Hands-on hierarchical Bayesian modeling of cosmic populations

Tom Loredo (CCAPS)
With Jessi Cisewski (Yale) (separate slides)

AAS231 Workshop — 2018-01-07

Thanks to our sponsor!

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Overall agenda

- Tom: Quick overviews of basic Bayesian inference & computation; hierarchical Bayes for cosmic populations
- Lab sessions; lunch on your own
- Tom's wrap-up
- Jessi: Approximate Bayesian Computation (ABC) lecture & lab
- We'll stay until about 5:30pm
- Lab materials:
  `https://github.com/tloredo/AAS231-CosmicPopulations`

Menu

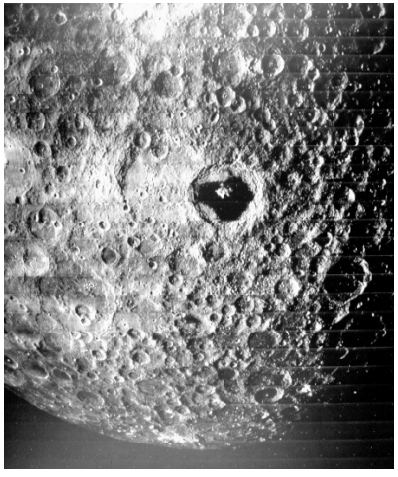
1. What/Why/How of cosmic demographics
2. Basic Bayesian inference (briefly!)
3. Bayesian computation (also briefly!)
4. Hierarchical Bayesian modeling
5. Selection effects: Thinned latent point processes

Menu

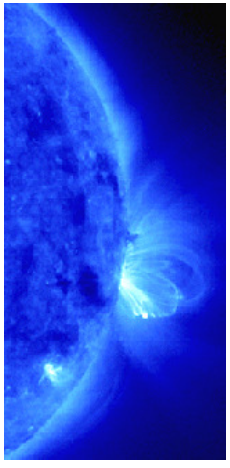
1. **What/Why/How of cosmic demographics**
2. Basic Bayesian inference (briefly!)
3. Bayesian computation (also briefly!)
4. Hierarchical Bayesian modeling
5. Selection effects: Thinned latent point processes

What: We survey everything in the sky!

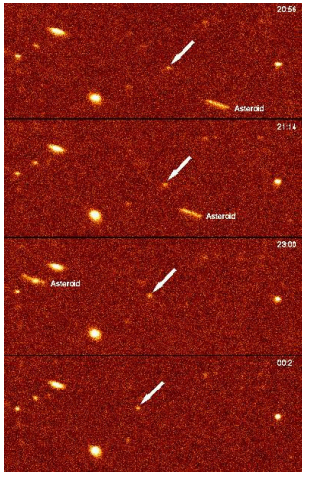
Lunar Craters



Solar Flares

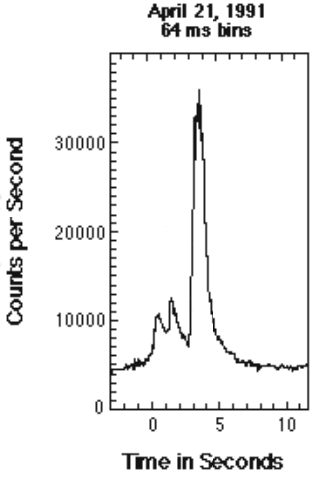


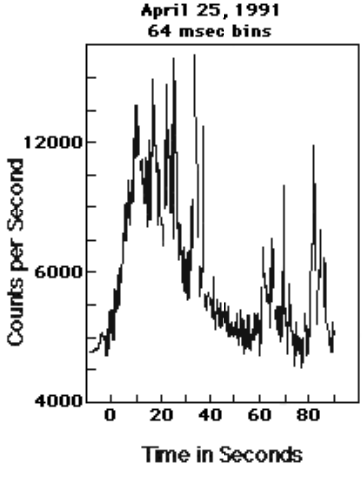
TNOs



Stars & Galaxies

GRBs

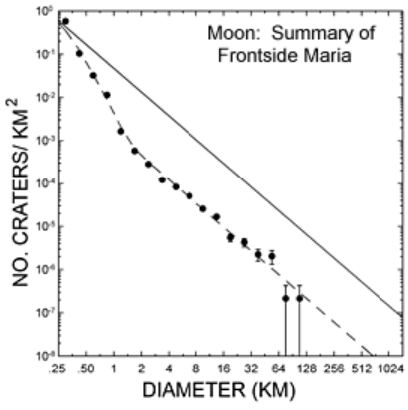




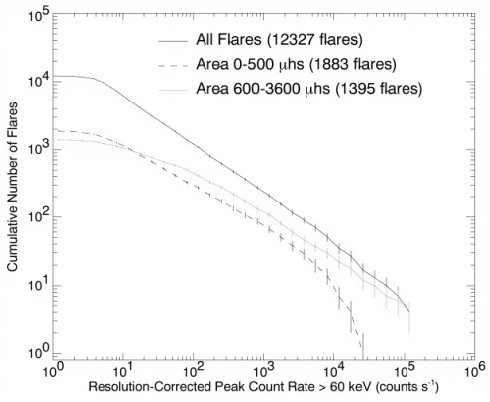
Why: Physics via population distributions

log(N)–log(S) curves, number counts, number-size dist'ns...

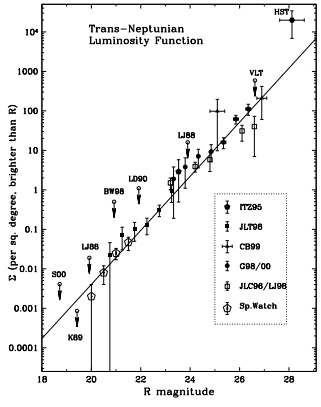
Lunar Craters



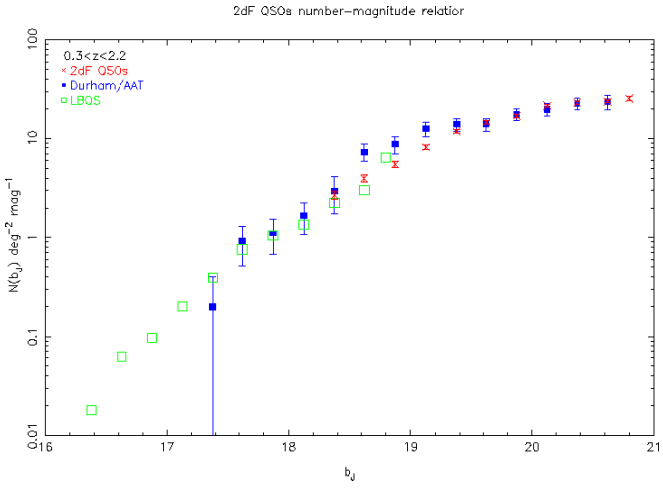
Solar Flares



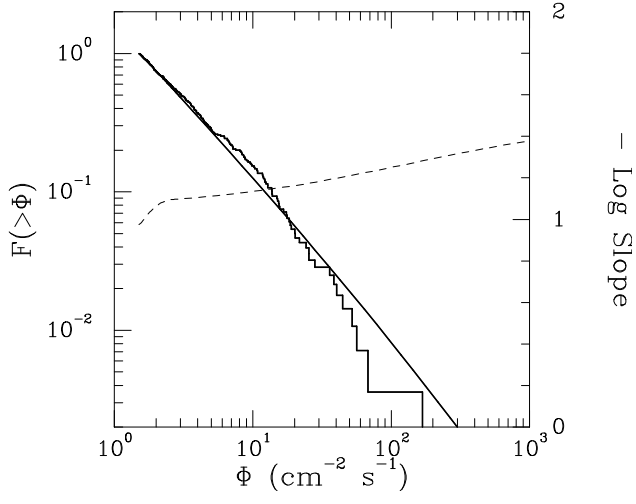
TNOs



Quasars

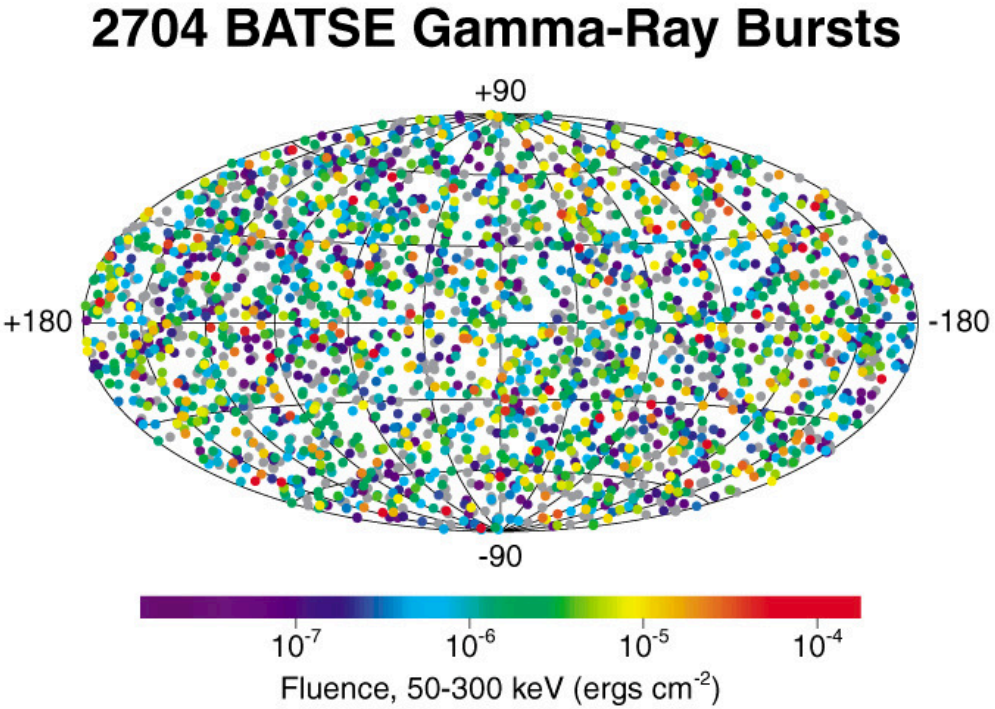


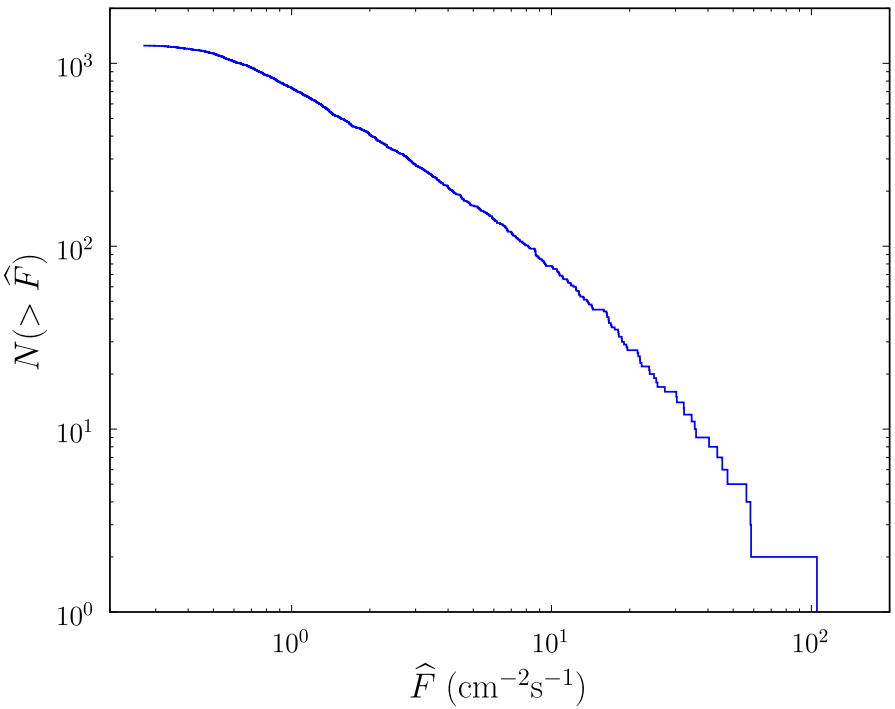
GRBs



Basic observables: Directions and fluxes

Peak fluxes and directions of GRBs from 4B catalog





The Three-Halves (or Five-Halves) Law

Assumptions

- Euclidean space: $F = \frac{L}{4\pi r^2}$
- Homogeneous and isotropic: $n(\vec{r}) = n_0$
- Standard candles: $f_L(L; \vec{r}) = \delta(L - L_0)$

Flux distribution

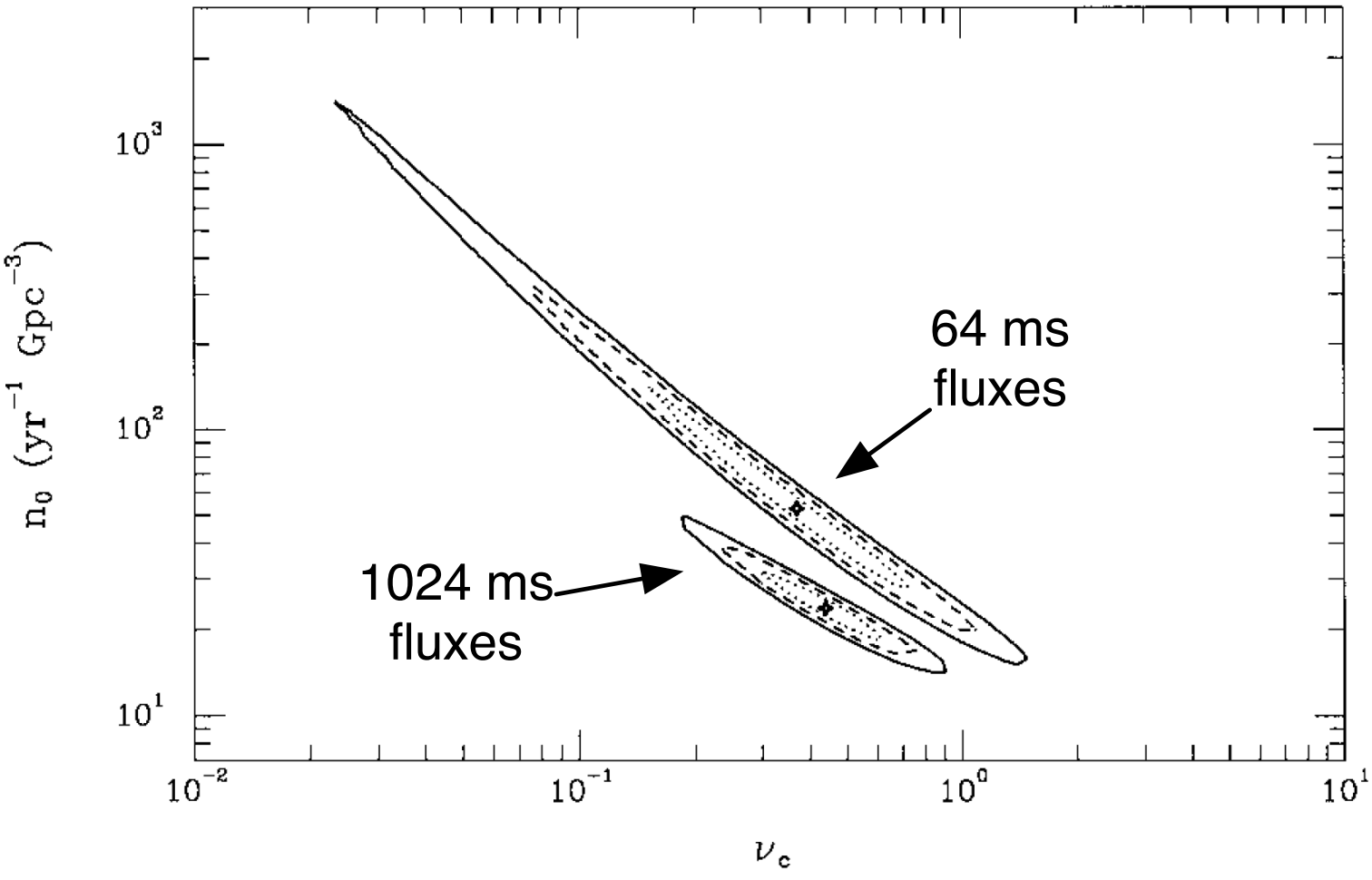
A precise flux measurement $\rightarrow r(F) = \left(\frac{L_0}{4\pi F}\right)^{1/2}$

$$\begin{aligned} \#\text{ with flux } > F &= \#\text{ closer than } r(F) \\ N_>(F) &= \frac{4\pi}{3}[r(F)]^3 n_0 \\ &\propto F^{-3/2} \end{aligned}$$

Differential distribution (surf. dens. per unit flux & steradian):

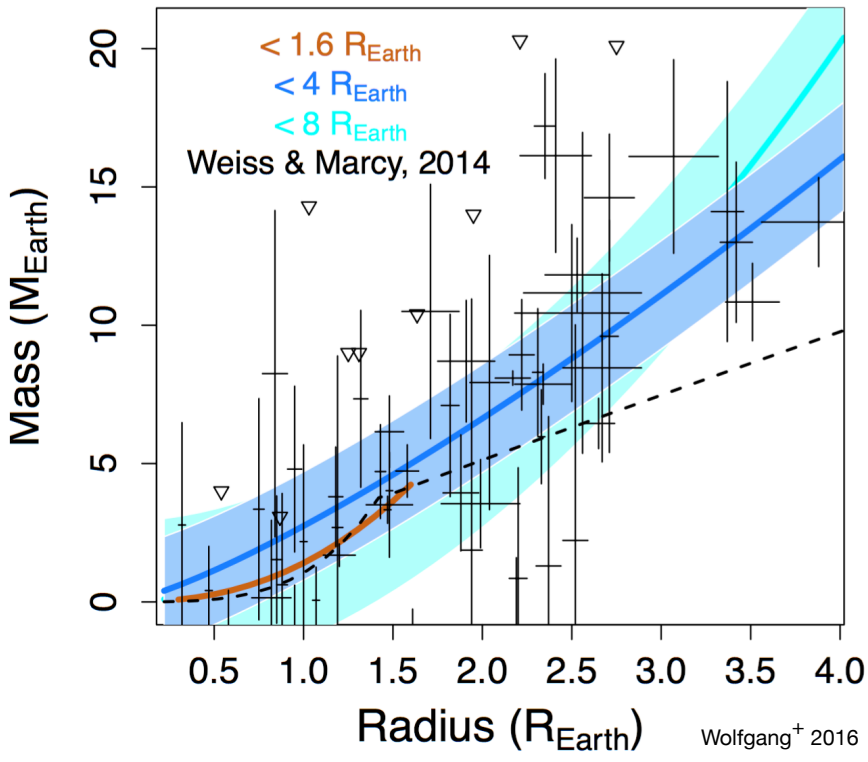
$$\Sigma(F) = -\frac{1}{4\pi}\frac{dN_>}{dF} \propto F^{-5/2}$$

Marginal posterior for standard-candle luminosity, rate



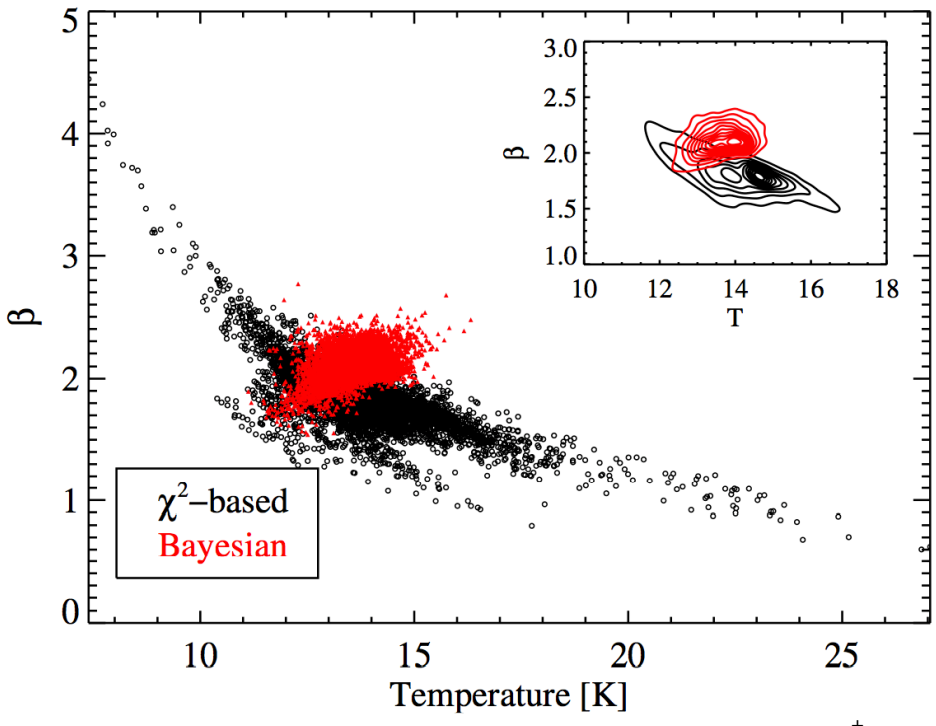
Joint and conditional distributions

Exoplanet mass-radius relation



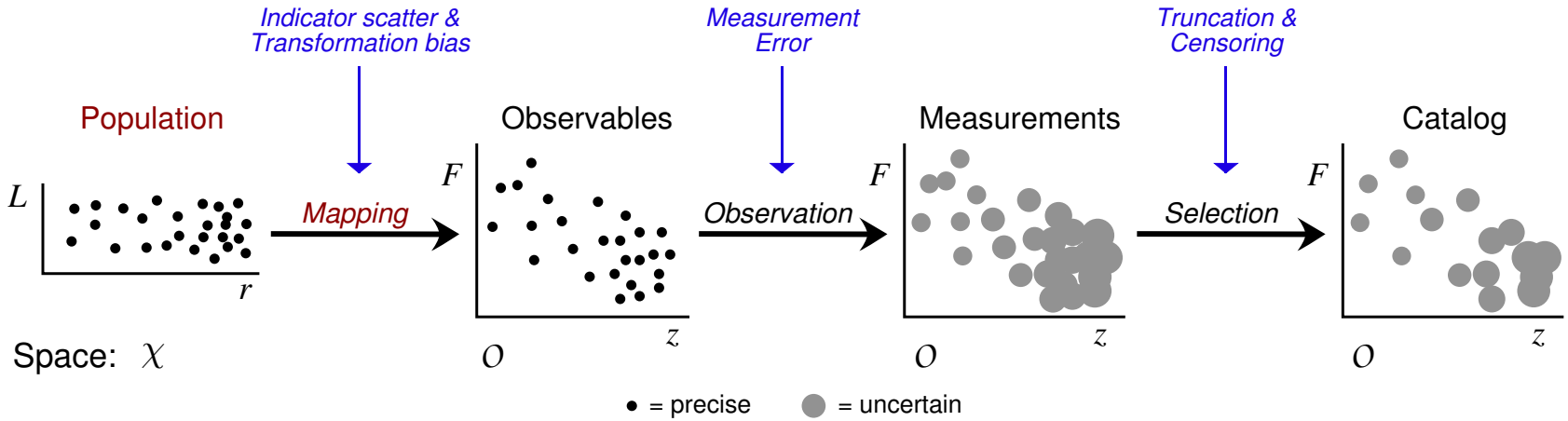
Wolfgang^{+} 2016

Dust cloud properties



Kelly^{+} 2016

Observing and modeling cosmic populations

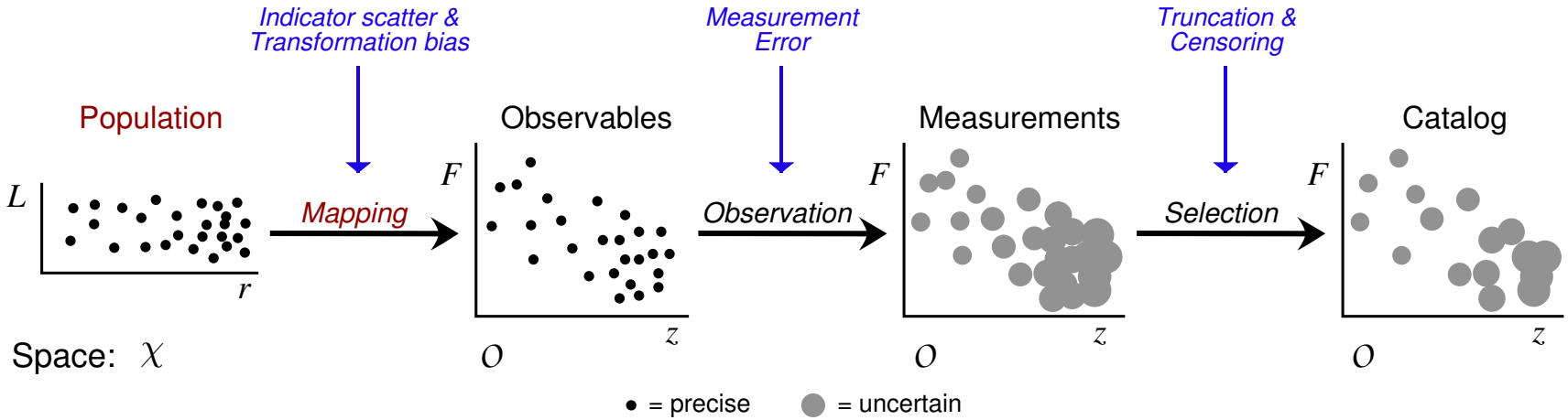


Science goals

- *Density estimation*: Infer the distribution of source characteristics, $p(\chi)$
- *Regression/Cond'l density estimation*: Infer relationships between different characteristics
- *Map regression*: Infer parameters defining the mapping from characteristics to observables

Notably, seeking improved point estimates of source characteristics is seldom a goal in astronomy

"Un-surveying"



Inverse methods

- Try to "correct" or "debias" data via adjustments/weights
- Focus on moments & empirical dist'n function (EDF)

Forward modeling methods — our focus!

