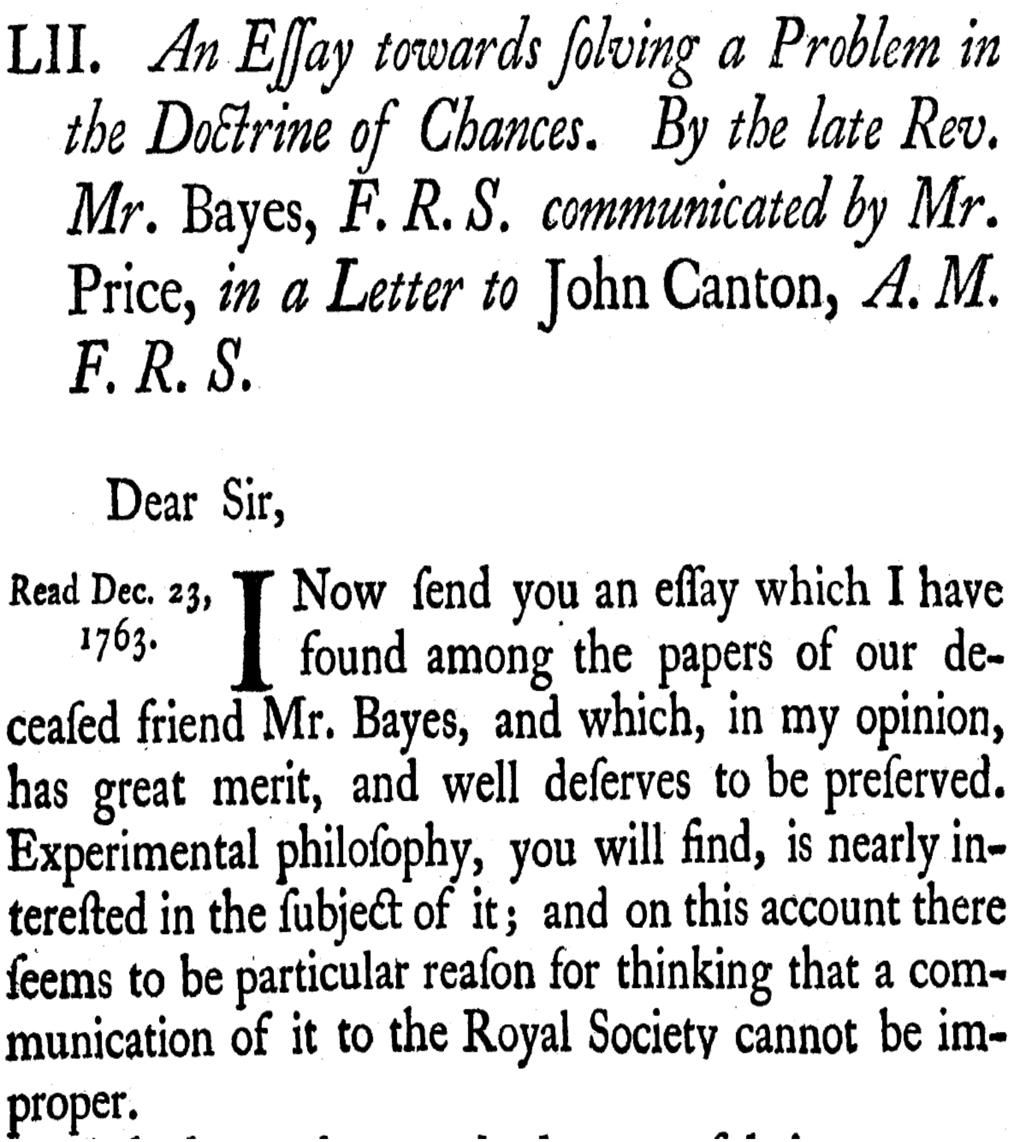
Bayesian inference in astronomy: past, present and future.

Sanjib Sharma
(University of Sydney)

