	CHA	CLE	MIA	OKC	ORL	PHI	UTA
BOS			93					
CHA						96		
CLE	104							
MIA							104	
OKC								119
ORL		89						
PHI				91				
UTA					111			

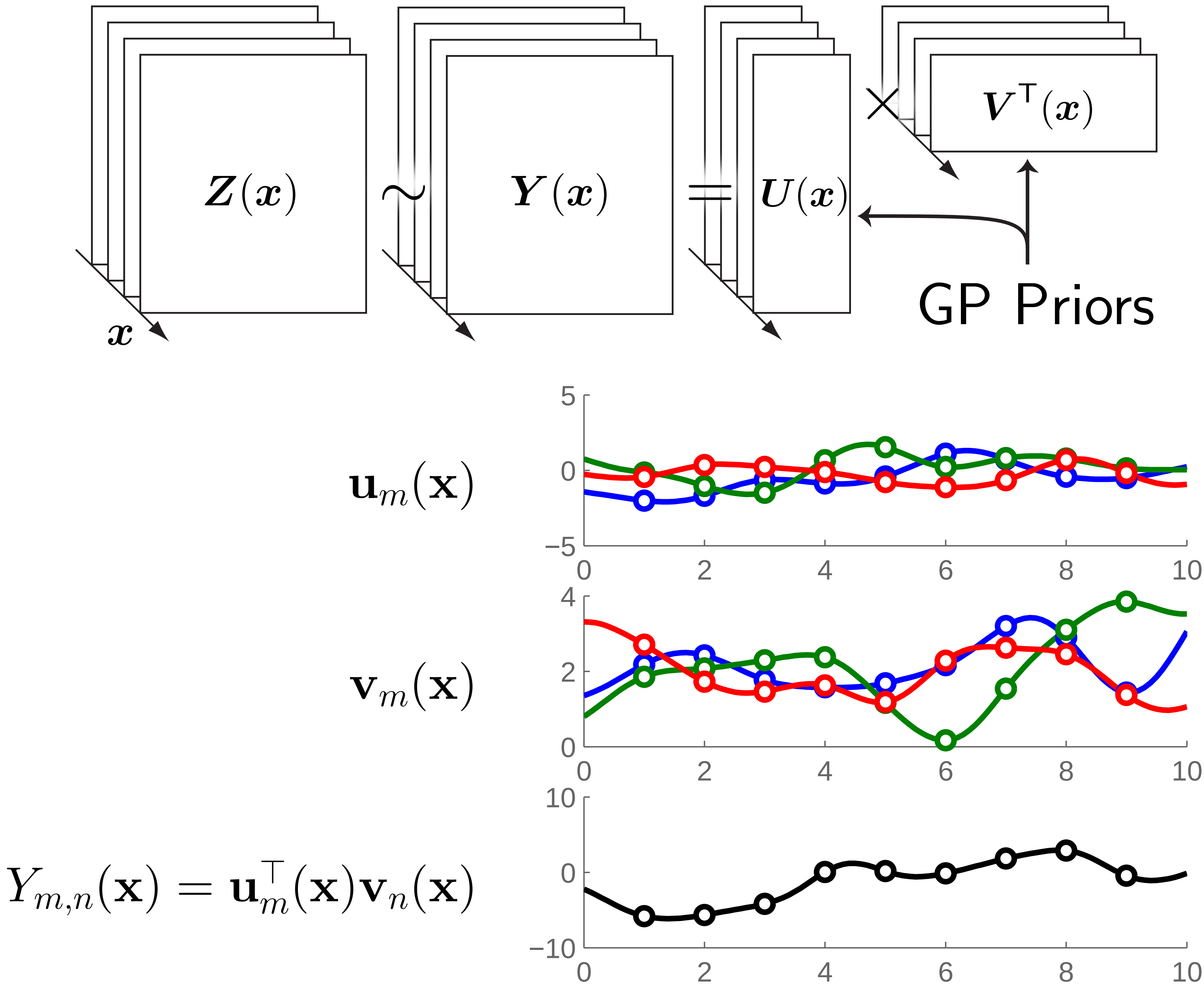
$Z_{m,n}$ = Score of team m against n.

$Z_{n,m}$ = Score of team n against m.

Offense

Defense

$$\underbrace{\boldsymbol{Z}}_{M \times N} \sim \underbrace{\boldsymbol{Y}}_{M \times N} = \underbrace{\boldsymbol{U}}_{M \times K} \times \underbrace{\boldsymbol{V}^{\mathsf{T}}}_{K \times N}$$



$Z(\boldsymbol{x})$
$\sim$
$Y(\boldsymbol{x})$
$=$
$U(\boldsymbol{x})$
$\times$
$V^\top(\boldsymbol{x})$
$\boldsymbol{x}$
GP Priors
$\mathbf{u}_m(\mathbf{x})$
$\mathbf{v}_m(\mathbf{x})$
$Y_{m,n}(\mathbf{x}) = \mathbf{u}_m^\top(\mathbf{x})\mathbf{v}_n(\mathbf{x})$

Roadmap

— Probability fundamentals

— Inferring a physical parameter

— Probabilistic models and machine learning

— Graphical models

— Monte Carlo basics, probabilistic inference in practice