- Try to predict data by applying survey process to model
- Focus on likelihood—*predictive probability* for survey data

(Analogous to "design-based" vs. "model-based" methods in survey sampling lit.)

Menu

1. What/Why/How of cosmic demographics
2. **Basic Bayesian inference (briefly!)**
3. Bayesian computation (also briefly!)
4. Hierarchical Bayesian modeling
5. Selection effects: Thinned latent point processes

Frequentist vs. Bayesian hypothesis appraisal

"The data D_{obs} support conclusion C . . . "

Frequentist appraisal

> *"C was selected with a procedure that's right 95% of the time over a set $\{D_{hyp}\}$ that includes D_{obs}."*
>
> Probabilities are properties of *procedures*, not of particular results

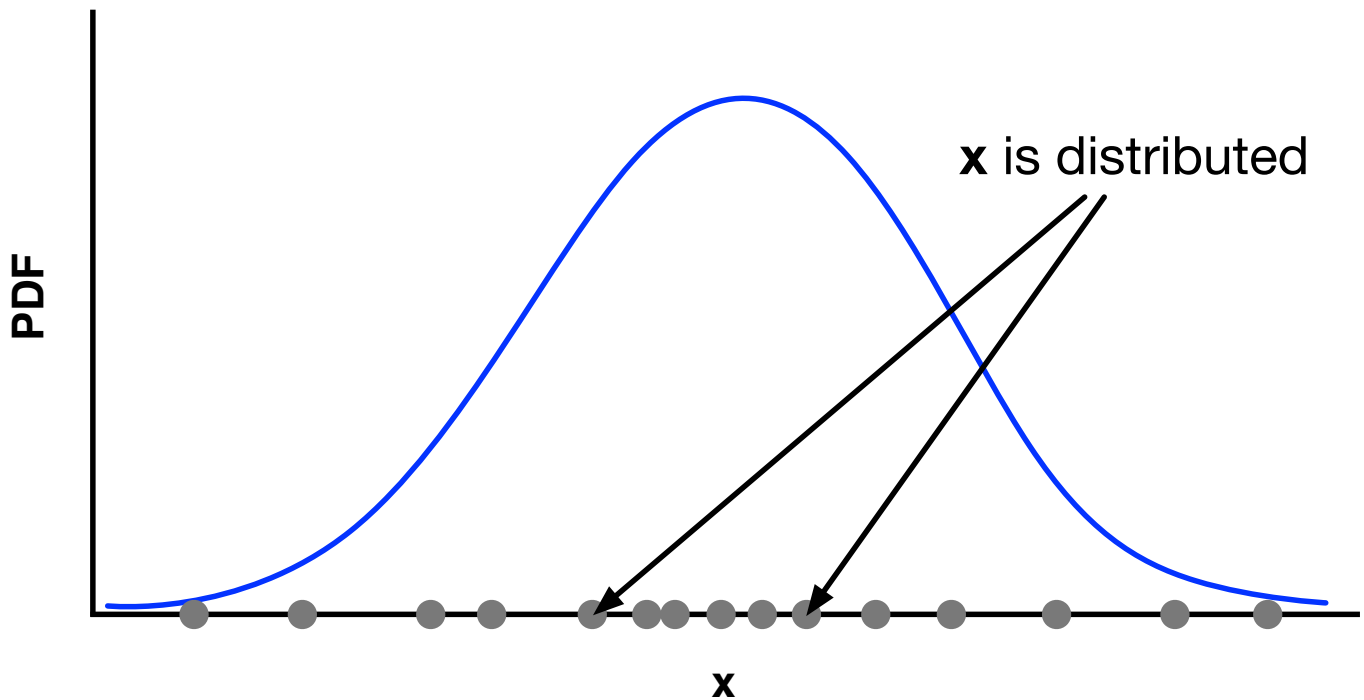
Bayesian appraisal

> *"The strength of the chain of reasoning from the model and D_{obs} to C is 0.95, on a scale where 1= certainty."*
>
> Probabilities are associated with *specific, observed data*
>
> Long-run performance must be separately evaluated (and is often good by frequentist criteria)

Interpreting PDFs

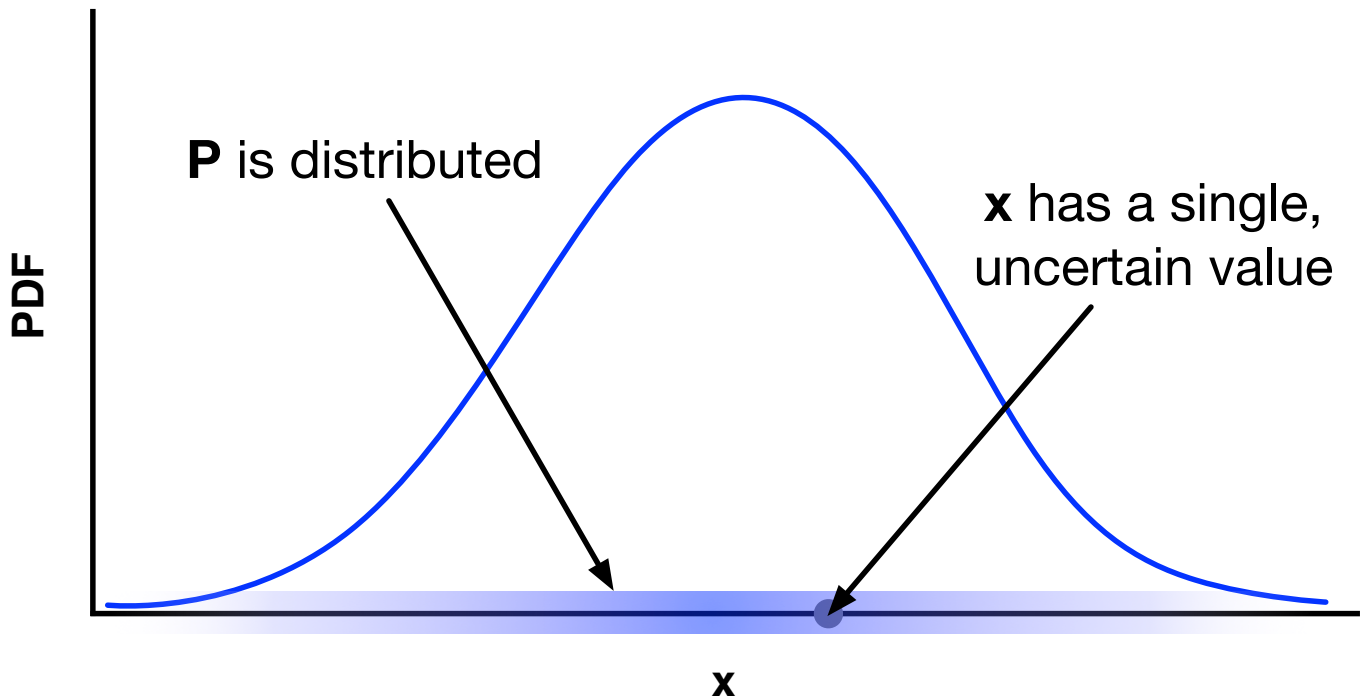
Frequentist

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* in a sequence of trials. $p(x)$ describes how the *values of* x would be distributed among infinitely many trials:



Bayesian

Probability *quantifies uncertainty* in an inductive inference. $p(x)$ describes how *probability* is distributed over the possible values x might have taken in the single case before us:



The weather forecaster

Joint Frequencies of Actual & Predicted Weather

	Actual		
Prediction	Snow	Sun	
Snow	**1/4**	**1/2**	3/4
Sun	**0**	**1/4**	1/4
	1/4	3/4	

Forecaster is right only 50% of the time

Observer notes a prediction of 'Sun' *every day* would be right 75% of the time, and applies for the forecaster's job

Should that observer get the job?

	Actual	
Prediction	Snow	Sun
Snow	1/4	1/2
Sun	0	1/4

Forecaster: You'll never be in an unpredicted snow storm

Observer: You'll be in an unpredicted storm 1 day out of 4

Bayesian viewpoint

The value of an inference typically lies in its usefulness for the case at hand

Long run performance is not an adequate criterion for assessing the usefulness of inferences for the case at hand

When long run performance is important, it needs to be separately evaluated

Probability & frequency: Bernoulli and Bayes

Frequency from probability

Bernoulli's *law of large numbers* (1713): In repeated IID binary trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \to \alpha \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be interpreted as a frequency distribution

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" (1763) $\to$ First use of *Bayes's theorem*:

Probability for success in next trial of IID sequence:

$$\mathbb{E}(\alpha) \to \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be estimated from a frequency distribution

Propositions and arguments

Proposition: Statement that may be true or false

Argument: Assertion that an *hypothesized conclusion*, H, follows from *premises*, $\mathcal{P} = \{A, B, C, \ldots\}$ (take "," = "and")

Notation:

$H \vert \mathcal{P}$:	Premises $\mathcal{P}$ imply H
	H may be deduced from $\mathcal{P}$
	H follows from $\mathcal{P}$
	H is true given that $\mathcal{P}$ is true

Arguments are (compound) propositions.

Central role of arguments $\rightarrow$ special terminology for true/false:

- A true argument is *valid*
- A false argument is *invalid* or *fallacious*

Representing induction with $[0,1]$ calculus

$P(H|\mathcal{P}) \equiv$ strength of argument $H|\mathcal{P}$

$$\begin{aligned} P &= 1 \rightarrow \text{Argument is } \textit{deductively valid} \\ &= 0 \rightarrow \text{Premises imply } \overline{H} \\ &\in (0,1) \rightarrow \text{Degree of deducibility} \end{aligned}$$

Mathematical model for induction

$$\text{'AND' (product rule):} \quad \begin{aligned} P(A,B|\mathcal{P}) &= P(A|\mathcal{P})\,P(B|A,\mathcal{P}) \\ &= P(B|\mathcal{P})\,P(A|B,\mathcal{P}) \end{aligned}$$

$$\text{'OR' (sum rule):} \quad \begin{aligned} P(A \vee B|\mathcal{P}) &= P(A|\mathcal{P}) + P(B|\mathcal{P}) \\ &\quad -P(A,B|\mathcal{P}) \end{aligned}$$

$$\text{'NOT':} \quad P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$$

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i|\ldots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$\mathcal{C} \equiv$ context, initial set of premises

'AND' (product rule): $P(H_i, D_{\rm obs}|\mathcal{C}) = P(H_i|\mathcal{C})\,P(D_{\rm obs}|H_i, \mathcal{C})$
$= P(D_{\rm obs}|\mathcal{C})\,P(H_i|D_{\rm obs}, \mathcal{C})$

'OR' (sum rule): $P(H_1 \vee H_2|\mathcal{C}) = P(H_1|\mathcal{C}) + P(H_2|\mathcal{C})$
$-P(H_1, H_2|\mathcal{C})$

'NOT': $P(\overline{H_i}|\mathcal{C}) = 1 - P(H_i|\mathcal{C})$

Three Important Theorems

Bayes's Theorem (BT)

Consider $P(H_i, D_{\text{obs}}|\mathcal{C})$ using the product rule:

$$\begin{aligned} P(H_i, D_{\text{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\text{obs}}|H_i, \mathcal{C}) \\ &= P(D_{\text{obs}}|\mathcal{C})\, P(H_i|D_{\text{obs}}, \mathcal{C}) \end{aligned}$$

Solve for the *posterior probability* (expands the premises!):

$$P(H_i|D_{\text{obs}}, \mathcal{C}) = P(H_i|\mathcal{C})\, \frac{P(D_{\text{obs}}|H_i, \mathcal{C})}{P(D_{\text{obs}}|\mathcal{C})}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

$$\textit{posterior} \propto \textit{prior} \times \textit{likelihood}$$

$$\text{norm. const. } P(D_{\text{obs}}|\mathcal{C}) = \textit{prior predictive}$$

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ ($\mathcal{C}$ asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i|\mathcal{C}) &= \sum_i P(B_i|A, \mathcal{C})P(A|\mathcal{C}) = P(A|\mathcal{C}) \\ &= \sum_i P(B_i|\mathcal{C})P(A|B_i, \mathcal{C})\end{aligned}$$

If we do not see how to get $P(A|\mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|\mathcal{C}) = \sum_i P(B_i|\mathcal{C})P(A|B_i, \mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A|\mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A|\mathcal{C}) = \sum_i P(A, B_i|\mathcal{C})$$

Example: Take $\mathcal{P} = \mathcal{C}$, $A = D_{\text{obs}}$, $B_i = H_i$; then

$$\begin{aligned} P(D_{\text{obs}}|\mathcal{C}) &= \sum_i P(D_{\text{obs}}, H_i|\mathcal{C}) \\ &= \sum_i P(H_i|\mathcal{C}) P(D_{\text{obs}}|H_i, \mathcal{C}) \end{aligned}$$

prior predictive for D_{obs} = Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Normalization

For *exclusive, exhaustive* H_i,

$$\sum_i P(H_i|\cdots) = 1$$

Inference With Parametric Models

Models M_i ($i = 1$ to N), each with parameters θ_i, each imply a *sampling dist'n* (conditional predictive dist'n for possible data):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the *observed* data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty).

Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript: $D = D_{obs}$.

Classes of Problems

Single-model inference

Premise = choice of single model (specific i)
Parameter estimation: What can we say about θ_i or $f(\theta_i)$?
Prediction: What can we say about future data D'?

Multi-model inference

Premise = $\{M_i\}$
Model comparison/choice: What can we say about i?
Model averaging:

- *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
  What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D', accounting for model uncertainty?

Model checking

Premise = $M_1 \vee$ "all" alternatives
Is M_1 adequate? (predictive tests, calibration, robustness)

Parameter Estimation

Problem statement

I = Model M with parameters θ (+ any add'l info)

H_i = statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"

Probability for any such statement can be found using a *probability density function* (PDF) for θ:

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \cdots) &= f(\theta) d\theta \\ &= p(\theta | \cdots) d\theta \end{aligned}$$

Posterior probability density

$$p(\theta | D, M) = \frac{p(\theta | M)\ \mathcal{L}(\theta)}{\int d\theta\ p(\theta | M)\ \mathcal{L}(\theta)}$$

Summaries of posterior

- "Best fit" values:
  - *Mode*, $\hat{\theta}$, maximizes $p(\theta|D, M)$
  - *Posterior mean*, $\langle\theta\rangle = \int d\theta\, \theta\, p(\theta|D, M)$
- Uncertainties:
  - *Credible region* Δ of probability C:
    $C = P(\theta \in \Delta|D, M) = \int_\Delta d\theta\, p(\theta|D, M)$
    Highest Posterior Density (HPD) region has $p(\theta|D, M)$ higher inside than outside
  - Posterior standard deviation, variance, covariances
- Marginal distributions
  - Interesting parameters ϕ, nuisance parameters η
  - *Marginal dist'n* for ϕ: $p(\phi|D, M) = \int d\eta\, p(\phi, \eta|D, M)$

Binary Outcomes: Parameter Estimation

M = Existence of two outcomes, S and F; for each case or trial, the probability for S is α; for F it is $(1 - \alpha)$

H_i = Statements about α, the probability for success on the next trial $\rightarrow$ seek $p(\alpha|D, M)$

D = Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood:

$$\begin{aligned} p(D|\alpha, M) &= p(\text{failure}|\alpha, M) \times p(\text{failure}|\alpha, M) \times \cdots \\ &= \alpha^n (1 - \alpha)^{N-n} \\ &= \mathcal{L}(\alpha) \end{aligned}$$

Prior

Starting with no information about α beyond its definition, use as an "uninformative" prior $p(\alpha|M) = 1$. Justifications:

- Intuition: Don't prefer any α interval to any other of same size
- Bayes's justification: "Ignorance" means that before doing the N trials, we have no preference for how many will be successes:

$$P(n \text{ success}|M) = \frac{1}{N+1} \qquad \rightarrow \qquad p(\alpha|M) = 1$$

Consider this a *convention*—an assumption added to M to make the problem well posed.

Prior Predictive

$$\begin{aligned} p(D|M) &= \int d\alpha\, \alpha^n (1-\alpha)^{N-n} \\ &= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!} \end{aligned}$$

A *Beta integral*, $B(a,b) \equiv \int dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Posterior

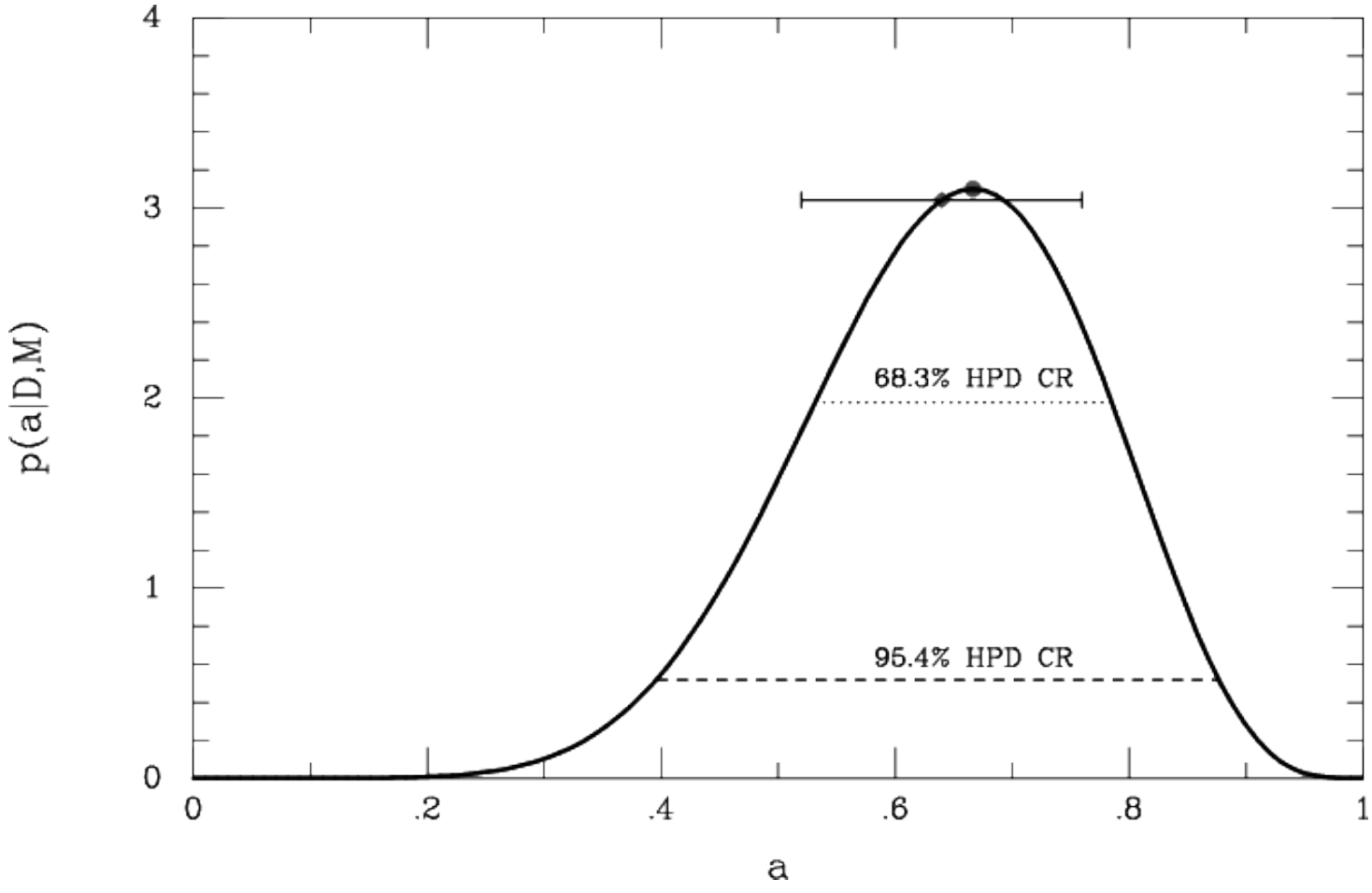
$$p(\alpha|D,M) \quad = \quad \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

A *Beta distribution*. Summaries:

- Best-fit: $\hat{\alpha} = \frac{n}{N} = 2/3$; $\langle\alpha\rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: $\sigma_\alpha = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$
  Find credible regions numerically, or with incomplete beta function

Note that the posterior depends on the data only through n, not the N binary numbers describing the sequence.

n is a (minimal) *sufficient statistic*.