January 2020

Past

Story of Mr Bayes: 1763



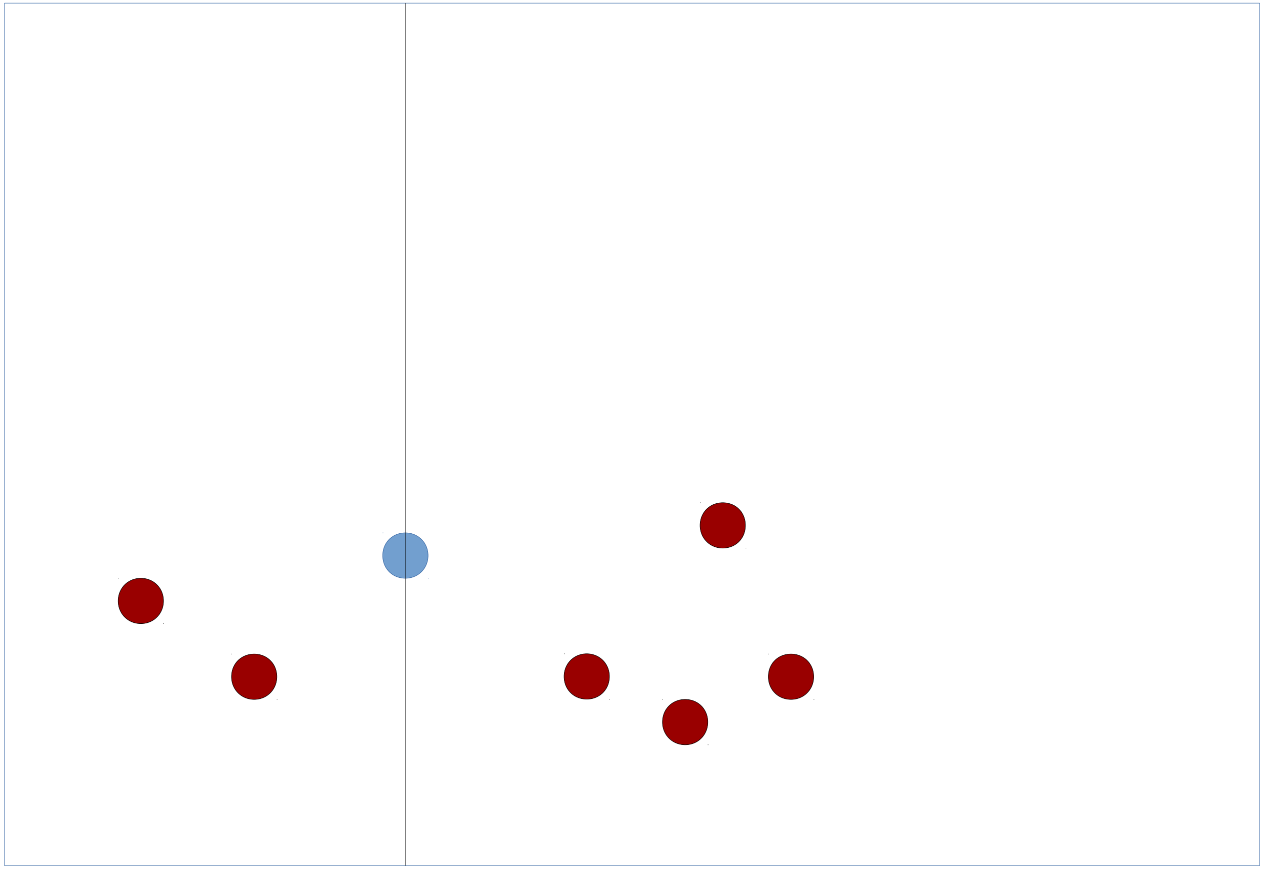
LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

Bayes problem.

- Location of blue ball based on how many balls are to right and how many to left.



Bernoulli trial problem

- A baised coin
  - If probability of head in a single trial is p.
  - What is the probability of k heads in n trials.
  - $P(k|p, n) = C(n, k)\, p^{k} (1-p)^{n-k}$

- The inverse problem
  - If k heads are observed in n trials.
  - What is the probability of occurence of head in a single trial.
    - $P(p|n, k) \sim P(k|n, p)$
    - P(Cause|Effect) ~ P(Effect| Cause)

Laplace 1774

- Independetly rediscoverded.

- In words rather than Eq, “Probability of a cause given an event /effect is proportional to the probability of the event given its cause”.
  - P(Cause|Effect) ~ P(Effect| Cause), $p(\theta|D) \sim p(D|\theta)$
  - Consider values for different θ then it becomes a dist.
  - Important point is LHS is conditioned on data.

- His friend Bouvard used his method to calculate the masses of Saturn and Jupiter.

- Laplace offered bets of 11000 to 1 odd and 1million to 1 that they were right to 1% for Saturn and Jupiter.
  - Even now Laplace would have won both bets.

1900-1950

- Largely ignored after Laplace till 1950.
- Theory of probability, 1939 by Harold Jeffrey
  - Main reference.
- In WW-II, used at Bletchley Park to decode German Enigma cipher.

- There were conceptual difficulties
  - Role of prior
  - Data is random or model parameter is random

1950 onwards

- Tide had started to turn in favor of Bayesian methods.
- Lack of proper tools and computational power main hindrance.
- Frequentist methods were simpler which made them popular.

Cox's Theorem: 1946

- Cox 1946 showed that sum and product rule can be derived from simple postulates. The rest of Bayesian probability follows from these two rules.