4
3
2
1
0
p(a|D,M)
0
.2
.4
.6
.8
1
a
68.3% HPD CR
95.4% HPD CR

Binary Outcomes: Binomial Distribution

Suppose $D = n$ (number of heads in N trials), rather than the actual sequence. What is $p(\alpha|n, M)$?

Likelihood

Let $\mathcal{S}$ = a sequence of flips with n heads.

$$p(n|\alpha, M) = \sum_{\mathcal{S}} p(S|\alpha, M)\; p(n|\mathcal{S}, \alpha, M)$$

$p(S|\alpha, M) = \alpha^n (1-\alpha)^{N-n}$

$p(n|\mathcal{S}, \alpha, M) = [\![\#\text{ successes} = n]\!]$

$$= \alpha^n (1-\alpha)^{N-n} C_{n,N}$$

$C_{n,N} = \#$ of sequences of length N with n heads.

$$\rightarrow p(n|\alpha, M) = \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

The *binomial distribution* for n given α, N.

Posterior

$$p(\alpha|n,M) = \frac{\frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}}{p(n|M)}$$

$$\begin{aligned} p(n|M) &= \frac{N!}{n!(N-n)!}\int d\alpha\, \alpha^n(1-\alpha)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

$$\rightarrow p(\alpha|n,M) = \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Same result as when data specified the actual sequence

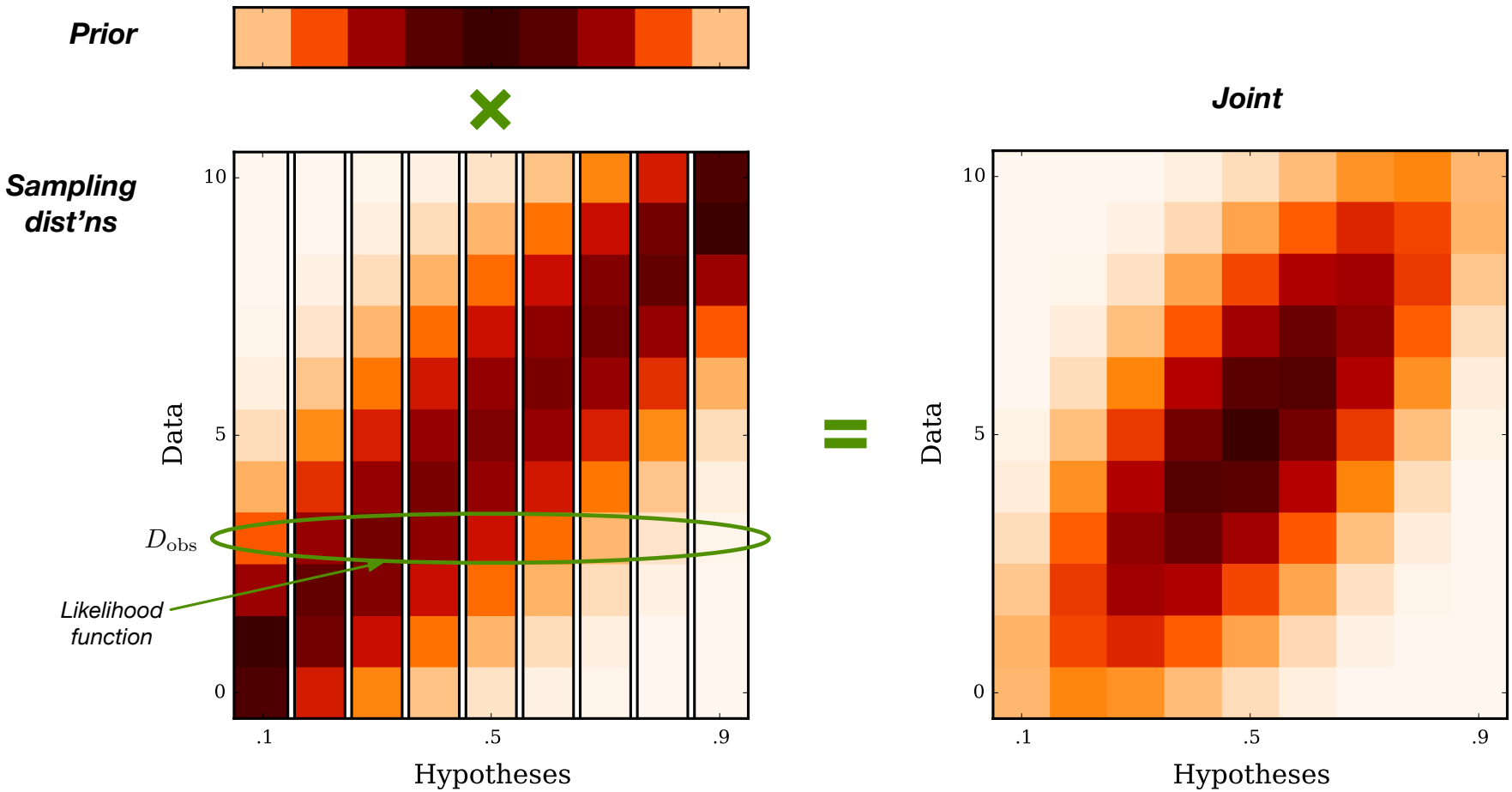
An example of the *likelihood principle*: Problems with proportional likelihood functions should produce the same inferences.

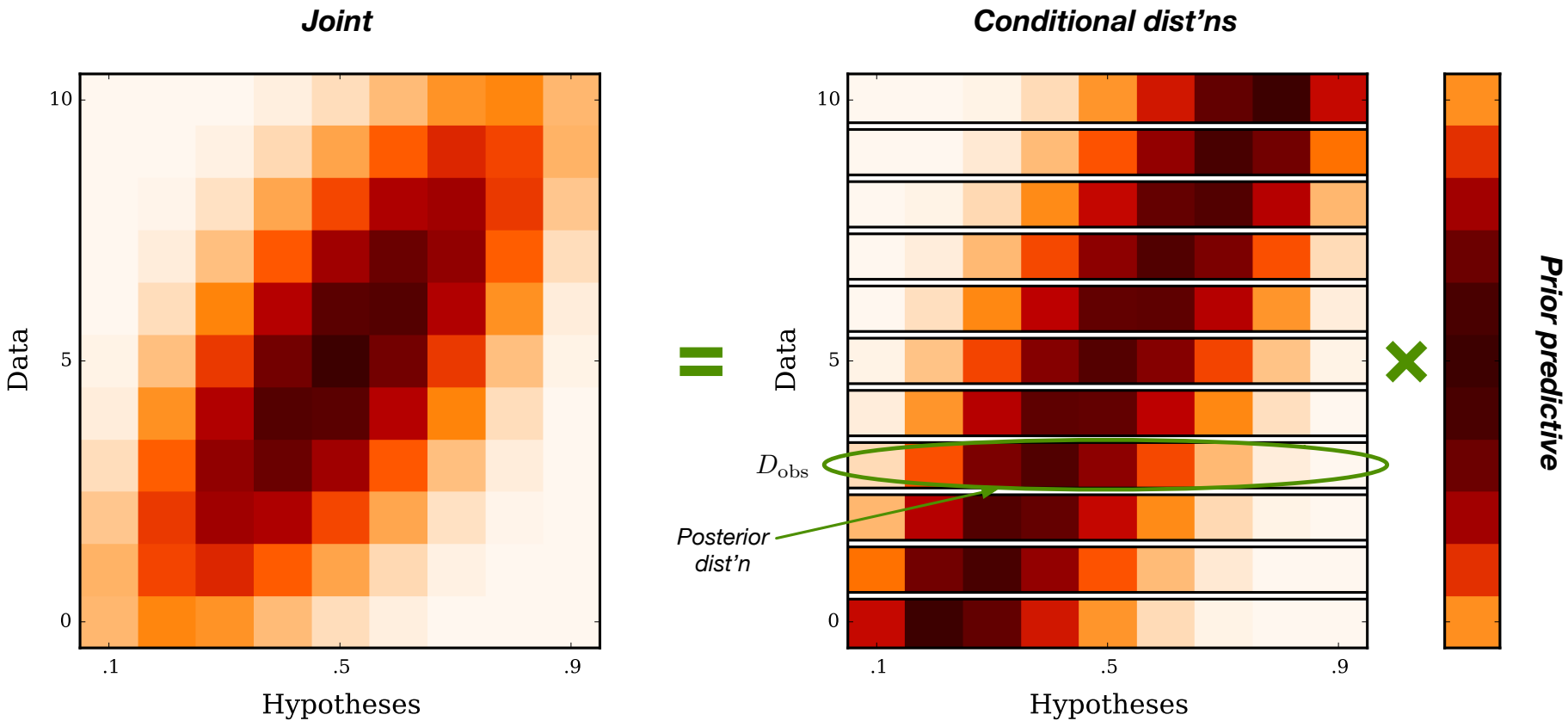
In Bayesian inference, it's a deduction, not a principle.

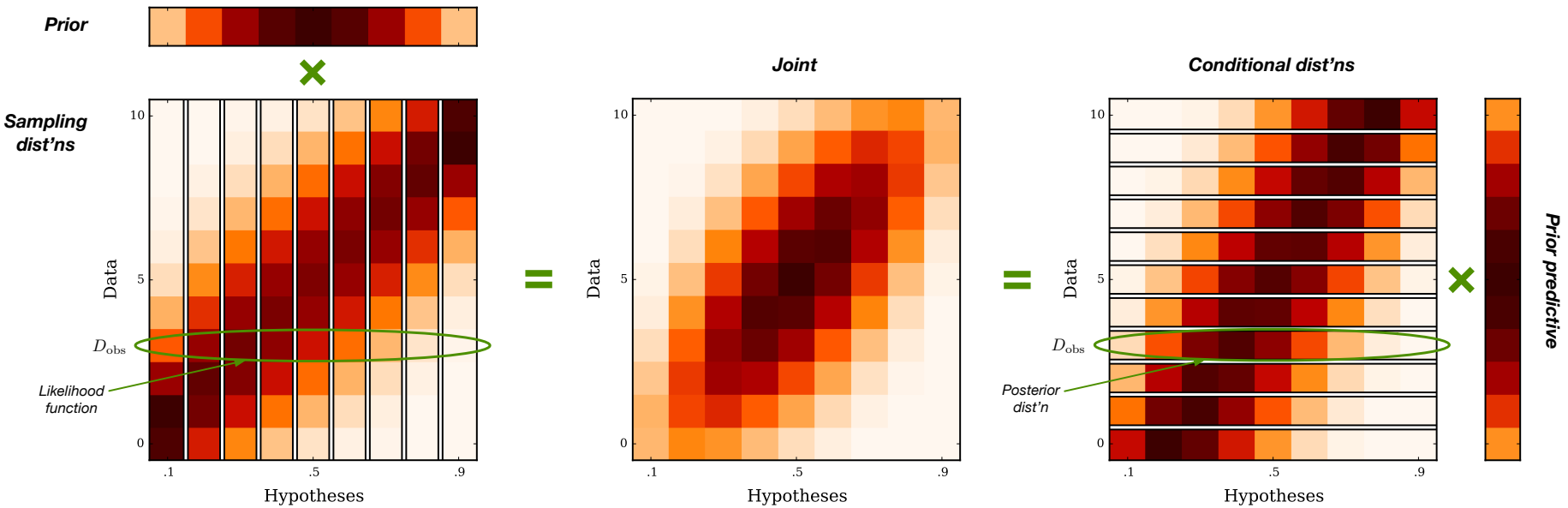
Inference as manipulation of the joint distribution

Bayes's theorem in terms of the *joint distribution*:

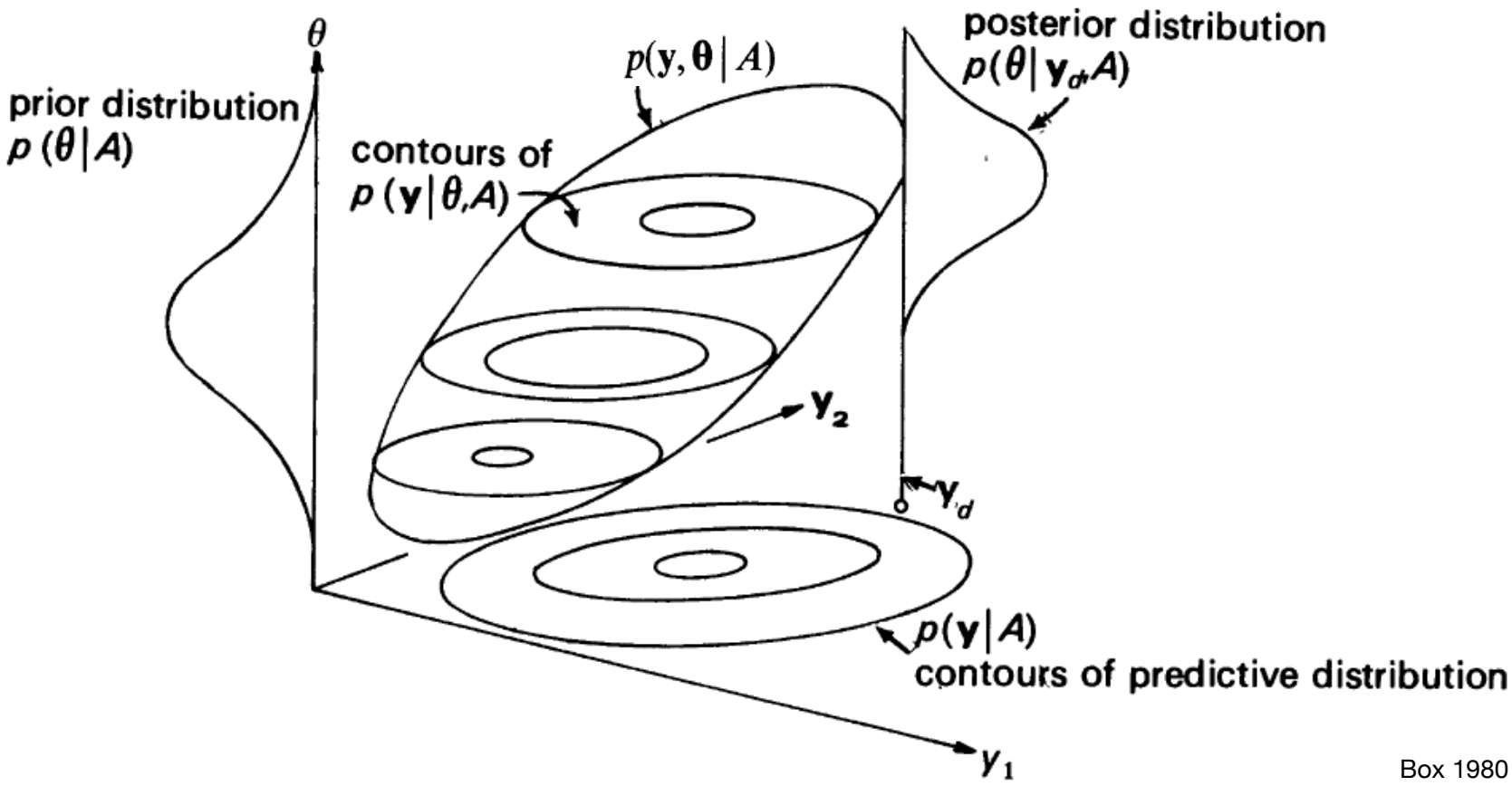
$$P(H_i|I) \times P(D_{\text{obs}}|H_i, I) = P(H_i, D_{\text{obs}}|I) = P(H_i|D_{\text{obs}}, I) \times P(D_{\text{obs}}|I)$$




Joint
Conditional dist'ns
Data
10
5
0
.1
.5
.9
Hypotheses
=
Data
D_{obs}
Posterior dist'n
Hypotheses
×
Prior predictive


Prior
Sampling dist'ns
Data
Hypotheses
D_{obs}
Likelihood function
Joint
Conditional dist'ns
Posterior dist'n
Prior predictive
10
5
0
.1
.5
.9

Continuous data, parameter spaces



Components of Bayes's theorem for a problem with a 1-D parameter space (θ) and a 2-D sample space (**y**), with observed data $\mathbf{y}_d$, and modeling assumptions *A*

Nuisance Parameters and Marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example

We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b.

We have additional data just about b.

What do the data tell us about s?

Marginal posterior distribution

To summarize implications for s, accounting for b uncertainty, *marginalize*:

$$\begin{aligned} p(s|D,M) &= \int db\; p(s,b|D,M) \\ &\propto p(s|M) \int db\; p(b|s,M)\, \mathcal{L}(s,b) \\ &= p(s|M)\mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function for s*:

$$\mathcal{L}_m(s) \equiv \int db\; p(b|s)\, \mathcal{L}(s,b)$$

Bayesian inference in one slide

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \quad \propto \quad p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis $\propto$ ability of hypothesis to *predict* the data

Law of total probability

$$p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) \quad = \quad \sum p(\text{Hypothesis} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Menu

1. What/Why/How of cosmic demographics
2. Basic Bayesian inference (briefly!)
3. **Bayesian computation (also briefly!)**
4. Hierarchical Bayesian modeling
5. Selection effects: Thinned latent point processes

Notation focusing on computational tasks

$$
\begin{aligned}
p(\theta|D,M) &= \frac{p(\theta|M)p(D|\theta,M)}{p(D|M)} \\
&= \frac{\pi(\theta)\mathcal{L}(\theta)}{Z} = \frac{q(\theta)}{Z}
\end{aligned}
$$

- $M =$ model specification
- D specifies observed data
- $\theta =$ model parameters
- $\pi(\theta) =$ prior pdf for θ
- $\mathcal{L}(\theta) =$ likelihood for θ (likelihood function)
- $q(\theta) = \pi(\theta)\mathcal{L}(\theta) =$ "quasiposterior"
- $Z = p(D|M) =$ (marginal) likelihood for the model

Parameter space integrals

For model with m parameters, we need to evaluate integrals like:

$$\int d^m\theta \, g(\theta)\, \pi(\theta)\, \mathcal{L}(\theta) \quad = \quad \int d^m\theta \, g(\theta)\, q(\theta)$$

- $g(\theta) = 1 \rightarrow Z = p(D|M)$ (norm. const., model likelihood)
- $g(\theta) = \theta \rightarrow$ posterior mean for θ
- $g(\theta) =$ 'box' $\rightarrow$ probability $\theta \in$ credible region
- $g(\theta) = 1$, integrate over subspace $\rightarrow$ marginal posterior
- $g(\theta) = \delta[\psi - \psi(\theta)] \rightarrow$ propagate uncertainty to $\psi(\theta)$

Except for optimization, Bayesian computation amounts to *computing the expectation of some function $g(\theta)$ with respect to the posterior dist'n for θ*

Contrast with frequentist computation, which integrates over *sample space*, e.g., via Monte Carlo simulation of data

Bayesian Computation Menu

Large sample size, N: Laplace approximation

- Approximate posterior as multivariate normal $\rightarrow$ det(covar) factors
- Uses ingredients available in χ^2/ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$ (better than $O(1/\sqrt{N})$)

Modest-dimensional models ($m \lesssim 10$ to 20)

- Quadrature, cubature, adaptive cubature
- IID Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($m \gtrsim 5$): Non-IID Monte Carlo

- Posterior sampling — create RNG that samples posterior
  - Markov Chain Monte Carlo (MCMC) is the most general framework
- Sequential Monte Carlo (SMC)
- Approximate(ly) Bayesian computation (ABC)
- ...

Modest-D: IID Monte Carlo Integration

$\int g \times p$ is just the *expectation of* g; suggests approximating with a *sample average* based on IID draws from p:

$$\int d\theta \, g(\theta)p(\theta) \approx \frac{1}{n} \sum_{\theta_i \sim p(\theta)} g(\theta_i) + O(n^{-1/2}) \quad \left[\begin{array}{c} \sim O(n^{-1}) \text{ with} \\ \text{quasi-MC} \end{array} \right]$$

This is like a cubature rule, with *equal weights* and *random nodes*

Ignores smoothness $\rightarrow$ poor performance in 1-D, 2-D vs. quadrature rules

Avoids curse of dimensionality: $O(n^{-1/2})$ regardless of dimension

Why/when it works

- Independent sampling & law of large numbers $\rightarrow$ asymptotic convergence in probability
- Error term is from CLT; requires finite variance

Practical problems

- $p(\theta)$ must be a density we can draw IID samples from—perhaps the prior or a simple posterior, but...
- $O(n^{-1/2})$ multiplier (std. dev'n of g) may be large

$\rightarrow$ *IID* Monte Carlo can be hard if dimension $\gtrsim$ 5–10*

*IID = independently, identically distributed

Posterior sampling

$$\int d\theta\, g(\theta)p(\theta|D) \approx \frac{1}{n} \sum_{\theta_i \sim p(\theta|D)} g(\theta_i) + O(n^{-1/2})$$

When $p(\theta)$ is a posterior distribution, drawing samples from it is called *posterior sampling*:

- *One set of samples* can be used for many different calculations (so long as they don't depend on low-probability events)
- This is the most promising and general approach for Bayesian computation in *high dimensions*—though with a twist (MCMC!)

Challenge: How to build a RNG that samples from a posterior?

In 1-D there are several options that inspire methods for higher-D

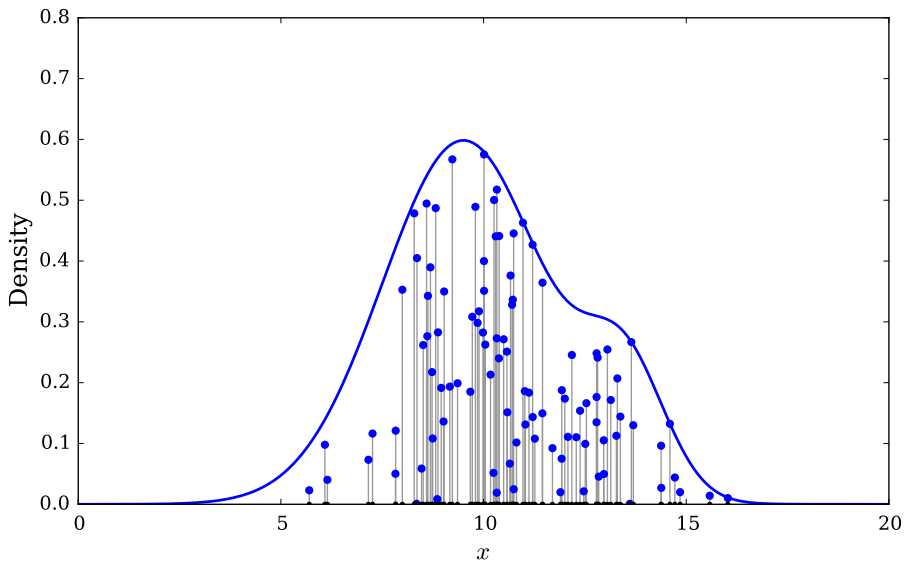
Accept-Reject Algorithm

Goal: Given $q(\theta) \equiv \pi(\theta)\mathcal{L}(\theta)$, build a RNG that draws samples from the probability density function (*PDF*)

$$f(\theta) = \frac{q(\theta)}{Z} \quad \text{with} \quad Z = \int d\theta\, q(\theta)$$

The probability for a region under the *PDF* is the *area (volume) under the curve (surface)*.

$\rightarrow$ Sample points uniformly in volume under q; their θ values will be draws from $f(\theta)$.

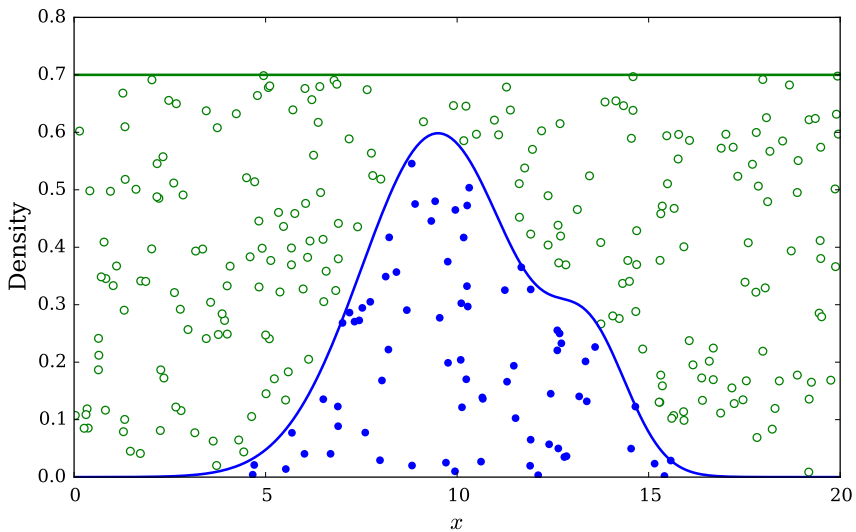


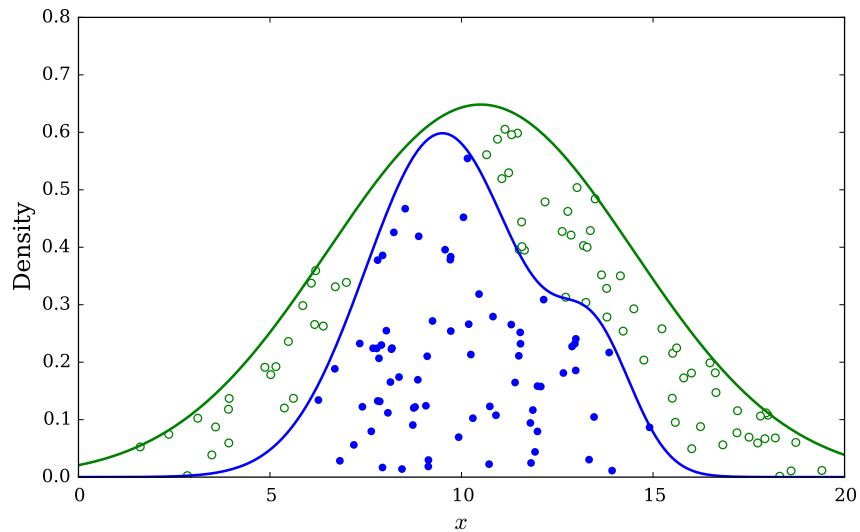
The fraction of samples with θ ("x" in the fig) in a bin of size $\delta\theta$ is the fractional area of the bin.

How can we generate points uniformly under the *PDF*?

Generate *candidate* points uniformly in a region enclosing $q(\theta)$ that you know how to sample from.

Keep the points that end up under q.





Take-away idea: *Propose candidates that may be accepted or rejected*

Accept-Reject Algorithm

1. Choose a tractable density $h(\theta)$ and a constant C so Ch bounds q
2. Draw a candidate parameter value $\theta' \sim h$
3. Draw a uniform random number, u
4. If $q(\theta') < Ch(\theta')$, record θ' as a sample
5. Goto 2, repeating as necessary to get the desired number of samples.

Efficiency = ratio of volumes, Z/C.

In problems of realistic complexity, the efficiency is intolerably low for parameter spaces of more than several dimensions.

Take-away idea: *Propose candidates that may be accepted or rejected*

Posterior sampling beyond 1-D

- Simplest option: *IID Monte Carlo*, for which we can demonstrate *convergence*
  - Weak LLN shows probability for large error between m and μ is bounded by σ^2/N
  - CLT shows the distribution for the error becomes Gaussian at large N, with $\sigma_m = \sigma/\sqrt{N}$
- IID sampling from $p(\theta)$ is feasible in 1-D, but increasingly challenging as dimension increases
  E.g.: Accept/reject will be inefficient unless we can find an envelope that resembles $p(\theta)$ — *hard!*

Markov Chain Monte Carlo*

Accept/Reject aims to produce *independent* samples—each new θ is chosen irrespective of previous draws

Since $\mathbb{E}(m) = \mu$ holds for any joint dist'n *with identical marginals matching* $p(\theta)$, consider *dependent* sampling

Choose new θ points in a way that

- Tends to *move toward* regions with higher probability than current
- Tends to *avoid* lower probability regions

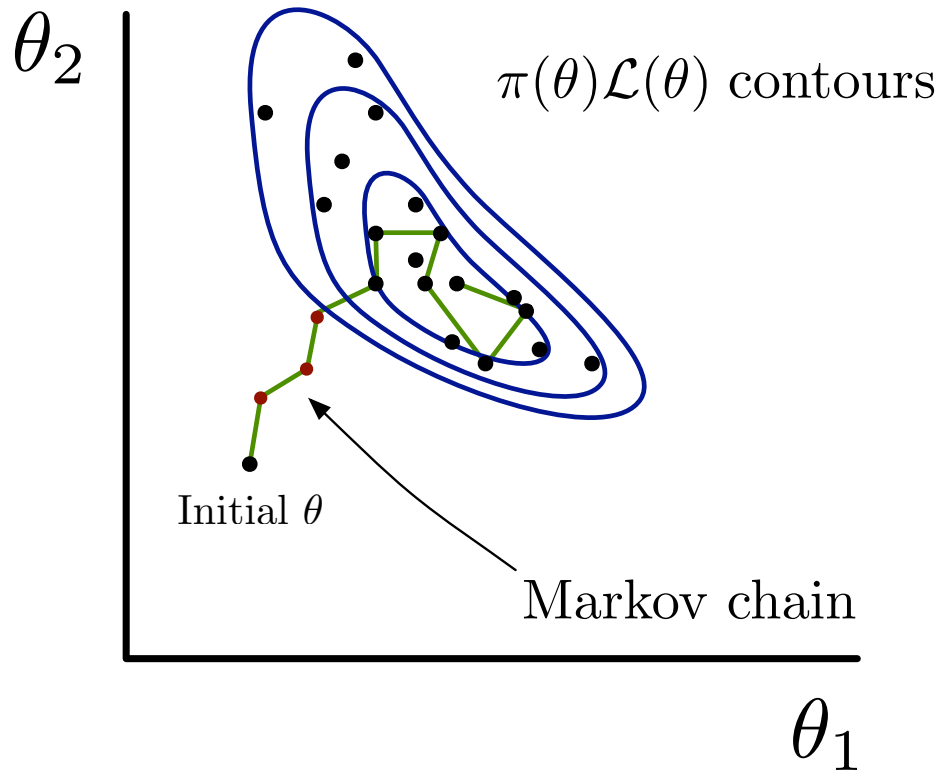
The simplest possibility is a *Markov chain*:

$$p(\text{next location}|\textit{current } \textbf{and } \textit{previous locations}) = p(\text{next location}|\textit{current location})$$

A Markov chain "has no memory"

*Chib & Greenberg (1995): "Understanding the Metropolis-Hastings Algorithm"

Goal: Posterior sampling via Markov chain Monte Carlo



Metropolis-Hastings algorithm

Given a target quasi-distribution $q(x)$ (it need not be normalized):

1. Specify a proposal distribution $k(y|x)$ (make sure it is irreducible and aperiodic).
2. Choose a starting point x; set $t = 0$ and $S_t = x$
3. Increment t
4. Propose a new state $y \sim k(y|x)$
5. If $q(x)k(y|x) < q(y)k(x|y)$, set $S_t = y$; goto (3)
6. Draw a uniform random number u
7. If $u < \frac{q(y)k(x|y)}{q(x)k(y|x)}$, set $S_t = y$; else set $S_t = x$; goto (3)

The rejection step makes this work by ensuring *detailed balance*

The art of MCMC is in *specifying the proposal distribution* $k(y|x)$

We want:

- New proposals to be accepted, so there is movement
- Movement to be significant, so we explore efficiently

These desiderata compete!

Random walk Metropolis (RWM)

Propose an *increment*, z, from the current location, not dependent on the current location, so $y = x + z$ with a specified PDF $K(z)$, corresponding to

$$k(y|x) = K(y - x)$$

The proposals would give rise to a *random walk* if they were all accepted; the M-H rule modifies them to be a kind of directed random walk

Most commonly, a symmetric proposal is adopted:

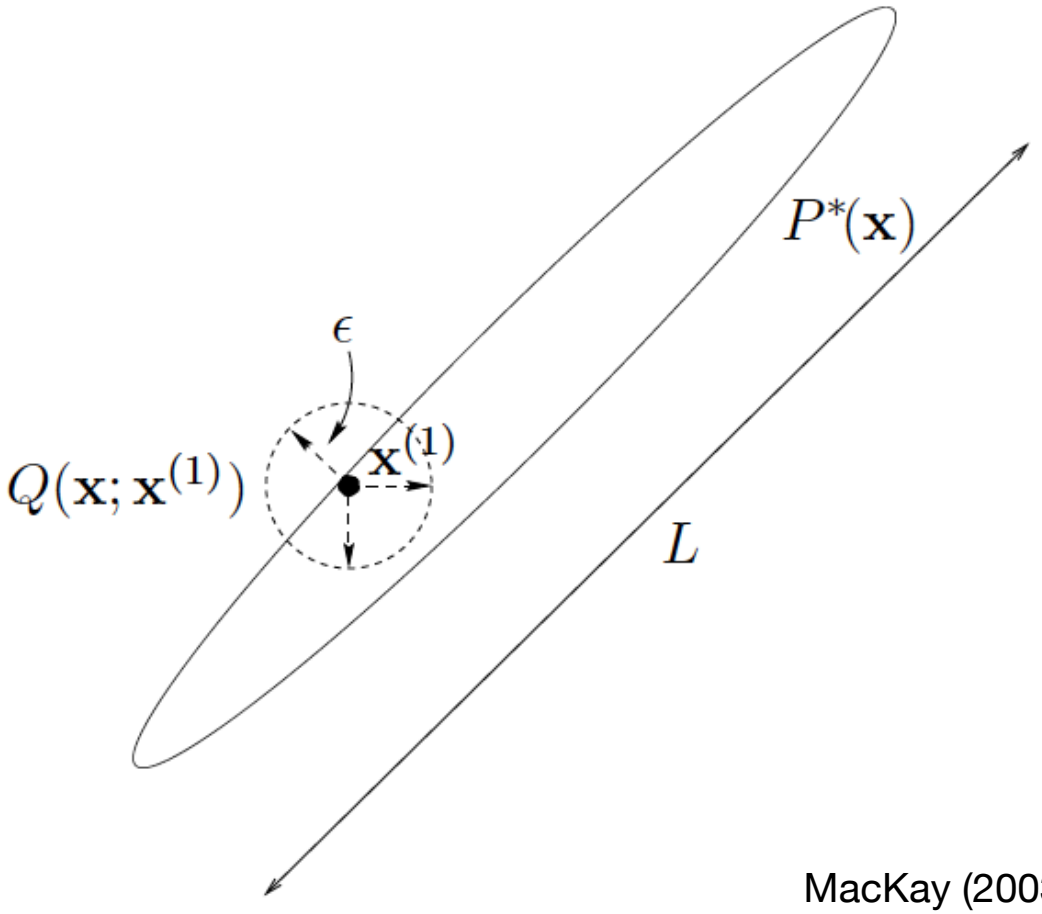
$$k(y|x) = K(|y - x|)$$

The acceptance probability simplifies:

$$\alpha(y|x) = \min\left[\frac{q(y)}{q(x)}, 1\right]$$

Key issues: shape and scale (in all directions) of $K(z)$

RWM in 2-D



MacKay (2003)

Small step size $\rightarrow$ good acceptance rate, but slow exploration

Random Walks

Random walk Metropolis and most other MCMC updates execute a *random walk* through parameter space:

- Moves are local, with a characteristic scale l
- Total distance traversed over time $t \propto \sqrt{t}$

This is a relatively slow (albeit steady) rate of exploration

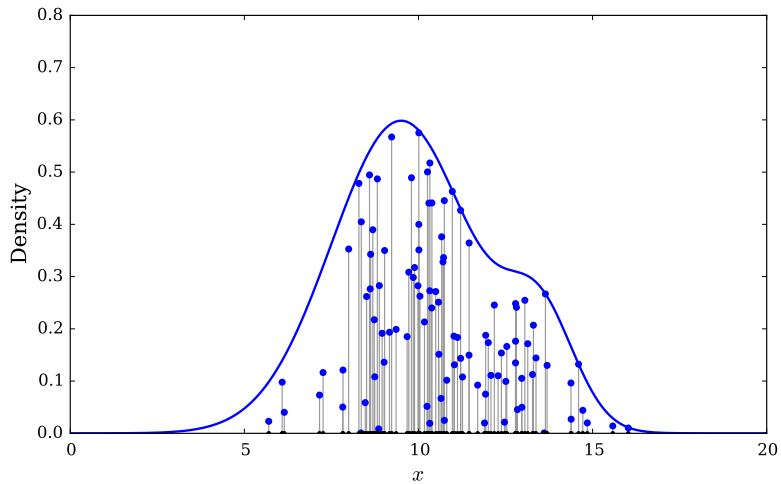
Multimodality $\rightarrow$ even slower exploration; only rare large jumps can move between modes

We need methods designed to make large moves

Auxiliary variables

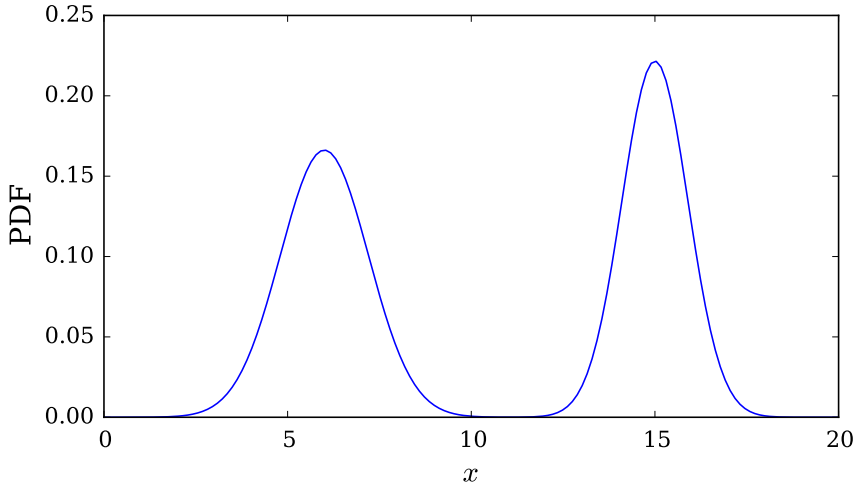
The accept/reject method for sampling a d-D density:

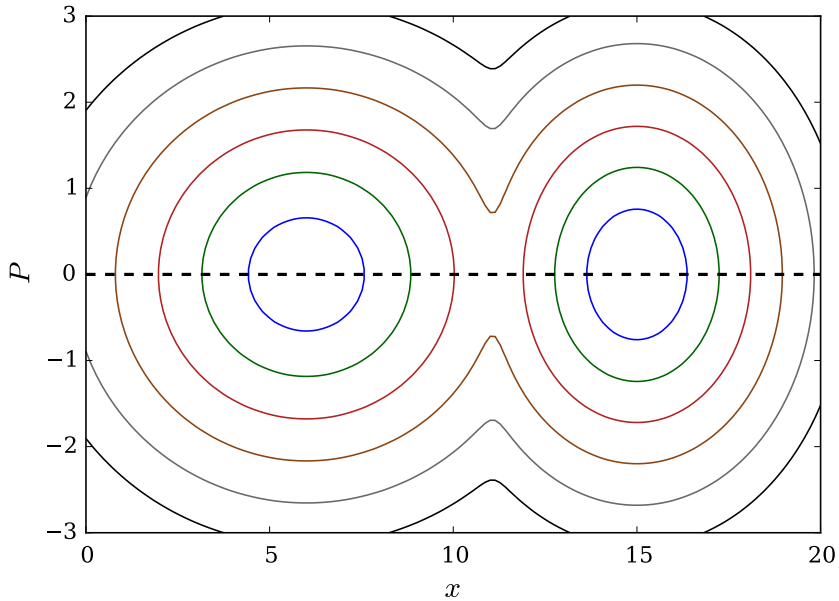
- Sample from a *uniform* $(d+1)$-D density (with a complicated boundary):



- Report the marginal samples for the d original dimensions

A paradoxical notion motivating some advanced MCMC methods is that making the problem "harder" (higher-dimensional) may actually make it *easier*





Double the dimensionality!

$$p(x, P) \propto q(x) \times f(P)$$

$$p(x) = \int dP\, p(x, P) \propto q(x)$$

$$p(P) = \int dx\, p(x, P) \propto f(P)$$

- Pick $P \sim f(P)$
- Move along a contour in phase space
- Drop P, keep x

Will work if the phase space motion corresponds to sampling $p(x, P)$

Hamiltonian (Hybrid) Monte Carlo

Give samples "momentum" so moves tend to go in the same direction a while; use derivatives to guide the evolution $\rightarrow$ suppress random walks

Adds d additional variables, P, with a joint Gaussian dist'n:

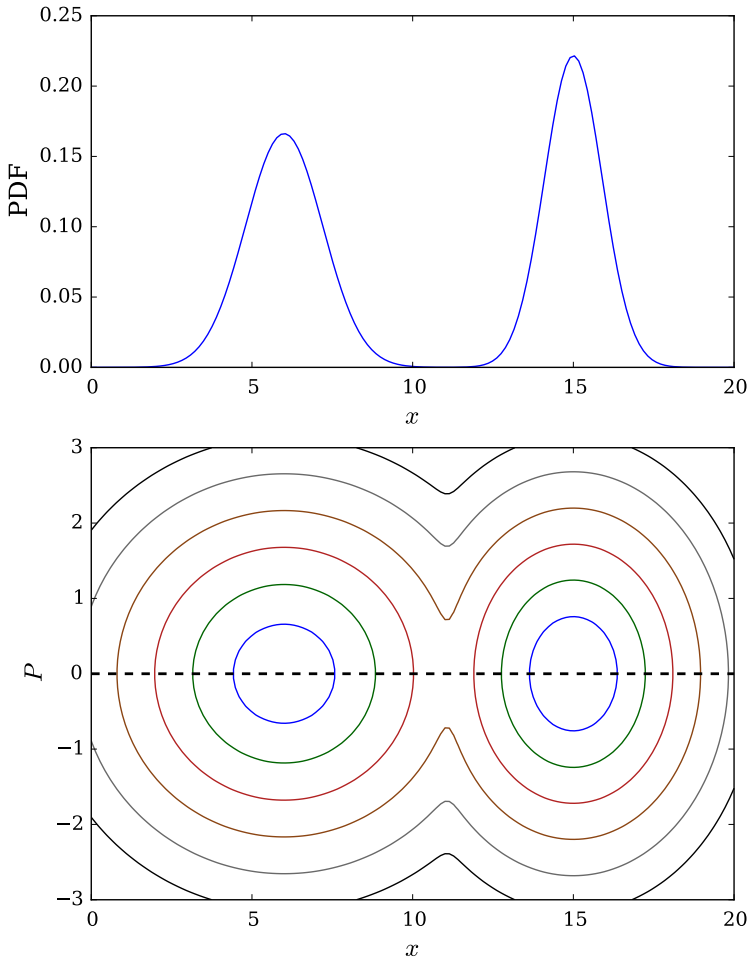
$$\log p(\theta, P) = -\left[U(\theta) + \frac{1}{2}P^2\right]; \qquad U(\theta) \equiv -\log q(\theta)$$

Sample P from a Gaussian, and use it to generate proposals via

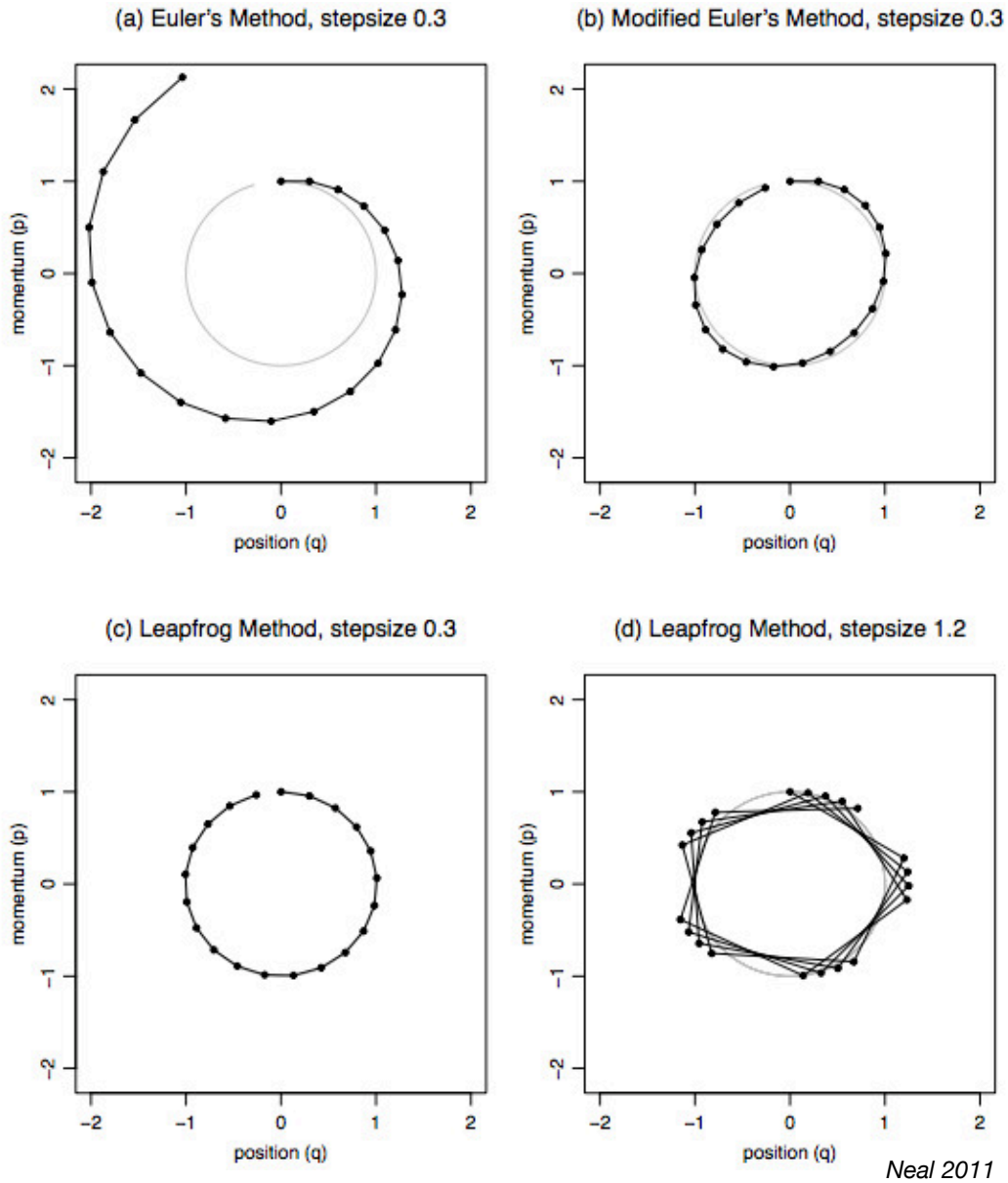
$$\dot{\theta} = P; \qquad \dot{P} = -\frac{\partial H}{\partial \theta}$$