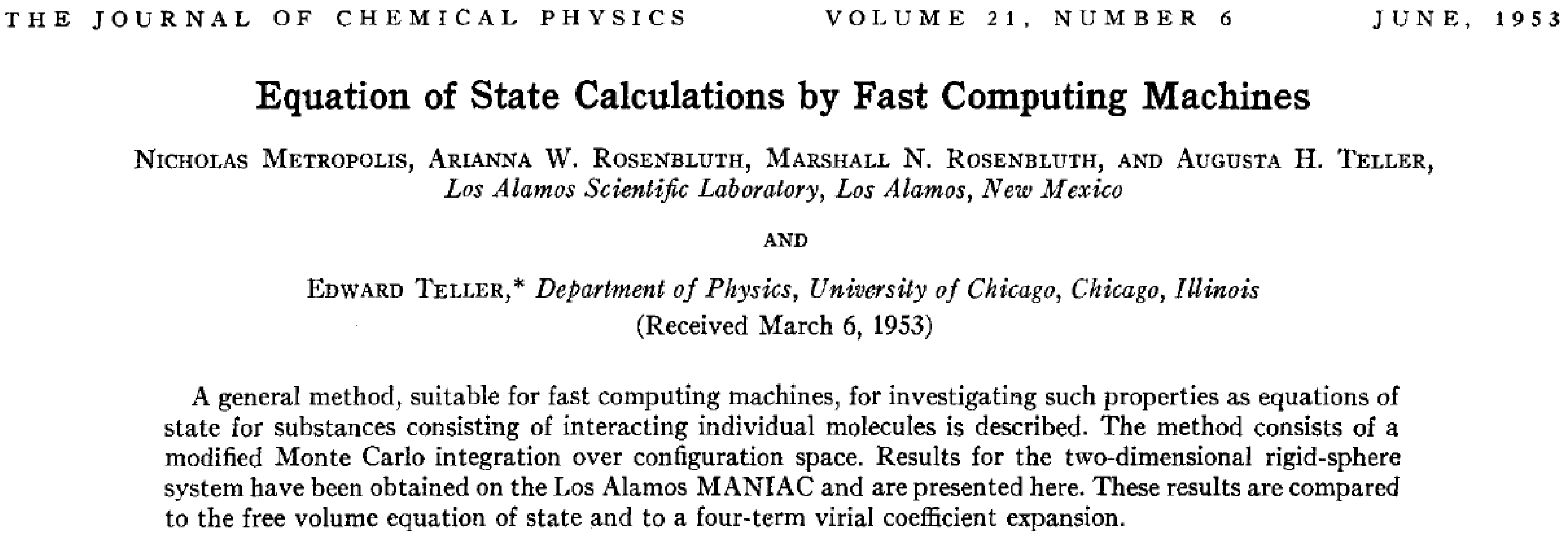
$$p(H\,|I) + p(\bar{H}\,|I) = 1 \qquad \text{Sum Rule,}$$

$$p(H, D|I) = p(H\,|D, I)p(D|I) = p(D|H, I)p(H\,|I) \qquad \text{Product Rule.}$$

$$p(H\,|D, I) = \frac{p(D|H, I)p(H\,|I)}{p(D|I)}, \qquad \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}},$$

$p(\theta|x) \sim p(x|\theta)p(\theta)$

Metropolis algorithm: 1953



THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

Who did what?

- Metropolis only was only responsible for providing computational time.
- Marshall Rosenbluth provided the solution to the problem
- Arianna Rosenbluth wrote the code.

Metropolis algorithm: 1953

- N interacting particles.
- A single configuration **ω**, can be completely specified by giving position and velocity of all the particles.
  - A point in R^{2N} space.
- **E(ω)**, total energy of the system
- For system in equilibrium **p(ω) ~ exp (- E(ω) / kT)**
- Computing any thermodynamic property, pressure, energy etc, requires integrals,which are analytically intractable

$$\bar{F} = \frac{\int F(\omega) \exp[-E(\omega)/kT]\mathrm{d}\omega}{Z},$$

- Start with arbitrary config N particles.
- Move each by a random walk and compute ΔE the change in energy between old and new config
- If: **ΔE < 0**, always accept.
- Else: accept stochastically with probability **exp (- ΔE / kT)**
- Immediate hit in statistical physics.

Hastings 1970

- The same method can be used to sample an arbitrary pdf **$p(\omega)$**
  - by replacing **$E(\omega)/kT \rightarrow -\ln p(\omega)$**
  - Had to wait till Hastings
- Generalized the algorithm and derived the essential condition that a Markov chain out to satisfy to sample the target distribution.
- Acceptance ratio not uniquely specified, other forms exist.
- His student Peskun 1973 showed that Metropolis gives the fastest mixing rate of the chain

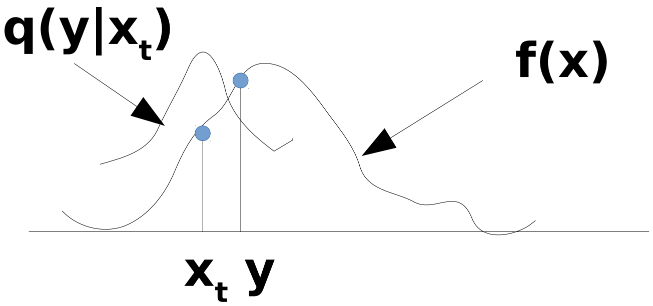
1980

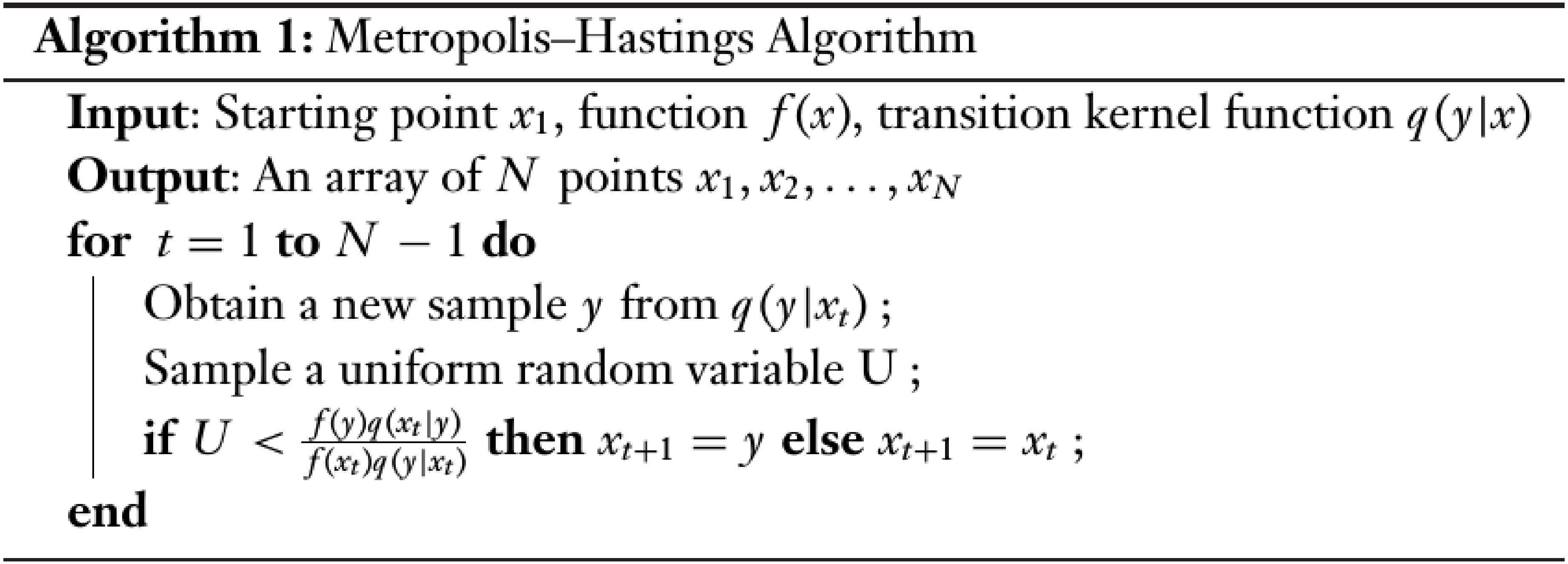
- Simulated annealing Kirkpatrick 1983
  - To solve combinatorial optimization problems using MH algorithm using ideas of annealing from solid state physics.
- Useful when we have multiple maxima and you want to
- select a globally optimum solution.
- Minimize an objective function **$C(\omega)$** by sampling from **$\exp(-C(\omega)/T)$** with progressively decreasing T.
-

1984

- Expectation Maximization (EM) algorithm
  - Dempster 1977
  - Provided a way to deal with missing data and hidden variables. Hierachical Bayesian models.
  - Vastly increased the range of problems that can addressed by Bayesian methods.