Hamiltonian dynamics $\rightarrow$ reversible, preserves volume, keeps p constant (proposals always accepted)

Navigating the phase space

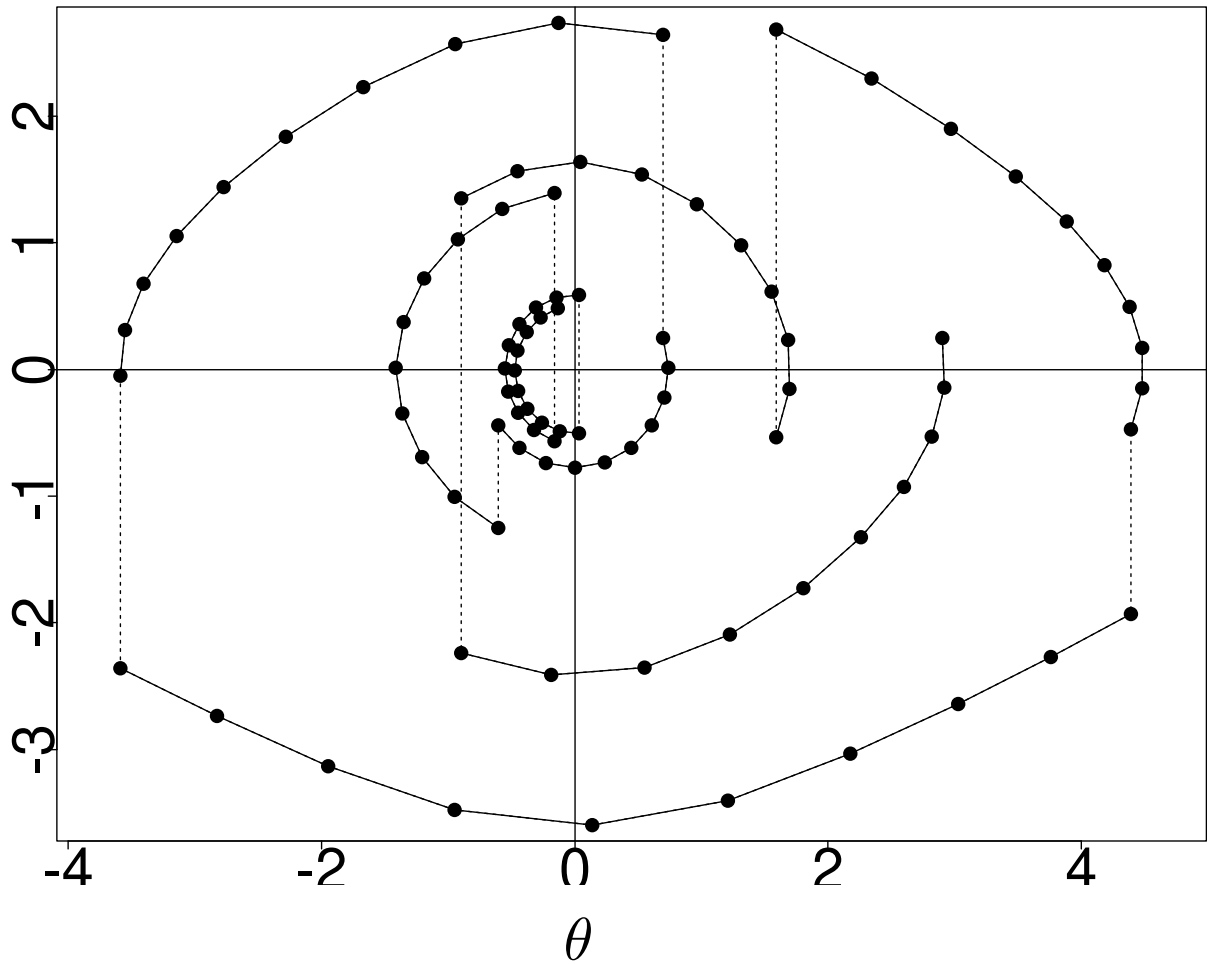


Numerical integration (1-D)



Neal 2011

Sampling a 1-D Student-t dist'n with dof$=5$



HMC vs. random walk (2-D)

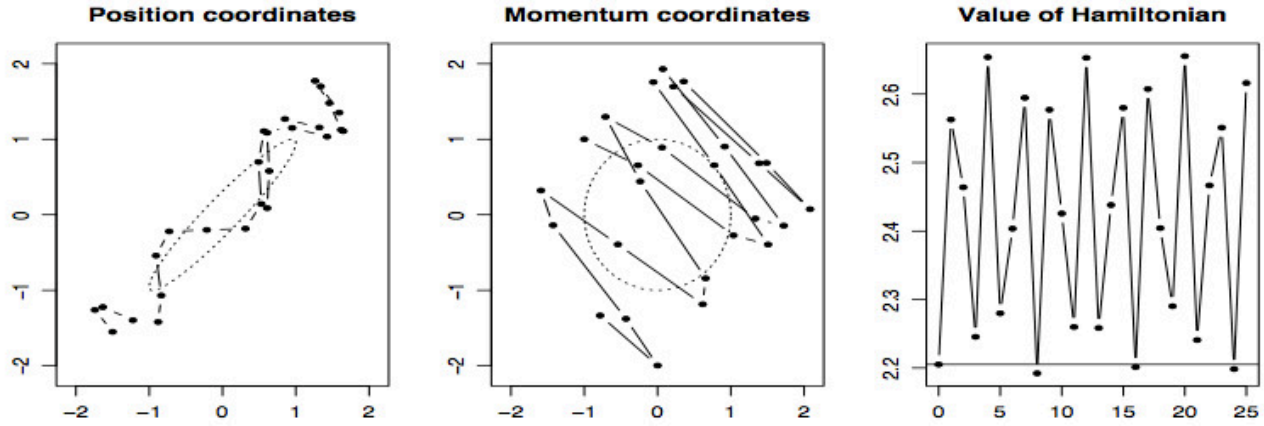
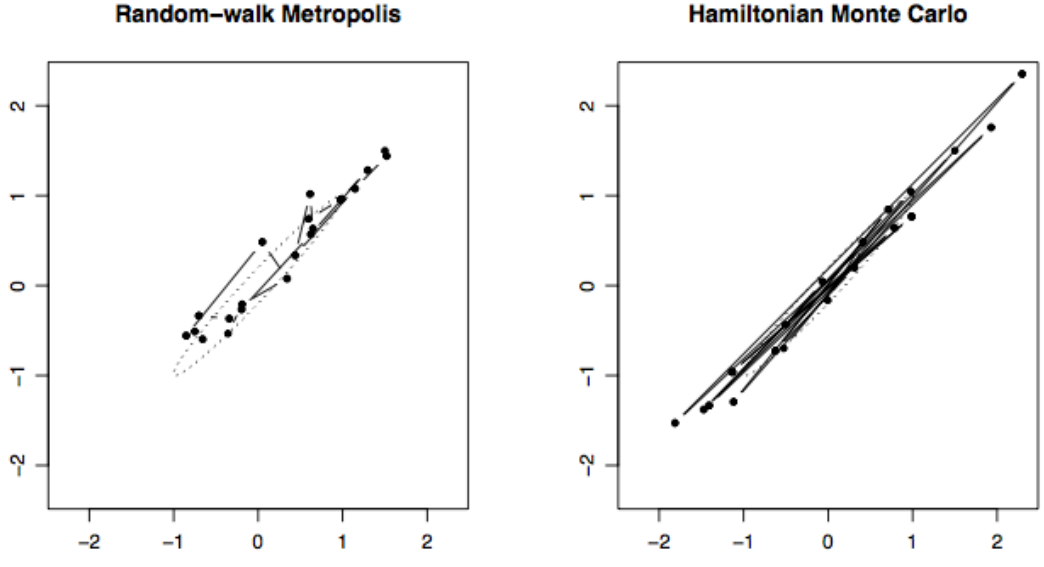


Figure 3: A trajectory for a 2D Gaussian distribution, simulated using 25 leapfrog steps with a stepsize of 0.25. The ellipses plotted are one standard deviation from the means. The initial state had $q = [-1.50, -1.55]^T$ and $p = [-1, 1]^T$.

20 iterations



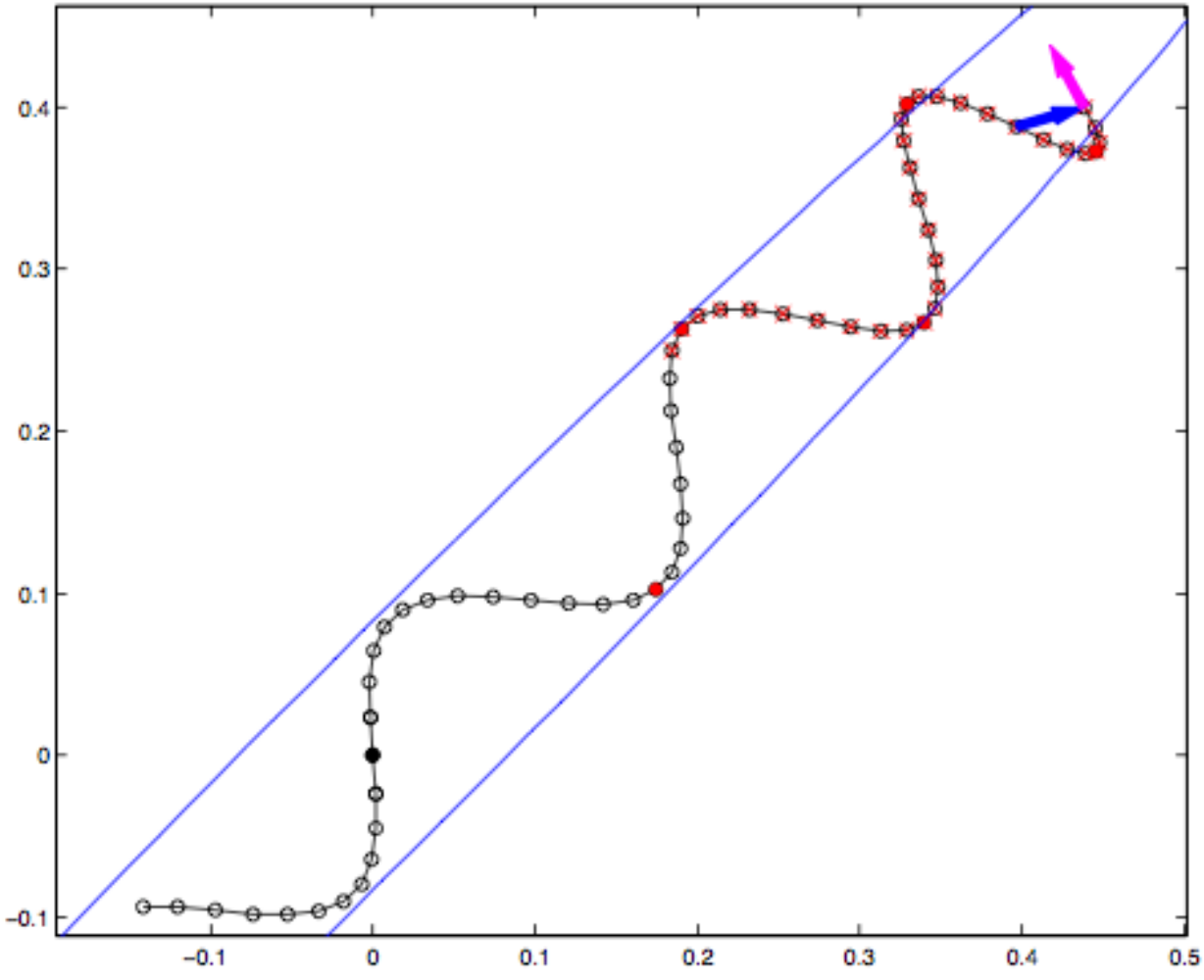
Neal 2011

Challenges for basic HMC

- Tuning parameters:
  - Choosing time step size, ϵ, and integration length, L
  - Handling problems with very different scales along different dimensions
- Computing the needed derivatives

Tuning integration length

No-U-Turn Sampler (NUTS)



Hoffman & Gelman 2013

Mass matrix = metric

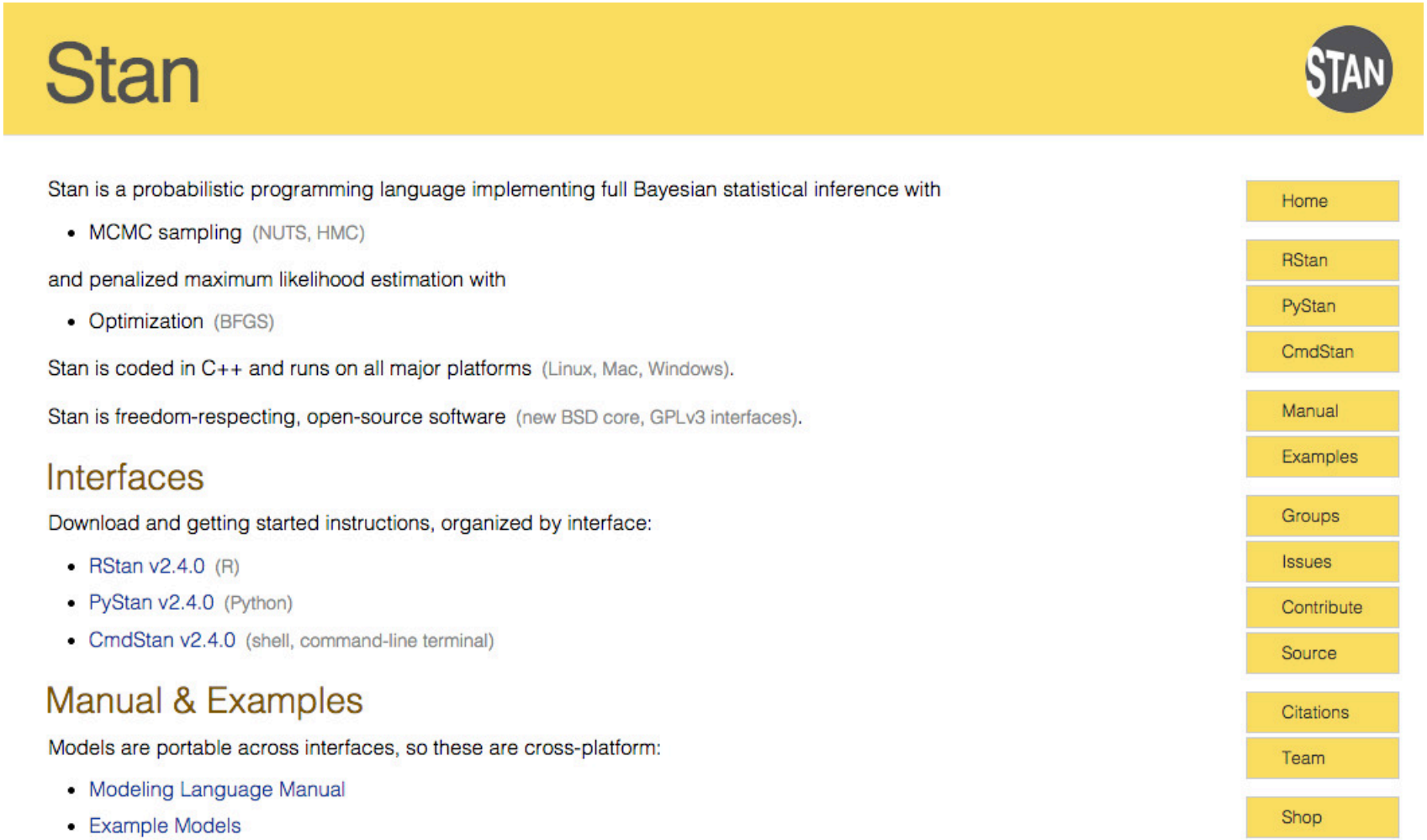
Add d additional variables, P, with a *correlated* Gaussian dist'n:

$$\log p(\theta, P) = -\left[U(\theta) + \frac{1}{2} P \cdot M^{-1} \cdot P\right]; \qquad U(\theta) \equiv -\log p(\theta)$$

M introduces d more tuning parameters!

- **Euclidean manifold HMC:** Use the Hessian at the mode
- **Riemannian manifold HMC:** Use position-dependent $M(\theta)$

Stan: mc-stan.org



http://mc-stan.org/

http://discourse.mc-stan.org/

Stan capabilities

- Hamiltonian Monte Carlo (HMC)
  - sample parameters on unconstrained space
    → transform + Jacobian adjustment
  - gradients of the model wrt parameters
    → automatic differentiation
  - sensitive to tuning parameters → **No-U-Turn Sampler**
- No-U-Turn Sampler (NUTS)
  - warmup: estimates mass matrix and step size
  - sampling: adapts number of steps
  - **maintains detailed balance**
- Optimization
  - BFGS, Newton's method

From Daniel Lee

Other capabilities in progress. . .

Stan Store

Stylish T-Shirt
$25.49

Stylish T-Shirt
$16.49

Stylish T-Shirt
$21.49

Stylish T-Shirt
$16.49

Stylish T-Shirt
$16.49

Stan Mug
$11.99

$15.49

$15.49

$27.49

$12.99

Menu

1. What/Why/How of cosmic demographics
2. Basic Bayesian inference (briefly!)
3. Bayesian computation (also briefly!)
4. **Hierarchical Bayesian modeling**
5. Selection effects: Thinned latent point processes

Joint and conditional distributions

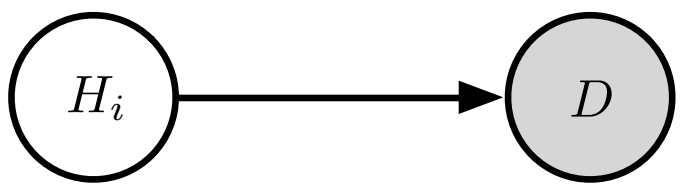
Bayesian inference is largely about the interplay between *joint*, *conditional*, and *marginal* distributions for related quantities

Ex: Bayes's theorem relating hypotheses and data ($|| \, \mathcal{C}$):

$$P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)} = \frac{P(H_i, D)}{P(D)} = \frac{\text{joint for everything}}{\text{marginal for knowns}}$$

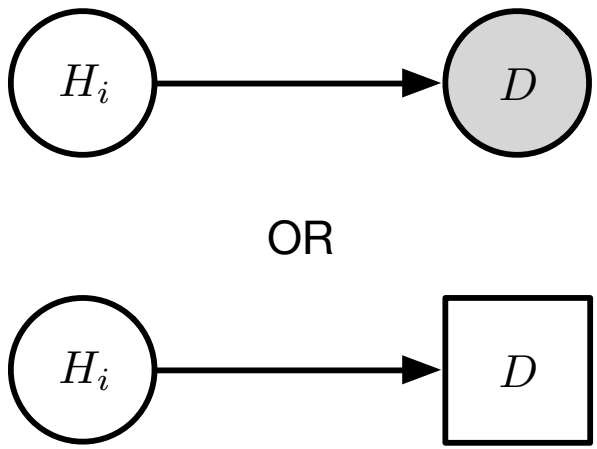
The usual form identifies an available factorization of the joint

Express this via a *directed acyclic graph* (DAG):



Joint distribution structure as a graph

- Graph = *nodes/vertices* connected by *edges/links*
- Circular/square nodes/vertices = a priori uncertain quantities (gray/square = becomes known as data)
- Directed edges specify conditional dependence
- Absence of an edge indicates conditional *in*dependence
  → *the most important edges are the missing ones*

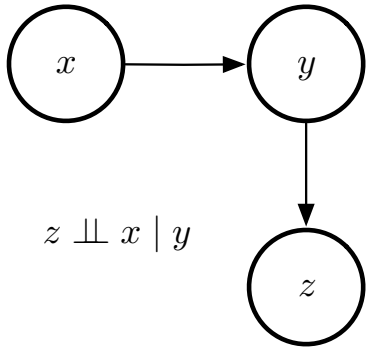


$$P(H_i, D) = P(H_i) \times P(D|H_i)$$

DAGs with missing edges

Conditional independence

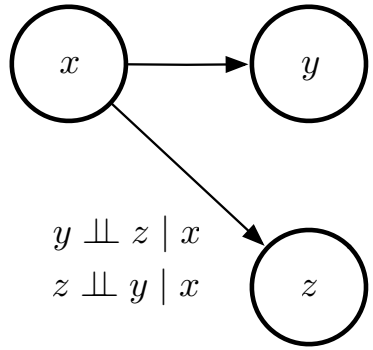
$p(x)\,p(y|x)\,p(z|y)$



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

"Causal chain"

$p(x)\,p(y|x)\,p(z|x)$



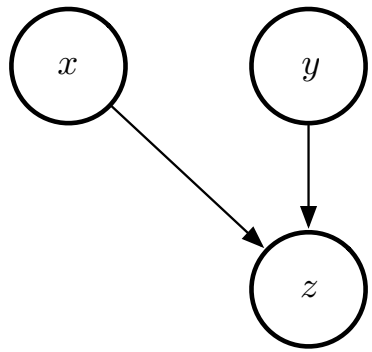
$y \perp\!\!\!\perp z \mid x$

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

"Common cause"