  - Deterministic and sensitive to initial condition.
  - Stochastic versions were developed
  - Data augmentation, Tanner and Wong 1987
- Geman and Geman 1984
  - Introduced Gibbs sampling in the context of image restoration.
  - First proper use of MCMC to solve a problem setup in Bayesian framework.

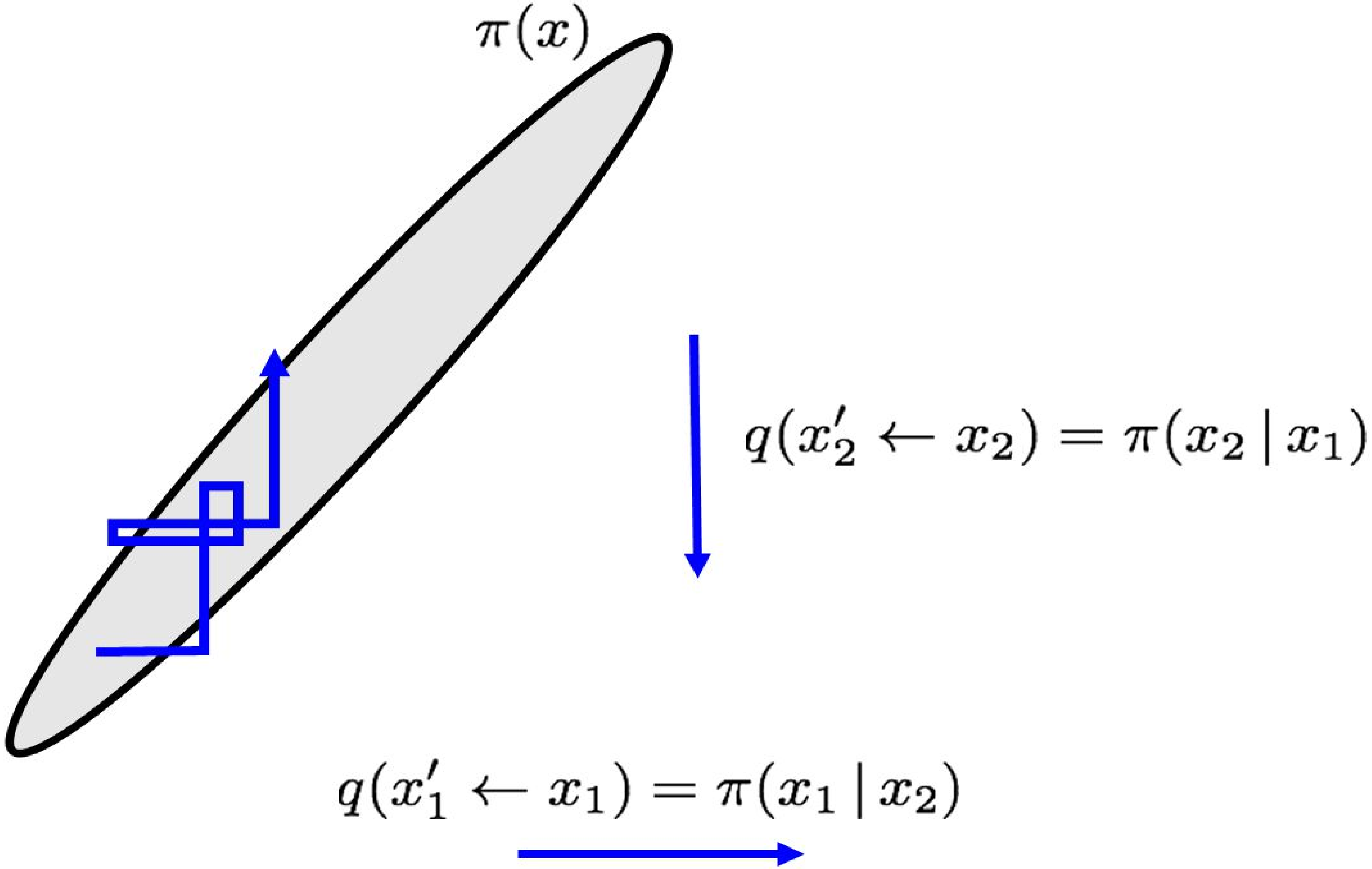
MH algorithm





```
Algorithm 1: Metropolis–Hastings Algorithm
Input: Starting point $x_1$, function $f(x)$, transition kernel function $q(y|x)$
Output: An array of $N$ points $x_1, x_2, \ldots, x_N$
for $t = 1$ to $N - 1$ do
    Obtain a new sample $y$ from $q(y|x_t)$ ;
    Sample a uniform random variable U ;
    if $U < \frac{f(y)q(x_t|y)}{f(x_t)q(y|x_t)}$ then $x_{t+1} = y$ else $x_{t+1} = x_t$ ;
end
```