Conditional dependence

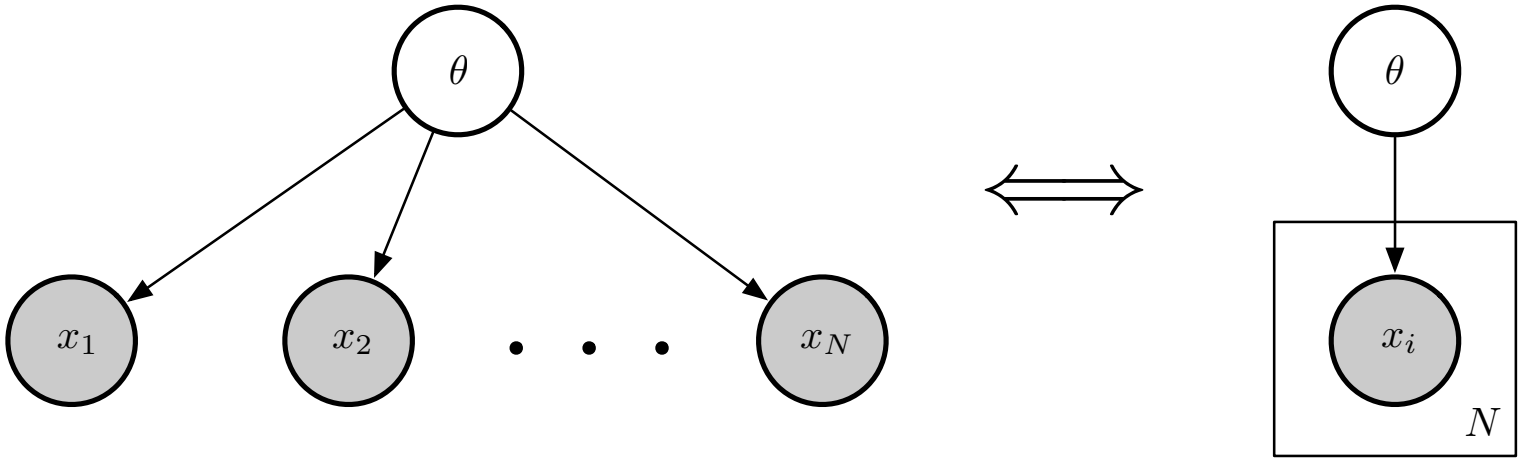
$p(x)\,p(y)\,p(z|x,y)$



"Common effects"

Bayes's theorem with IID samples

For model with parameters θ predicting data $D = \{x_i\}$ that are IID given θ:



$$p(\theta, D) = p(\theta)p(\{x_i\}|\theta) = p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

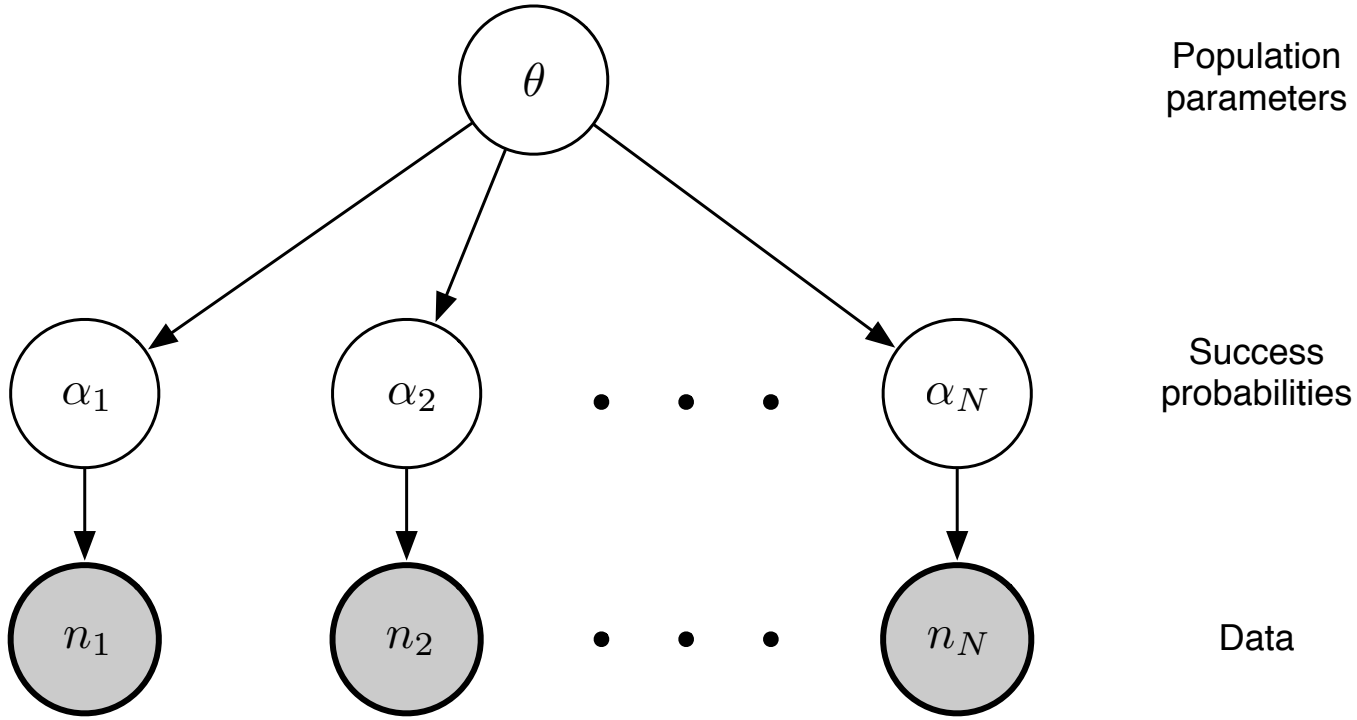
To find the posterior for the unknowns (θ), divide the joint by the marginal for the knowns ($\{x_i\}$):

$$p(\theta|\{x_i\}) = \frac{p(\theta)\prod_{i=1}^{N} p(x_i|\theta)}{p(\{x_i\})} \quad \text{with} \quad p(\{x_i\}) = \int d\theta\, p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

A population of coins/flippers

Each flipper+coin flips different number of times

- What do we learn about the *population* of coins—the distribution of αs?
- How does population membership effect inference for a single coin's α?



$$p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta)\ p(n_i|\alpha_i)$$
$$= \pi(\theta) \prod_i p(\alpha_i|\theta)\ \ell_i(\alpha_i)$$

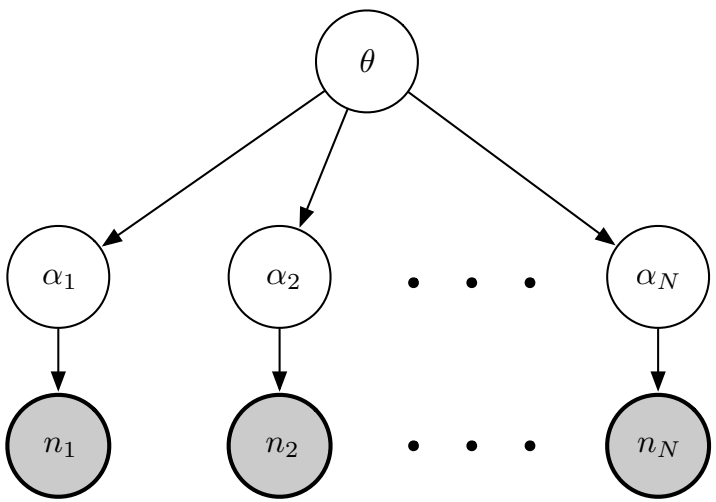
Terminology: θ are *hyperparameters*, $\pi(\theta)$ is the *hyperprior*

A simple multilevel model: beta-binomial

Goals:

- Learn a population-level "prior" by pooling data
- Account for population membership in member inferences

Qualitative



Population parameters

Success probabilities

Data

$$p(\theta, \{\alpha_i\}, \{n_i\}) = \pi(\theta) \prod_i p(\alpha_i|\theta)\ p(n_i|\alpha_i)$$
$$= \pi(\theta) \prod_i p(\alpha_i|\theta)\ \ell_i(\alpha_i)$$

Quantitative

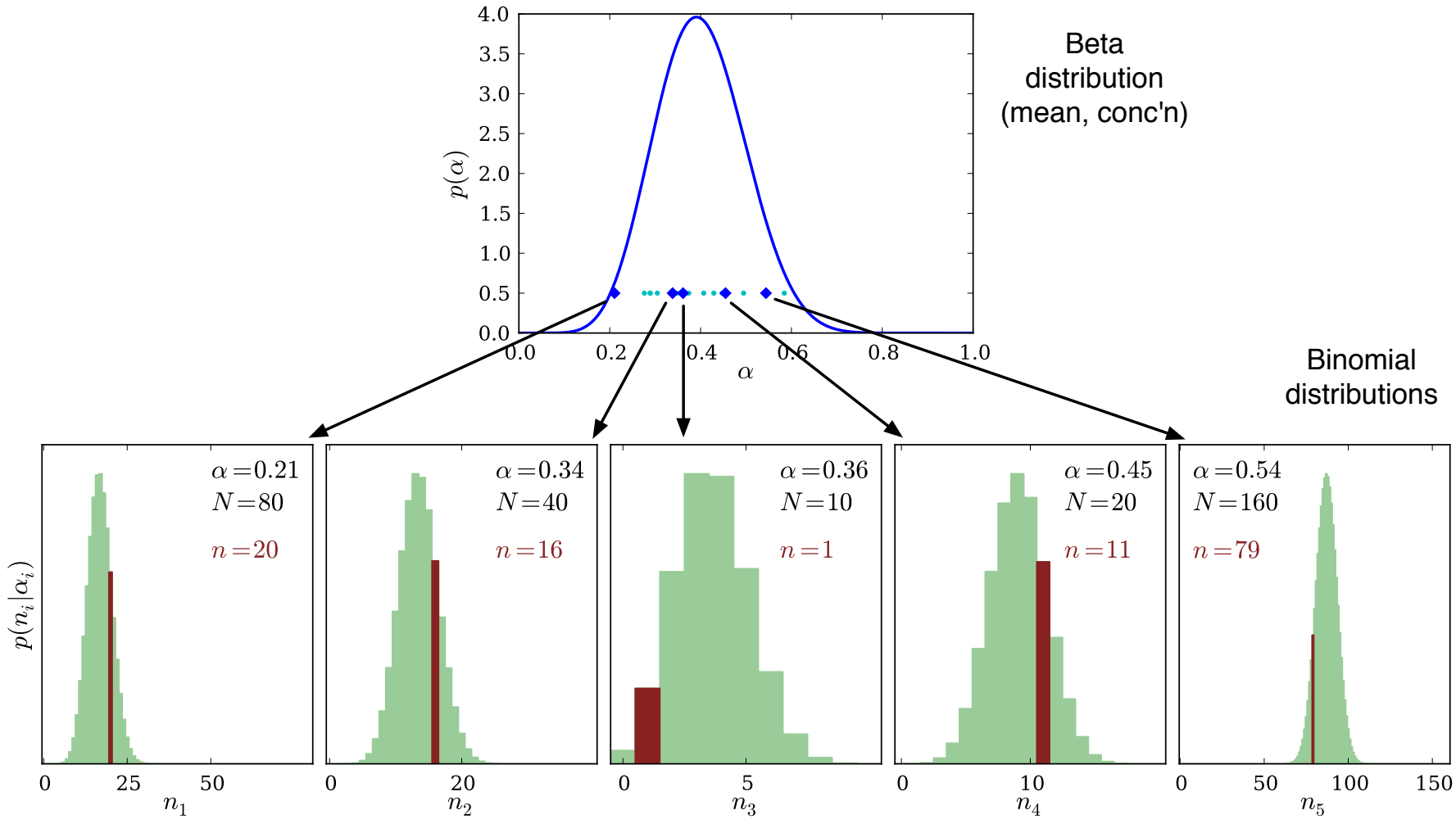
$$\theta = (a, b) \text{ or } (\mu, \sigma)$$

$$\pi(\theta) = \text{Flat}(\mu, \sigma)$$

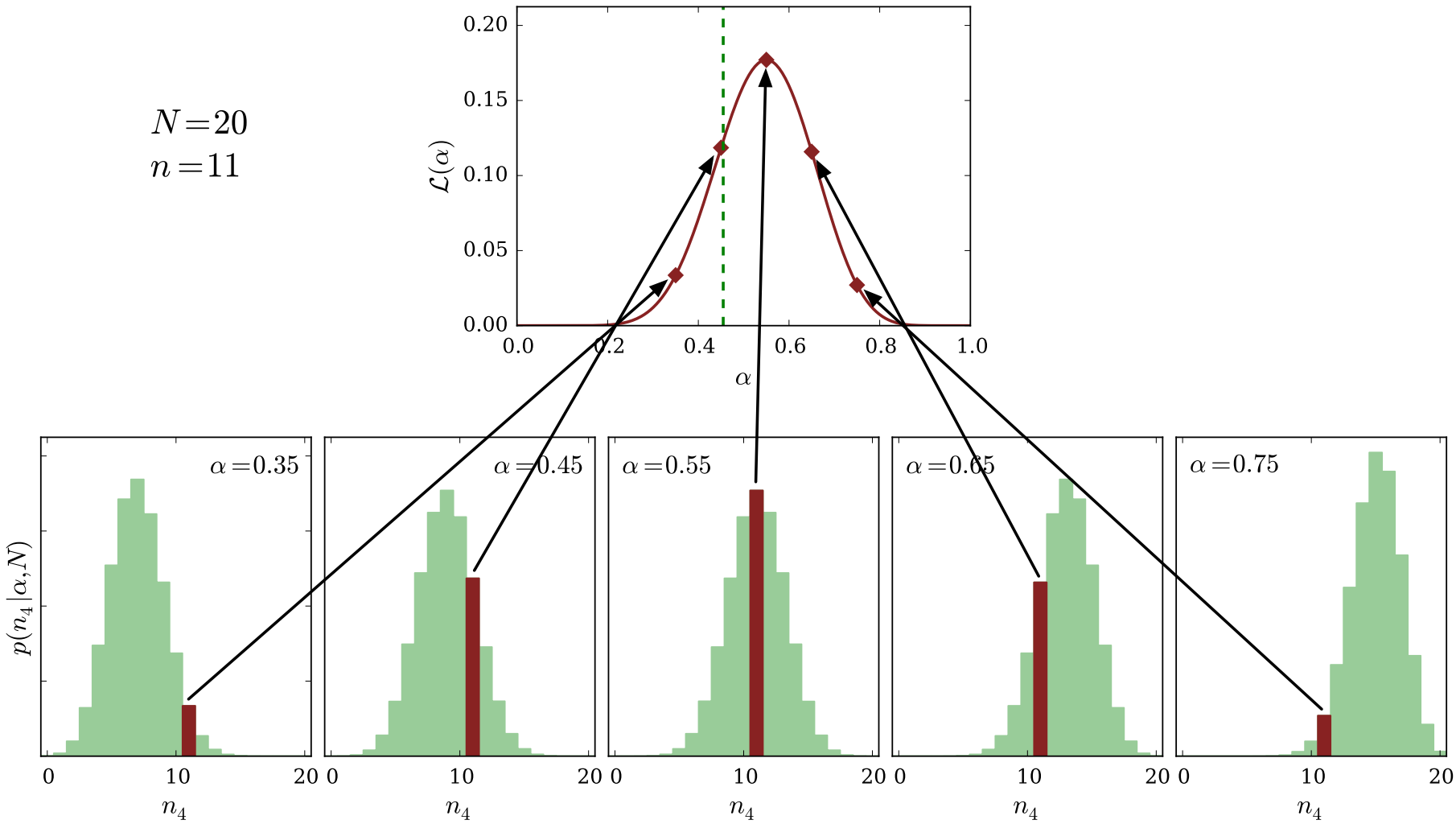
$$p(\alpha_i|\theta) = \text{Beta}(\alpha_i|\theta)$$

$$p(n_i|\alpha_i) = \binom{N_i}{n_i} \alpha_i^{n_i} (1 - \alpha_i)^{N_i - n_i}$$

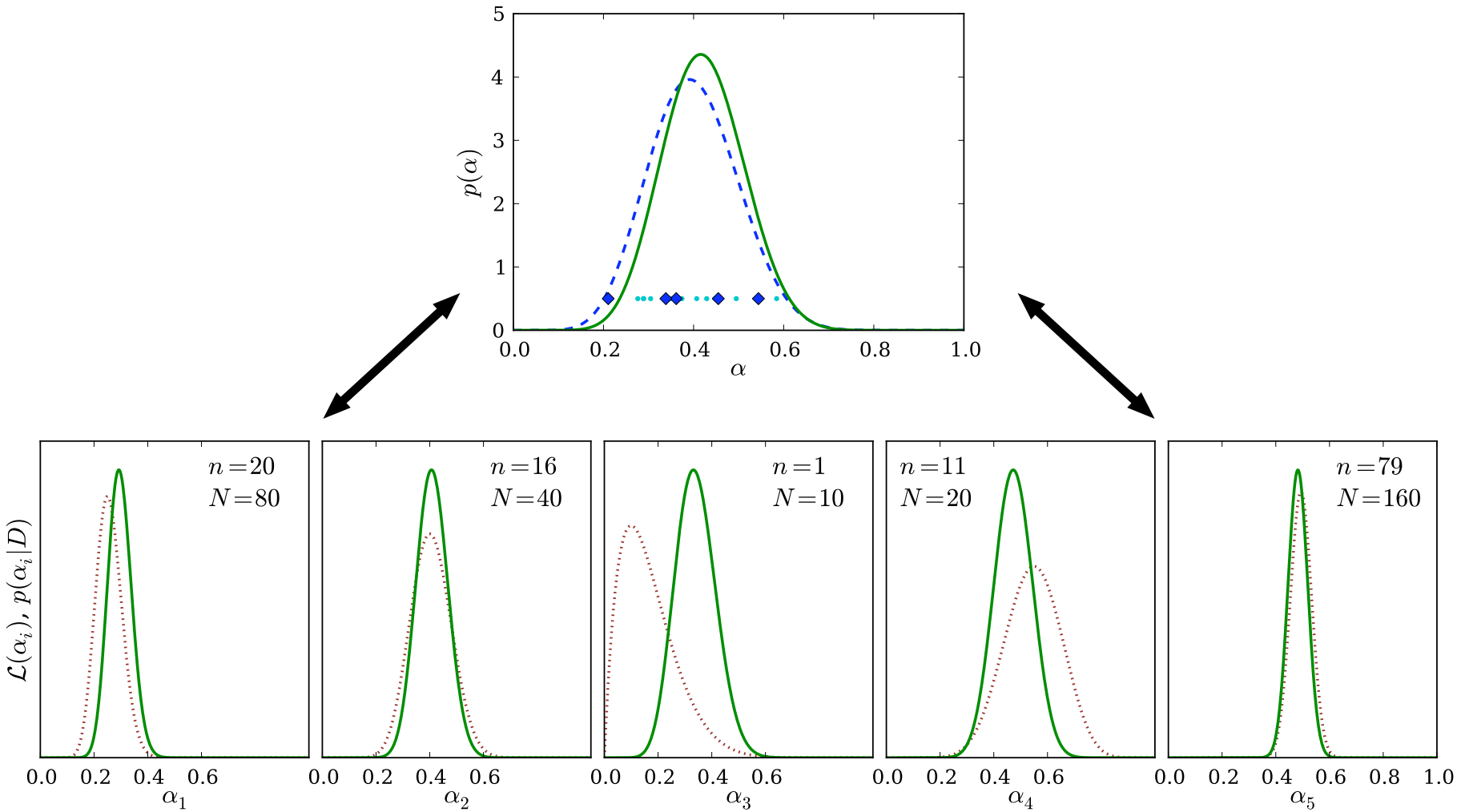
Generating the population & data



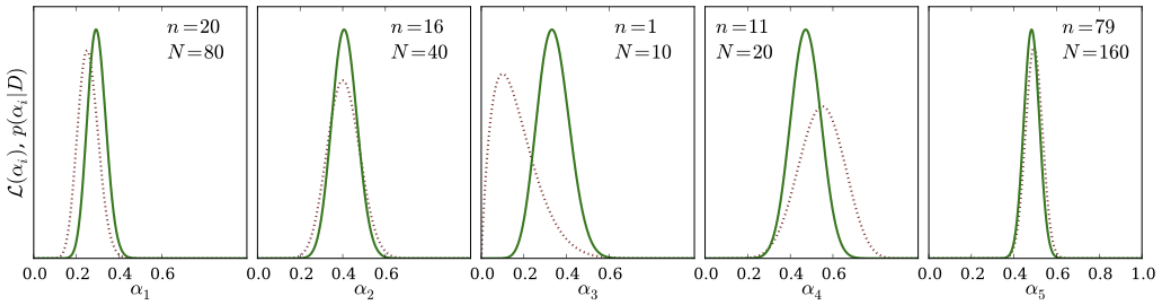
Likelihood function for one member's α

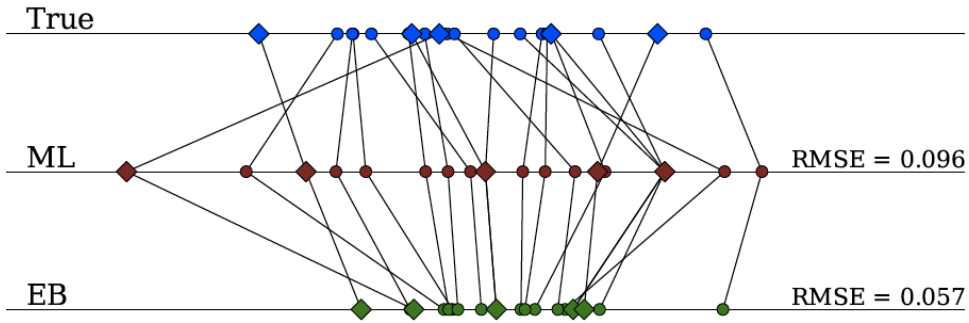


Learning the population distribution



Lower level estimates





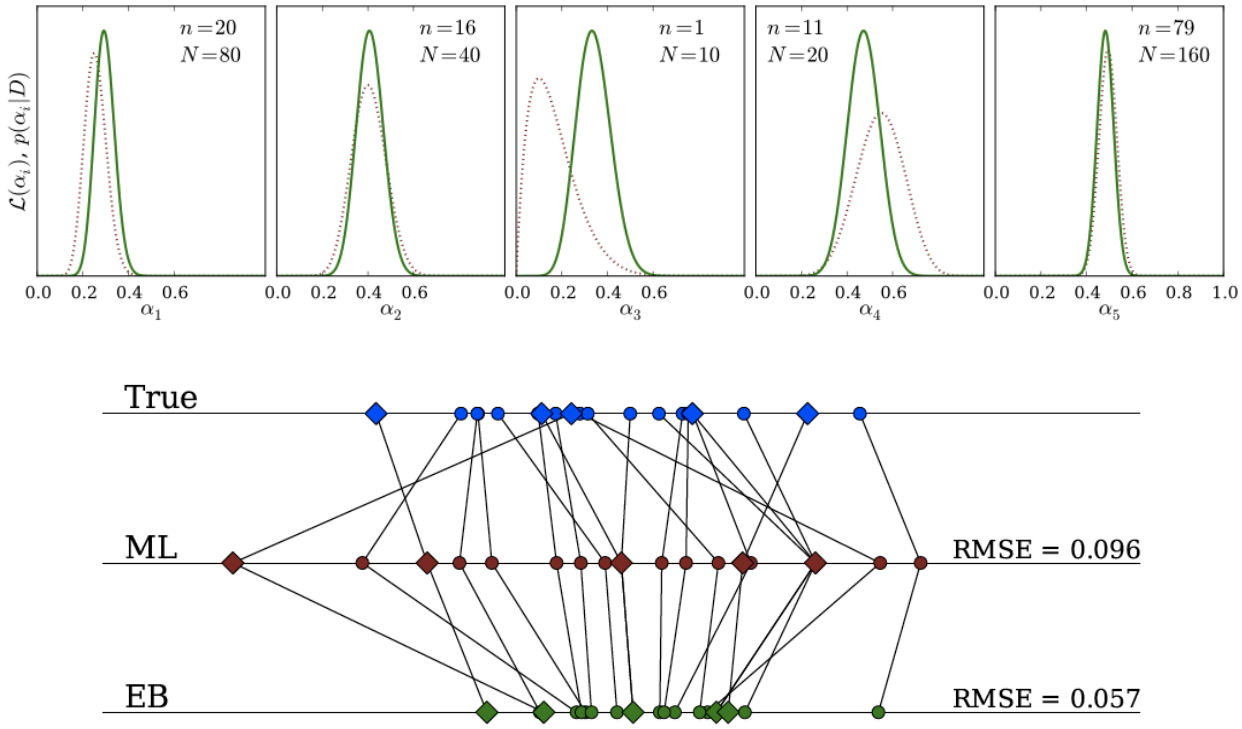
Two approaches

- **Hierarchical Bayes (HB):** Calculate marginals

$$p(\alpha_j|\{n_i\}) \propto \int d\theta\, \pi(\theta) \prod_{i \neq j} \int d\alpha_i\, p(\alpha_i|\theta)\, p(n_i|\alpha_i)$$