Gibbs Sampling



$$q(x_2' \leftarrow x_2) = \pi(x_2 \,|\, x_1)$$

$$q(x_1' \leftarrow x_1) = \pi(x_1 \,|\, x_2)$$

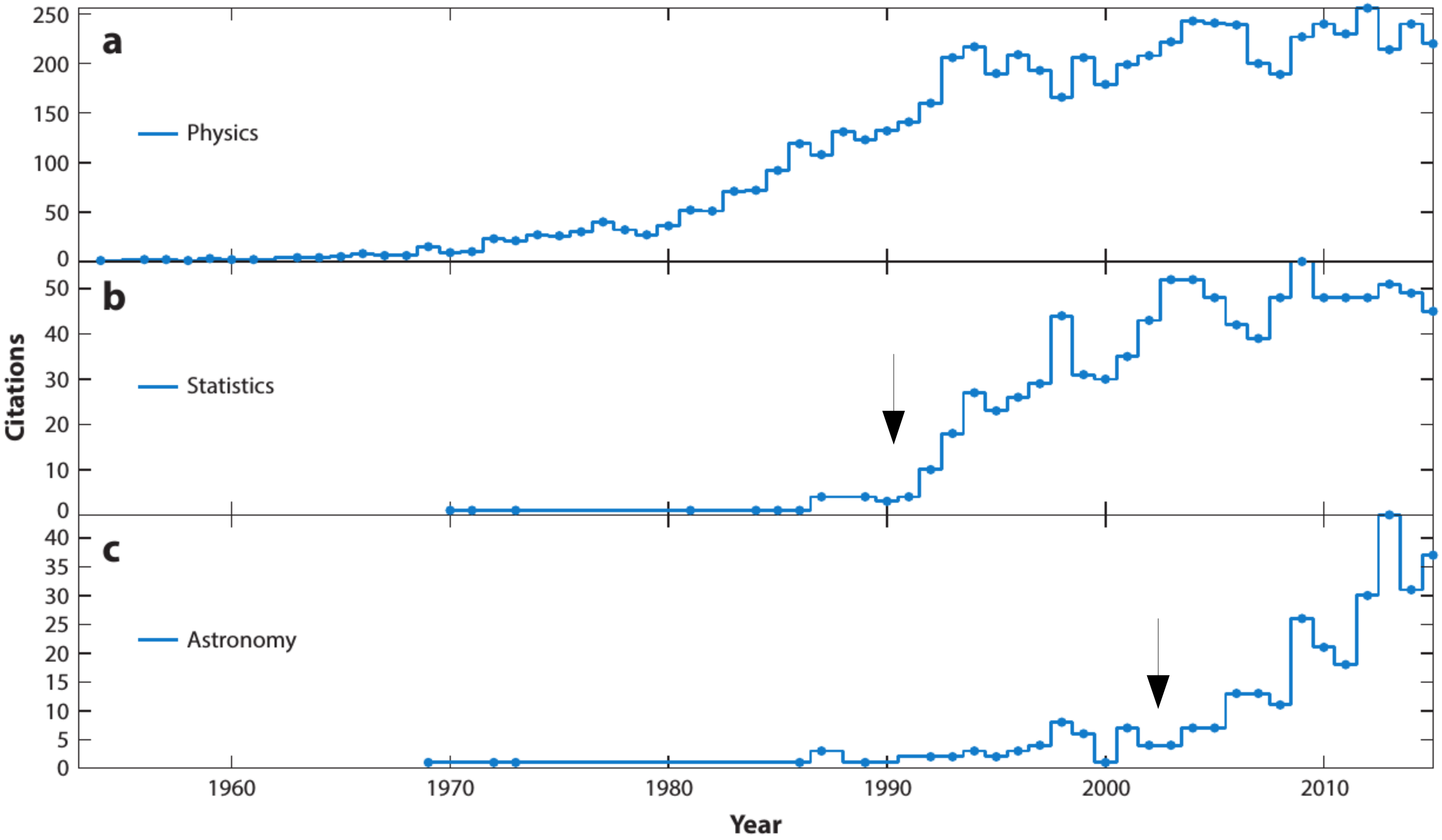
Image: Ryan Adams

1990

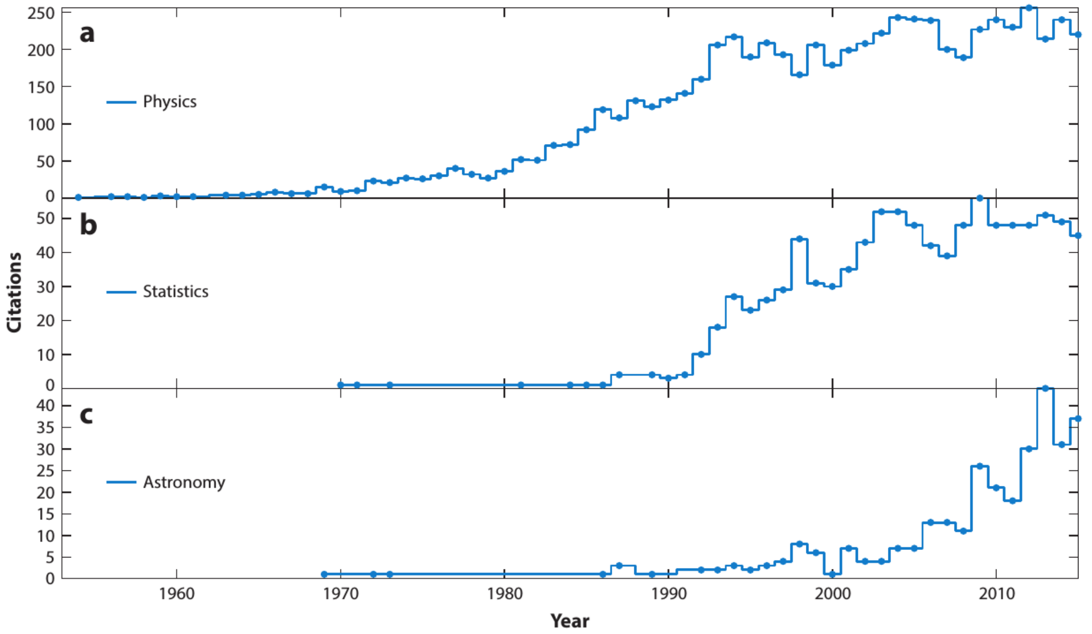
- Gelfand and Smith 1990
  - Largely credited with revolution in statistics,
  - Unified the ideas of Gibbs sampling, DA algorithm and EM algorithm.
  - It firmly established that Gibbs samling and MH based MCMC algorithms can be used to solve a wide class of problems that fall in the category of hierarchical bayesian models.
-

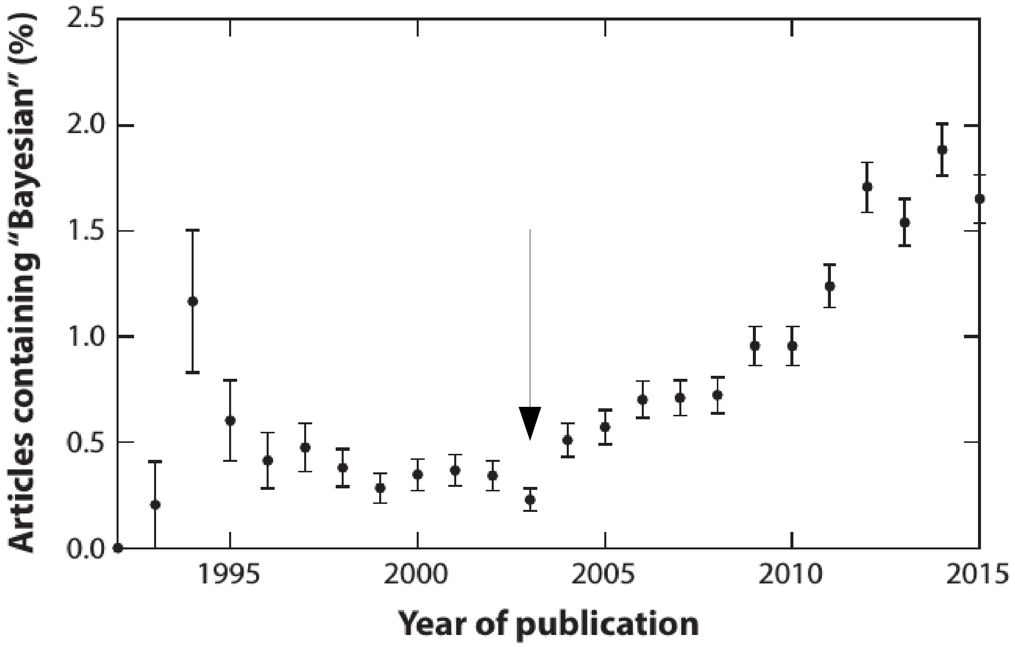
Citation history of Metropolis et al/ 1953

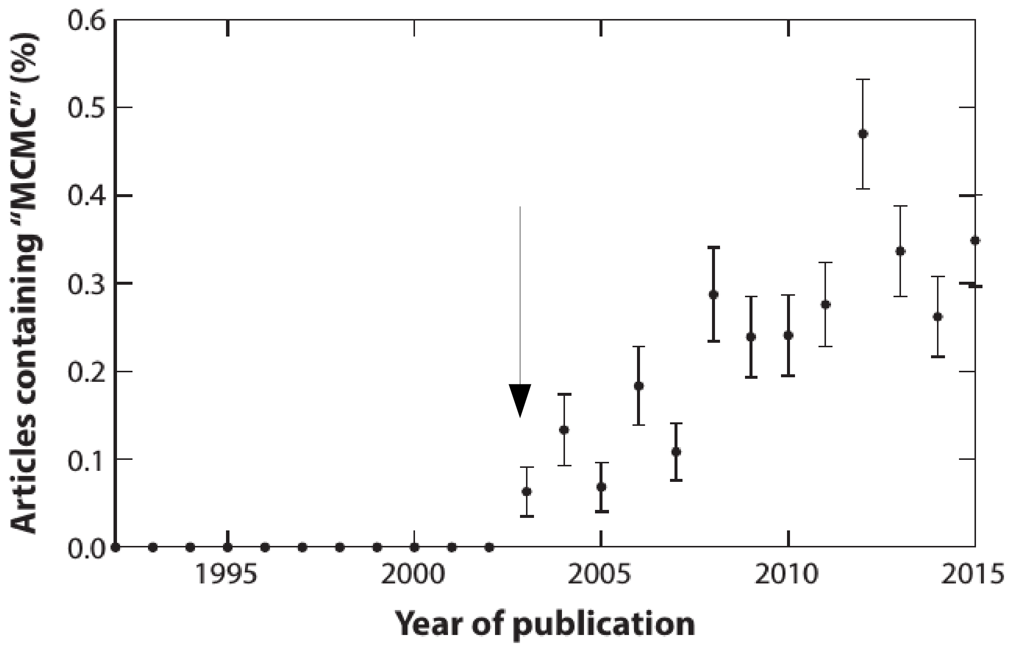
- Physics: well known from 1970-1990
- Statistics: only 1990 onwards
- Astronomy: 2002 onwards



Astronomy's conversion- 2002







Astronomy: 1990-2002

- Loredo 1990
  - Influential article on Bayesian probability theory
- Saha & Williams 1994
  - Galaxy kinematics from absorption line spectra.
- Christensen & Meyer 1998
  - Gravitational wave radiation
- Christensen et al. 2001 and Knox et al. 2001
  - Comsological parameter estimation using CMB data
- Lewis & Bridle 2002
  - Galvanized the astronomy community more than any other paper.

•Lewis & Bridle 2002

- Laid out in detail the Bayesian MCMC framework
- Applied it to one of the most important data sets of the time, the CMB data.
- Used it to address a significant scientific question- fundamanetal parameters of the universe.
- Made the code publicly available
  - Making it easier for new entrants.

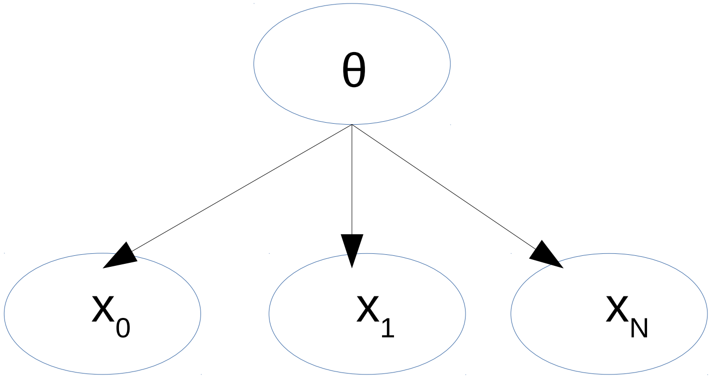
Metropolis in practise

- Requires tuning of proposal distribution
  - Too wide,
    - acceptance ratio close to zero, too many rejections, move far but rarely
  - Too small
    - acceptance ratio close to 1, move frequently but does not travel far.
- Solutions
  - Adaptive Metropolis
    - Tune based on past estimate of covariance, violates Markovian property, Trick is that adaptation becomes slow and slow with time.
  - Ensemble and affine invariant samplers