- **Empirical Bayes (EB):** Plug in an optimum $\hat{\theta}$ and estimate $\{\alpha_i\}$
  View as approximation to HB, or a frequentist procedure that estimates a prior from the data

Lower level estimates



Bayesian outlook

- Marginal posteriors are *narrower* than likelihoods
- Point estimates tend to be closer to true values than MLEs (averaged across the population)
- Joint distribution for $\{\alpha_i\}$ is *dependent*

Frequentist outlook

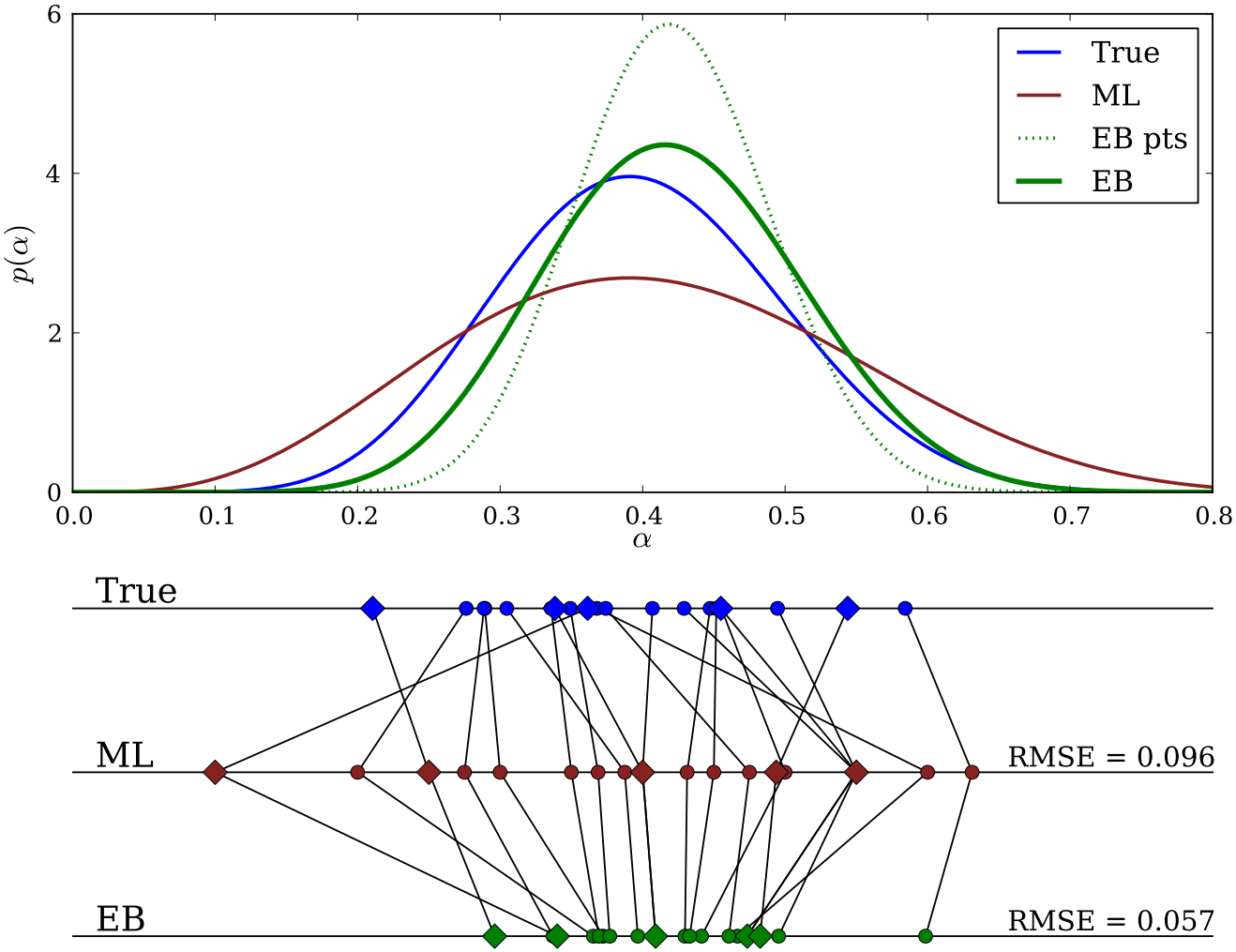
- Point estimates are biased
- Reduced variance $\rightarrow$ estimates are closer to truth on average (lower MSE in repeated sampling)
- Bias for one member estimate depends on data for all other members

Lingo

- Estimates *shrink* toward prior/population mean
- Estimates "muster and *borrow strength*" across population (Tukey's phrase); increases accuracy and precision of estimates
- Efron* describes shrinkage as a consequence of accounting for *indirect evidence*

*Bradley Efron (2010): "The Future of Indirect Evidence"

Population and member estimates



Competing data analysis goals

"Shrunken" member estimates provide improved & reliable estimate for population member properties

But they are *under-dispersed* in comparison to the true values $\rightarrow$ not optimal for estimating *population* properties*

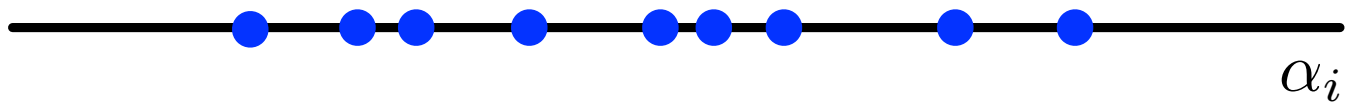
No point estimates of member properties are good for all tasks!

We should view population data tables/catalogs as providing
descriptions of member likelihood functions,
not "estimates with errors"

*Louis (1984); Eddington noted this in 1940! Cf. "corrections" for Eddington/Malmquist/Lutz-Kelker bias

Measurement error perspective

If the data provided *precise* $\{\alpha_i\}$ values (coin measurements, flip physics), we could easily model them as points drawn from a (beta) population PDF with params θ:



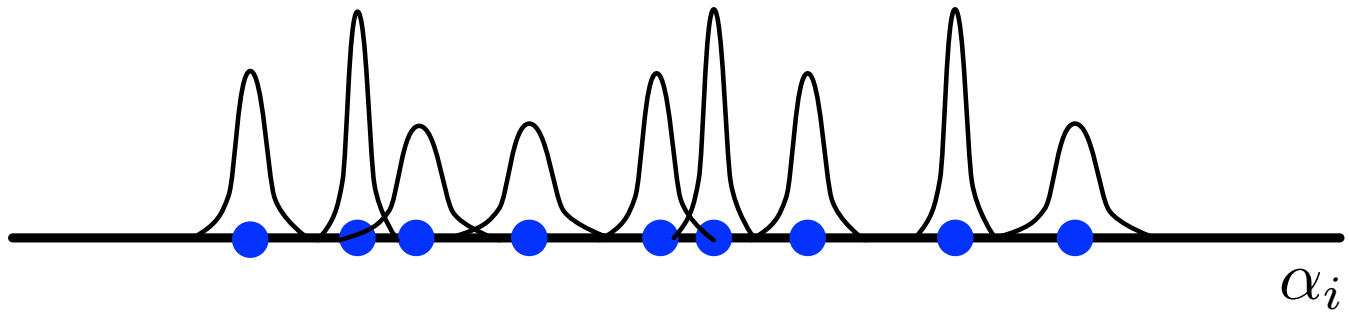
$$D = \{\alpha_i\}$$

$$p(D|\theta) = \prod_i p(\alpha_i|\theta)$$

$$= \prod_i \mathrm{Beta}(\alpha_i|\theta)$$

(A *binomial point process*)

Here the finite number of flips provide *noisy measurements of each* α_i, described by the member likelihood functions $\ell_i(\alpha_i)$;



$$D = \{n_i\}$$

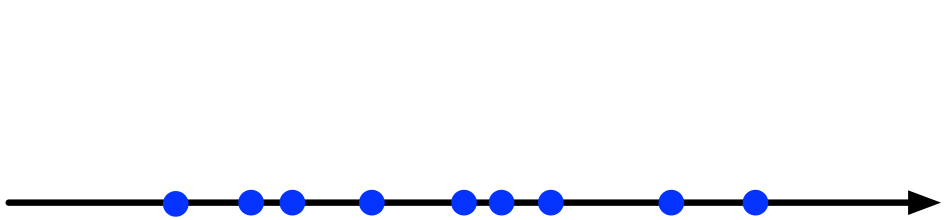
$$\begin{aligned} p(D|\theta) &= \prod_i \int d\alpha_i \; p(D, \{\alpha_i\}|\theta) \\ &= \prod_i \int d\alpha_i \; p(\alpha_i|\theta)\, p(n_i|\theta) \\ &= \prod_i \int d\alpha_i \; \mathrm{Beta}(\alpha_i|\theta)\; \mathrm{Binom}(n_i|\theta) \end{aligned}$$

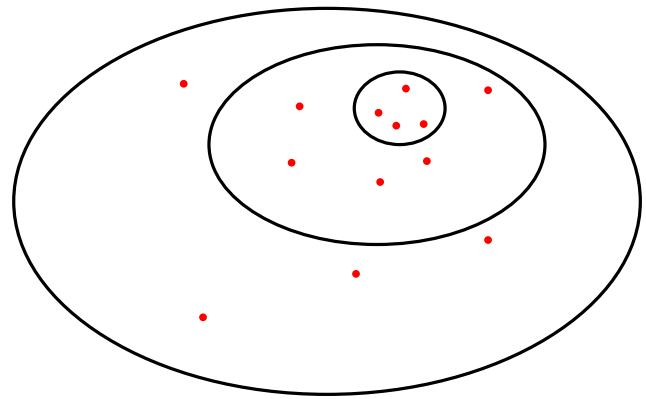
This is a prototype for *measurement error problems*

Basic demography with with cond. indep. samples

Suppose $f(x|\theta)$ is a distribution for an observable, x (scalar or vector):

- $x_i = F_i$: Number-size dist'n/number counts/log N–log S
- For $x_i = (z_i, F_i)$: Luminosity function





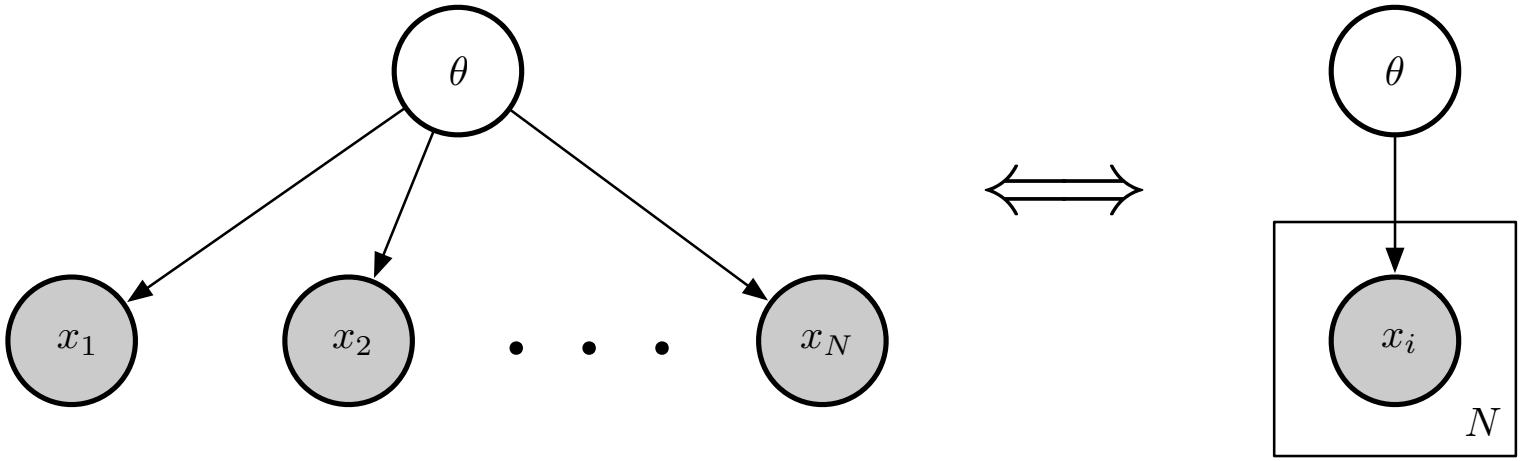
From N precisely measured samples, $\{x_i\}$, we can infer θ from

$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$

$$p(\theta|\{x_i\}) \propto p(\theta)\mathcal{L}(\theta) = p(\theta, \{x_i\})$$

A *binomial point process* (Poisson if N is random)

Graphical representation



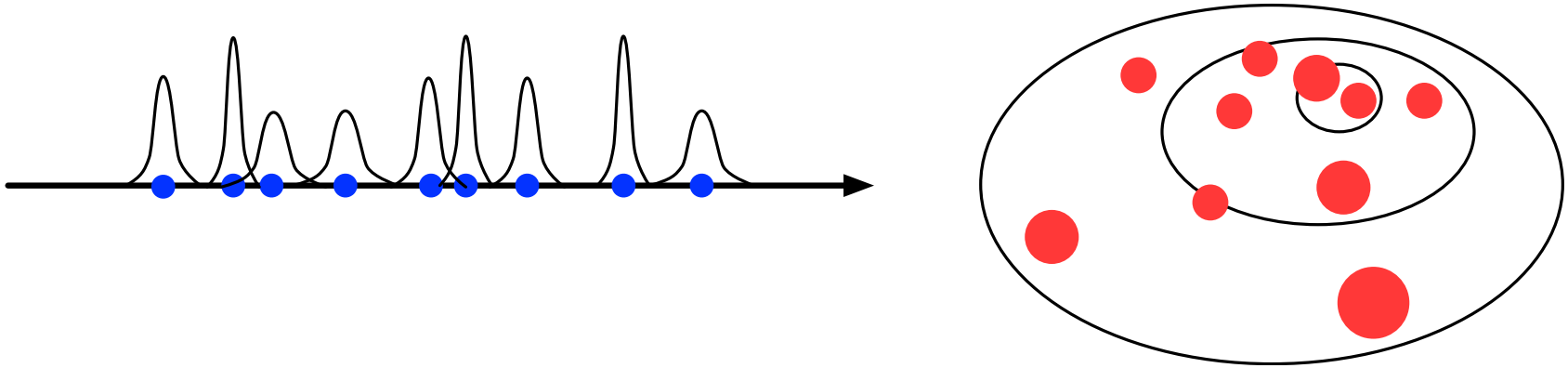
Joint distribution:

$$p(\theta, \{x_i\}) \;=\; p(\theta)\, p(\{x_i\}|\theta) \;=\; p(\theta) \prod_i f(x_i|\theta)$$

Posterior from BT:

$$p(\theta|\{x_i\}) = \frac{p(\theta, \{x_i\})}{p(\{x_i\})}$$

But what if the x data are *noisy*, $D_i = \{x_i + \epsilon_i\}$?



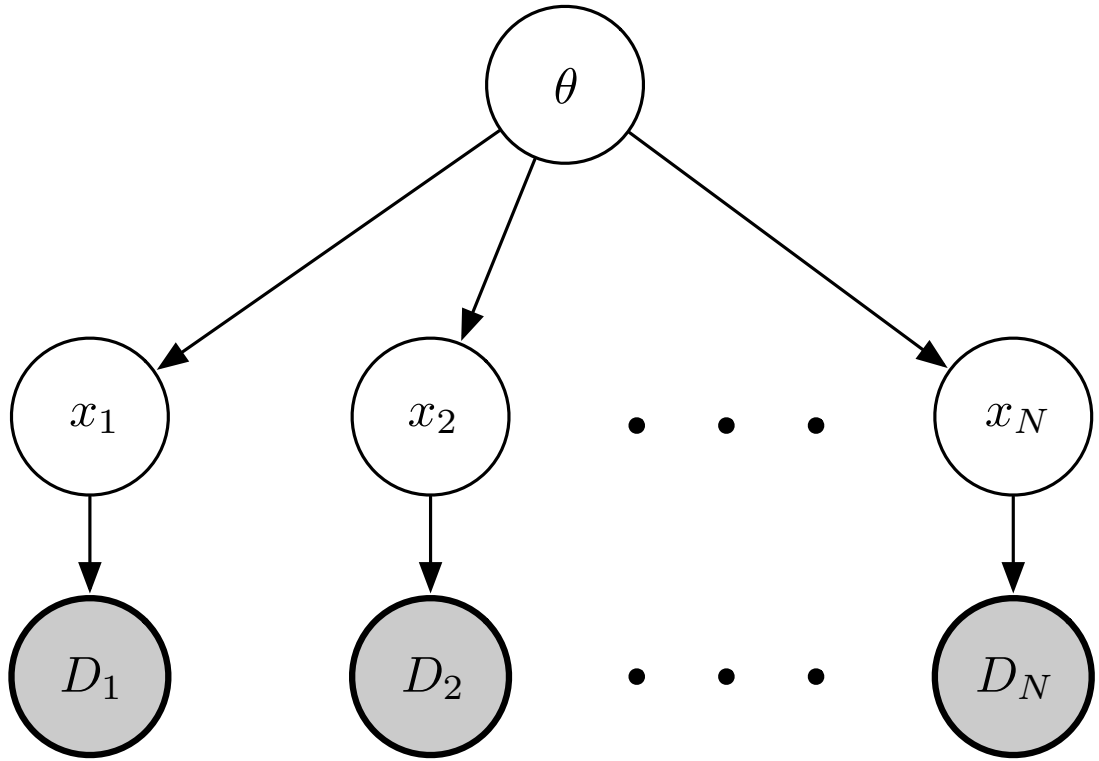
$\{x_i\}$ are now *uncertain (latent) parameters*
We should somehow incorporate $\ell_i(x_i) = p(D_i|x_i)$:

$$\begin{aligned} p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\, p(\{x_i\}|\theta)\, p(\{D_i\}|\{x_i\}) \\ &= p(\theta) \prod_i f(x_i|\theta)\, \ell_i(x_i) \end{aligned}$$

with *member likelihood functions* $\ell_i(x_i)$

Marginalize over $\{x_i\}$ to summarize inferences for θ.
Marginalize over θ to summarize inferences for $\{x_i\}$.

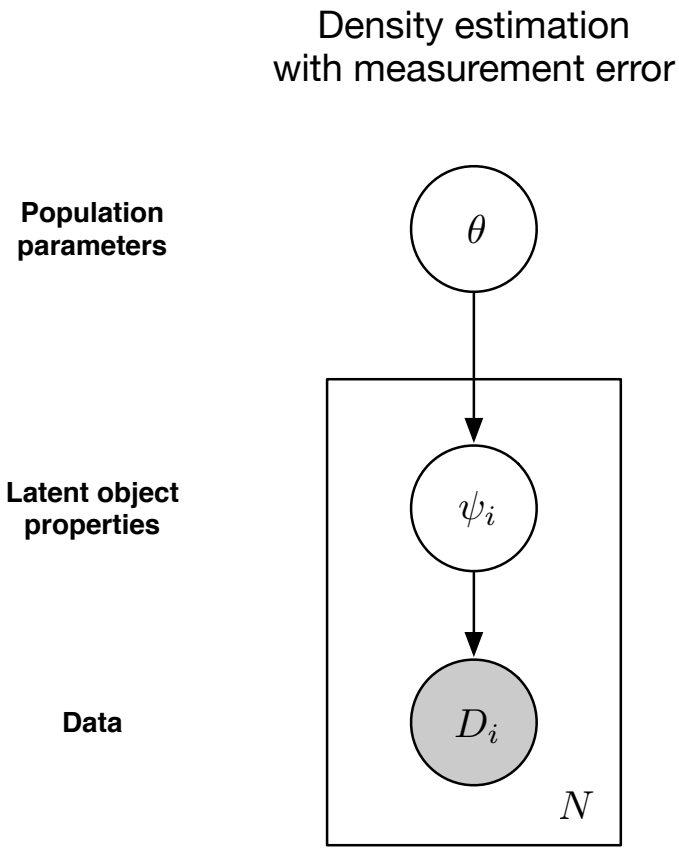
Graphical representation



$$
\begin{aligned}
p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\; p(\{x_i\}|\theta)\; p(\{D_i\}|\{x_i\}) \\
&= p(\theta) \prod_i f(x_i|\theta)\; p(D_i|x_i) \;=\; p(\theta) \prod_i f(x_i|\theta)\; \ell_i(x_i)
\end{aligned}
$$

(sometimes called a "two-level MLM" or "two-level hierarchical model")

Basic cosmic demography

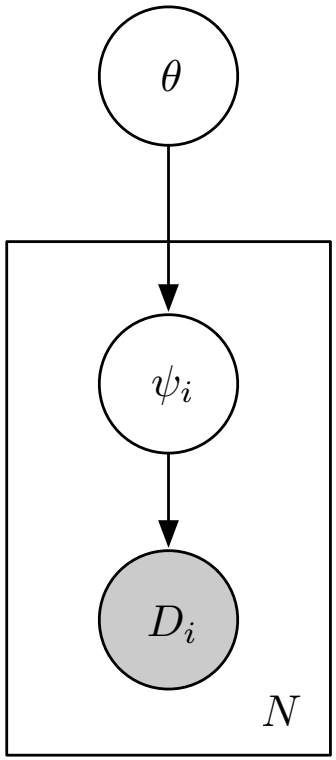


For $\psi_i = F_i$: Number-size dist'n/number counts/log N–log S

For $\psi_i = (z_i, F_i)$: Luminosity function

For $\psi_i = (M_i, R_i)$: Exoplanet mass-radius dist'n...

Joint dist'n and conditional independence



A graphical (multilevel/hierarchical) model specifies the joint dist'n for data and parameters (pop'n and latent). Here $\psi = \{\psi_i\}$:

$$\begin{aligned} p(\theta, \psi, D) &= p(\theta) \prod_{i=1}^{N} p(\psi_i|\theta)\, p(D_i|\psi_i) \qquad || \; M \\ &\propto \pi(\theta) \prod_{i} f(\psi_i; \theta)\, \ell_i(\psi_i) \end{aligned}$$

with *member likelihood functions* (not PDFs!)

$$\ell_i(\psi_i) \;\propto\; p(D_i|\psi_i)$$

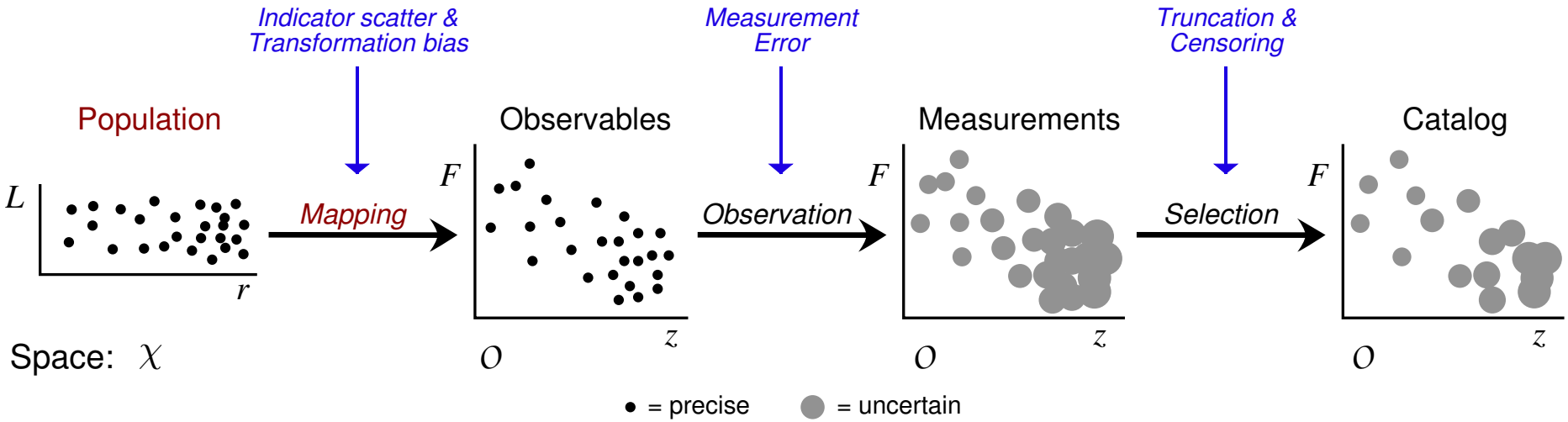
Bayes's theorem gives the posterior for all params:

$$p(\theta, \psi|D) \;=\; \frac{p(\theta, \psi, D)}{p(D)} \;\propto\; p(\theta, \psi, D)$$

Menu

1. What/Why/How of cosmic demographics
2. Basic Bayesian inference (briefly!)
3. Bayesian computation (also briefly!)
4. Hierarchical Bayesian modeling
5. **Selection effects: Thinned latent point processes**

The surveying process

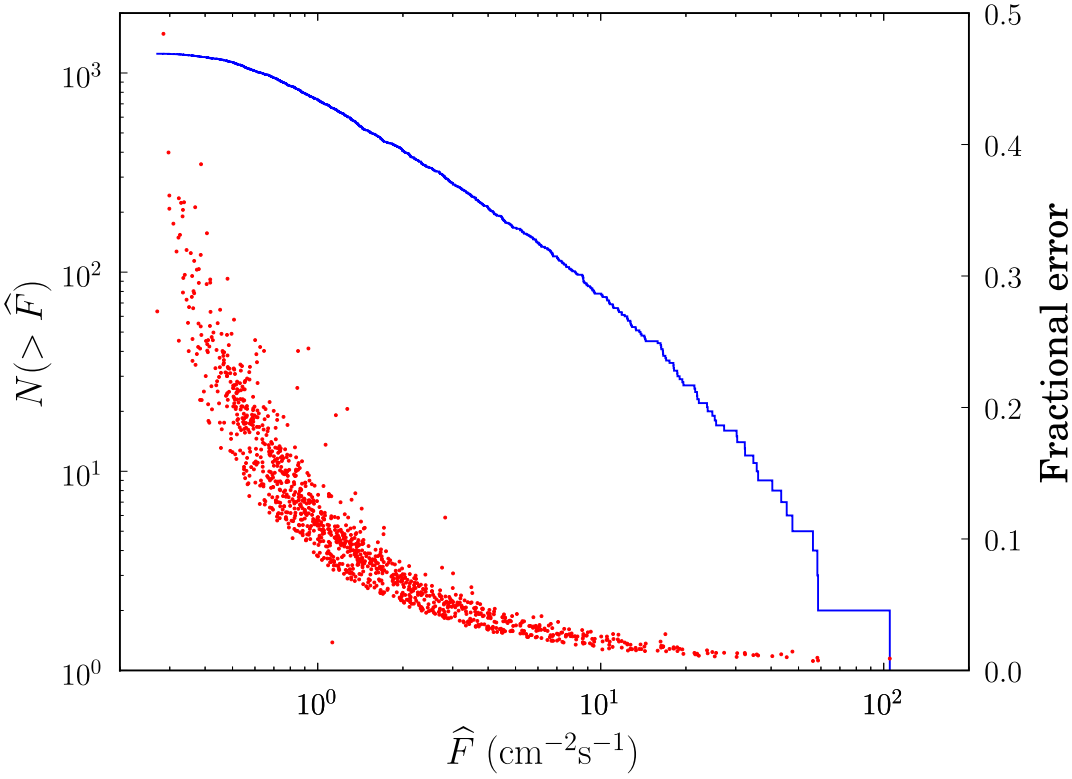


For simplicity we here ignore *contamination*

Important features

- Catalogs provide summaries of the actual raw survey data, produced via nontrivial analysis pipelines
- Detection is typically accomplished via some kind of scanning procedure
- The same survey data is used both for object detection, and for characterization (estimating object characteristics)

Survey distortions



- *"Scatter" effects* (measurement error, etc.) — *insidious*
  Typically ignored ("they'll average out" — *NOT!*)
- *Selection effects* (truncation, censoring) — *obvious* (usually)
  Typically treated by "debiasing" data
  Most sophisticated: product-limit estimators
  These *don't work* when there is also measurement error

Selection effects

Selection effects enter via *detection criteria*

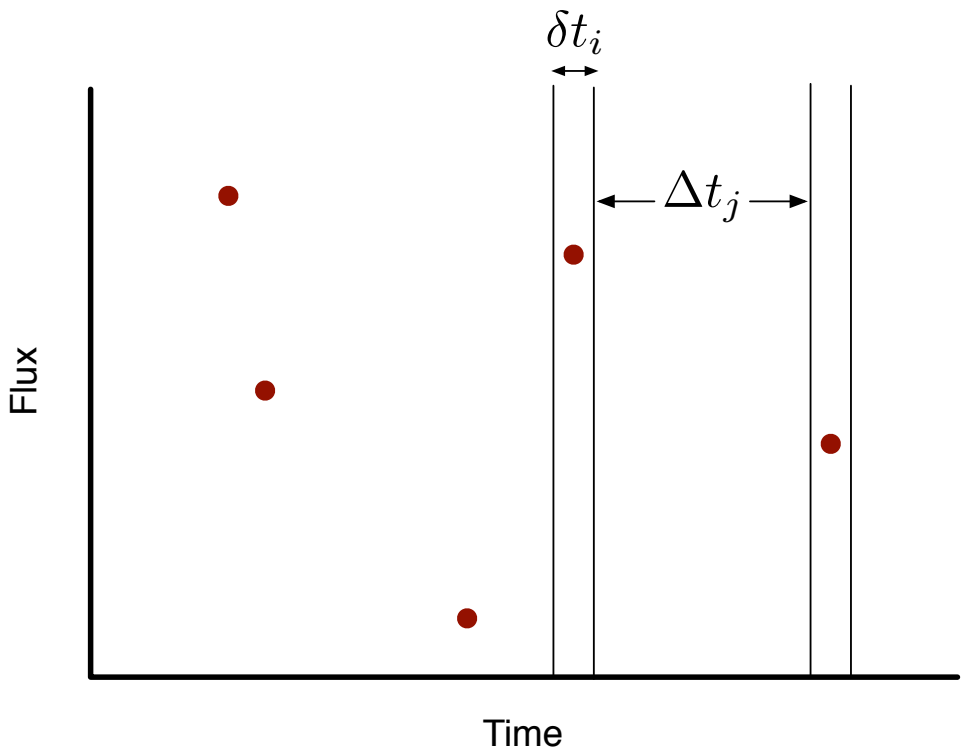
Detection is typically done by *scanning* over some domain:

- Imaging surveys: Scan an aperture or candidate source location over 2-D sky direction (image coordinates)
- Transient surveys: Scan a time window over candidate event times

To account for selection effects, explicitly model the scan process

Ideal survey: Marked Poisson point process

Data = Precise detections $\{t_i, F_i\}$ *and* non-detection info $\{\Delta t_j\}$



GRB rate: $R\rho(F;\mathcal{P})$

Non-detections:
$p_j = e^{-R\Delta t_j}$

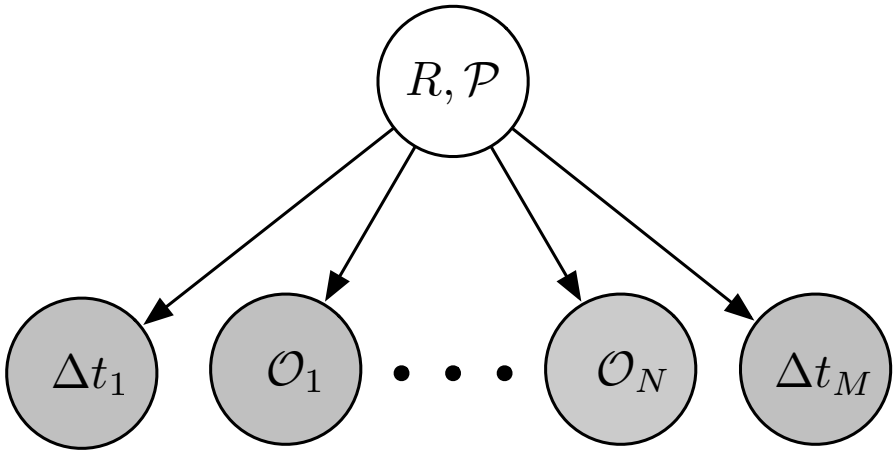
Detections:
$p_i = (R\delta t_i)\, e^{-R\delta t_i}\rho(F_i)$

Total duration T

$$\mathcal{L}(R,\mathcal{P}) \equiv p(D|R,\mathcal{P}) = e^{-RT} \times \prod_i (R\delta t)\rho(F_i;\mathcal{P})$$

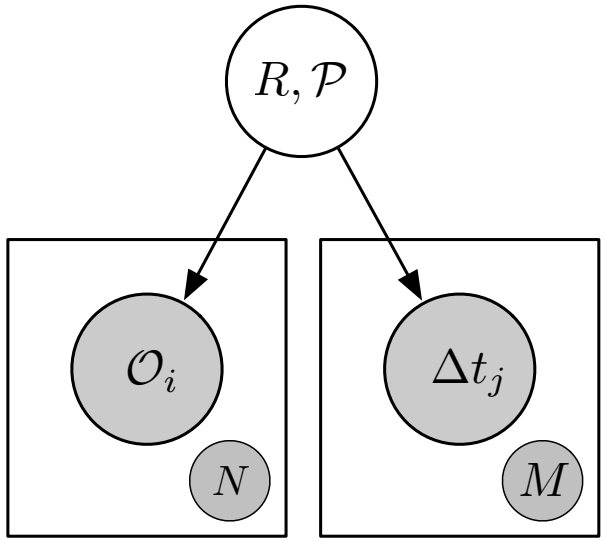
Likelihood for a *marked nonhomogeneous Poisson point process*

Graphical model



Population parameters

Precisely measured source observables
$\mathcal{O}_i = t_i, F_i, \vec{n}_i, \ldots$

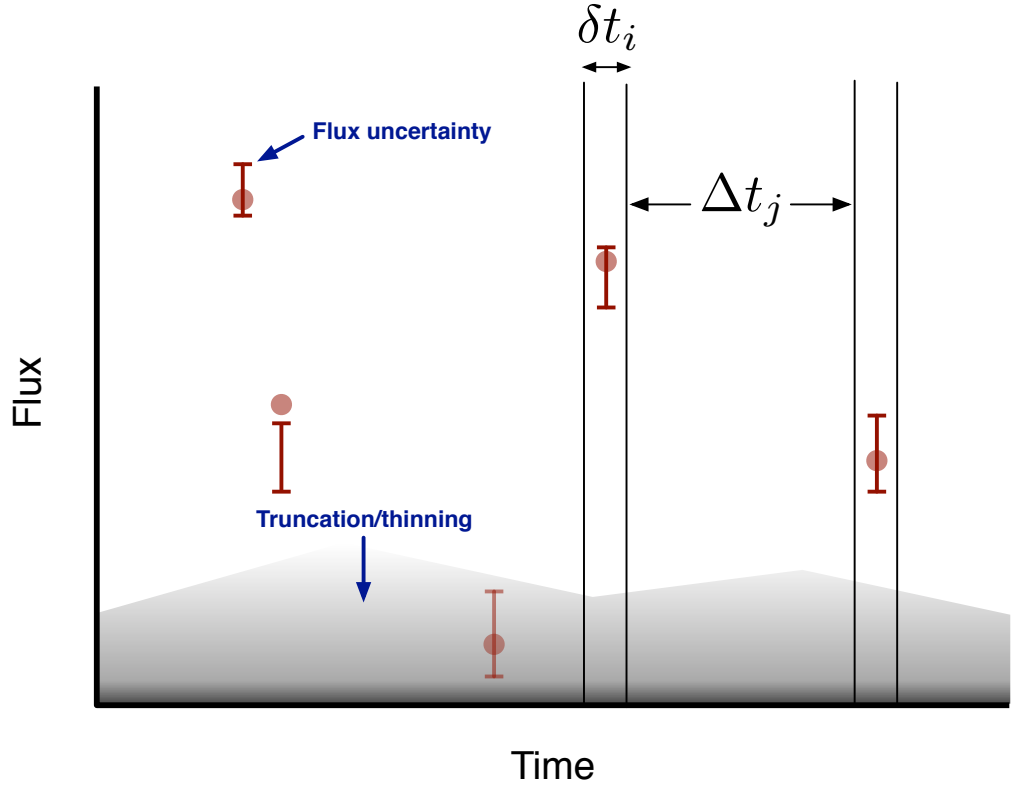


Data $D = \{\mathcal{O}_i\}, \{\Delta t_j\}$

$$p(R, \mathcal{P}, D) = p(R, \mathcal{P}) \times \left[\prod_i p(\mathcal{O}_i | R, \mathcal{P})\right] \times \left[\prod_j p(\Delta t_j | R, \mathcal{P})\right]$$

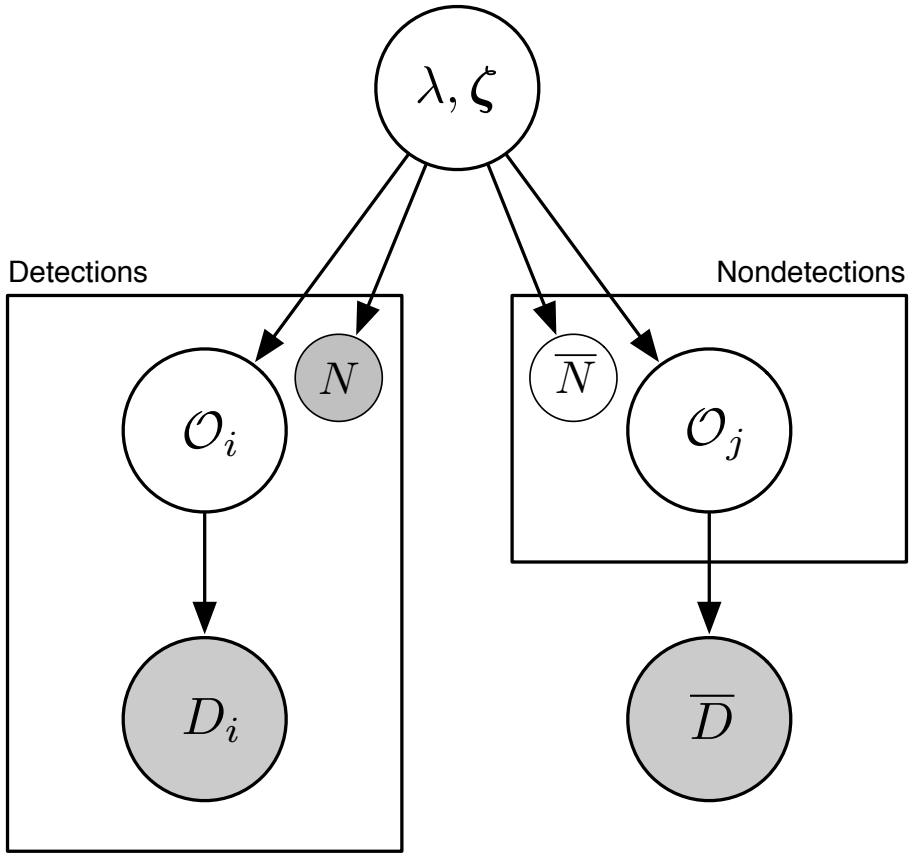
Actual survey: Thinned latent point process

Data = Detection times and *flux likelihoods*, $\ell_i(F)$ (Gaussians)
+ exposure/efficiency (from *threshold history*)



Model as a *thinned point process* in the scan space (time), with other variables as latent *marks* (flux)

Multilevel graphical model



Population parameters

Latent object observables

$\mathcal{O}_i = x_i, F_i, r_i, \ldots$

Data

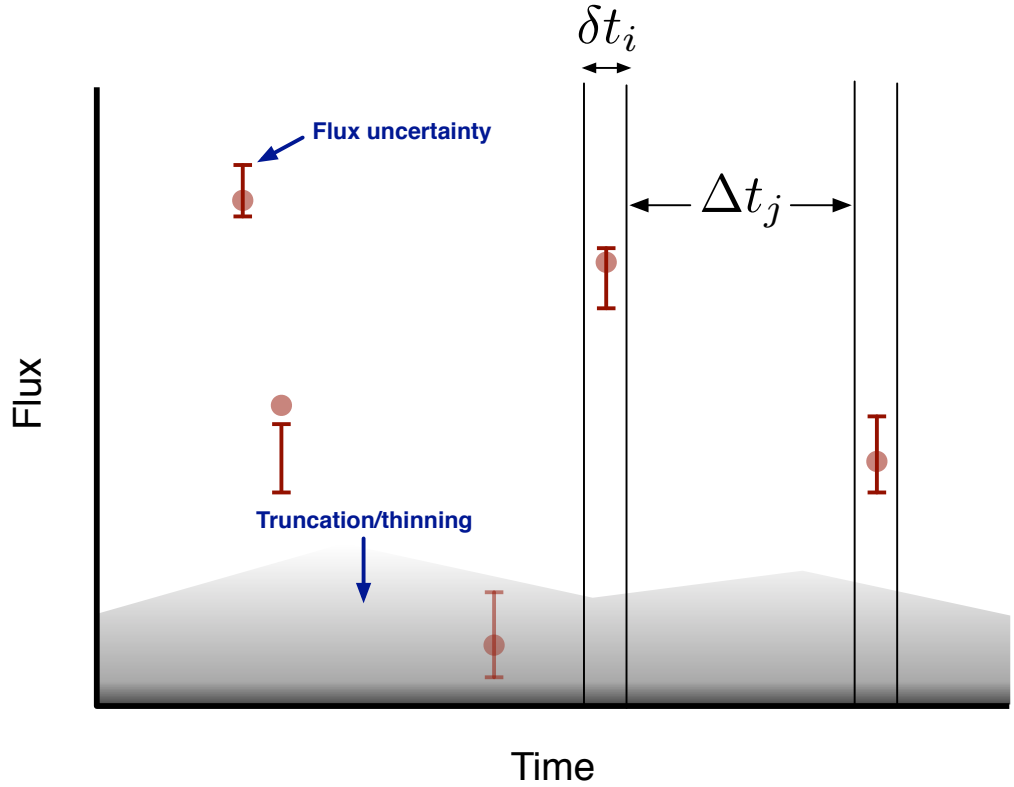
Summarize via source likelihoods

$\ell_i(\mathcal{O}_i) \equiv p(D_i|\mathcal{O}_i)$

Summarize via exposure/efficiency

$\eta(\mathcal{O})$

Actual survey: Thinned latent point process



$$p(R, \mathcal{P}, \{F_i\}, D) = p(R, \mathcal{P})\, e^{-\mu(R,\mathcal{P})} \prod_i (R\delta t)\ell_i(F_i)\rho(F_i; \mathcal{P})$$

$$\mu(R, \mathcal{P}) \equiv \int_T dt \int dF\ \eta(t, F)\, R\, \rho(f; \mathcal{P}); \qquad \ell_i(F_i) \equiv p(D_i|F_i)$$

Benefits and requirements of HB cosmic demography

Benefits

- Selection effects quantified by *non-detection data*
  - *vs.* $V/V_{\max}$ *and "debiasing" approaches*
- Source uncertainties propagated via *marginalization*
  - *Adaptive generalization of Eddington/Malmquist "corrections"*
  - *No single adjustment addresses source & pop'n estimation*

Requirements

- Data summaries for non-detection intervals (exposure, efficiency)
- *Likelihood functions* (*not* posterior PDFs) for detected source characteristics
  (Perhaps a role for *interim priors*)