Present

Bayesian hierarchical models

- $p(\theta \mid \{x_i\}) \sim p(\theta) \prod_i p(x_i \mid \theta)$

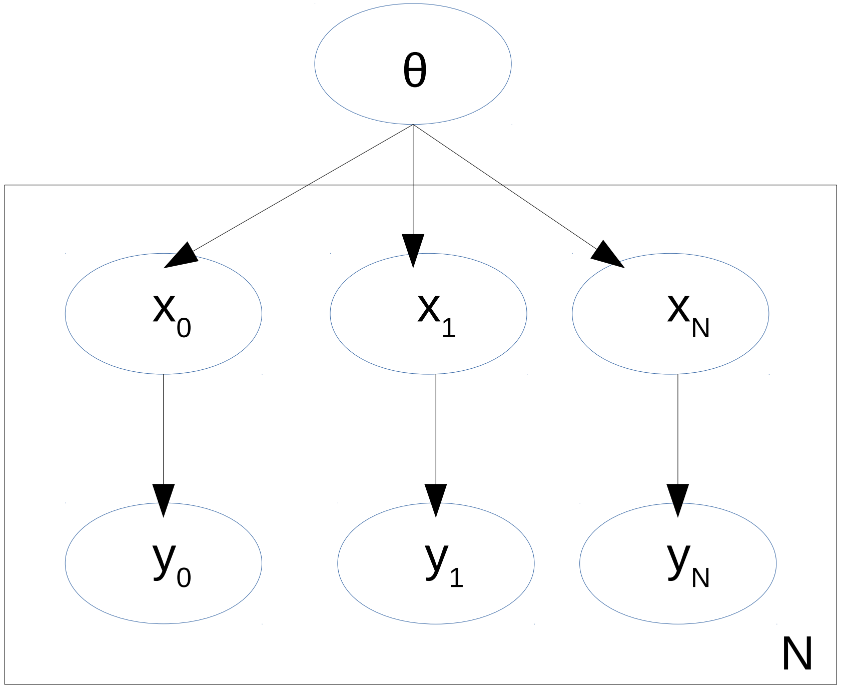


- $p(\theta, \{x_i\} \mid \{y_i\}) \sim p(\theta) \prod_i p(x_i \mid \theta)\ \ p(y_i \mid x_i, \sigma_{yi})$

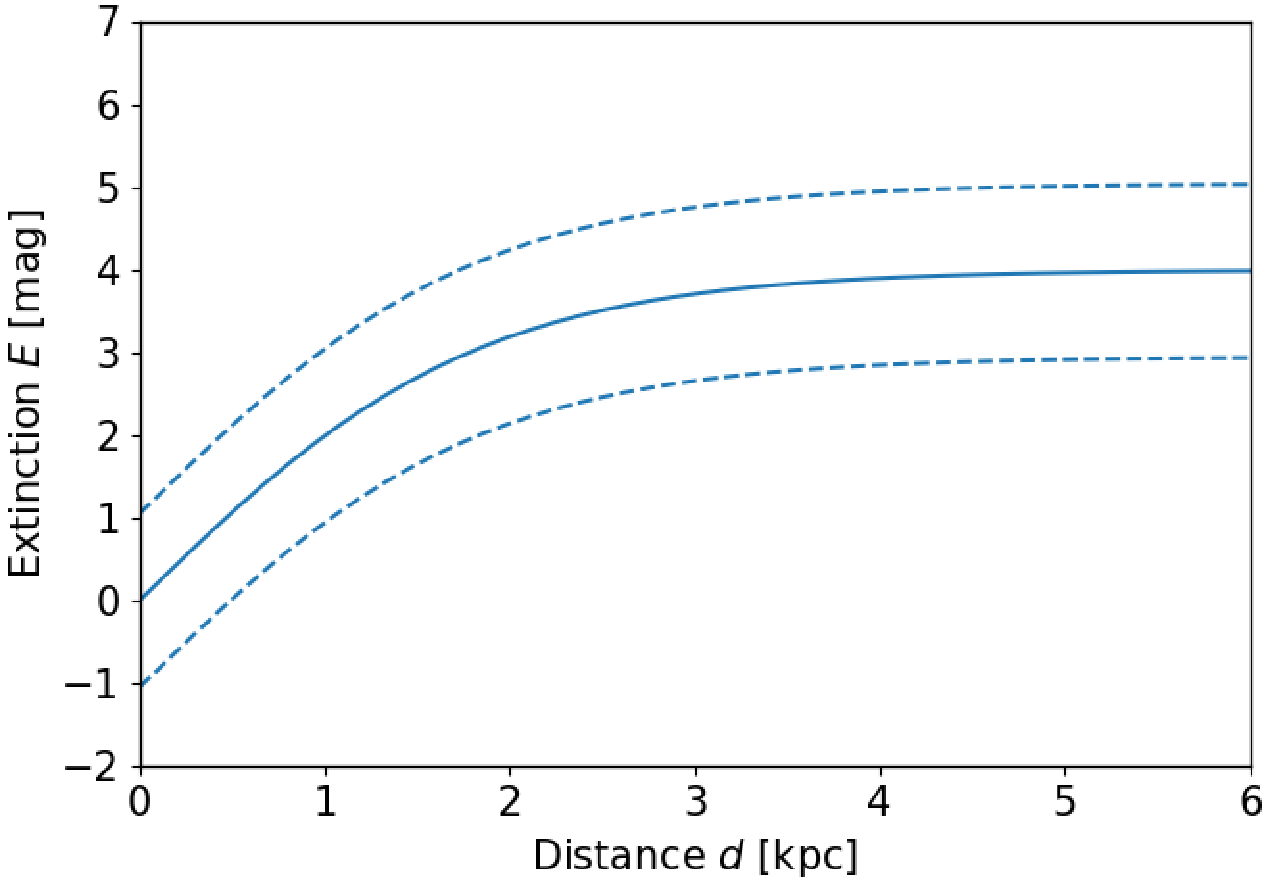
Level-0: Population →

Level-1: Individual Object-intrinsic →

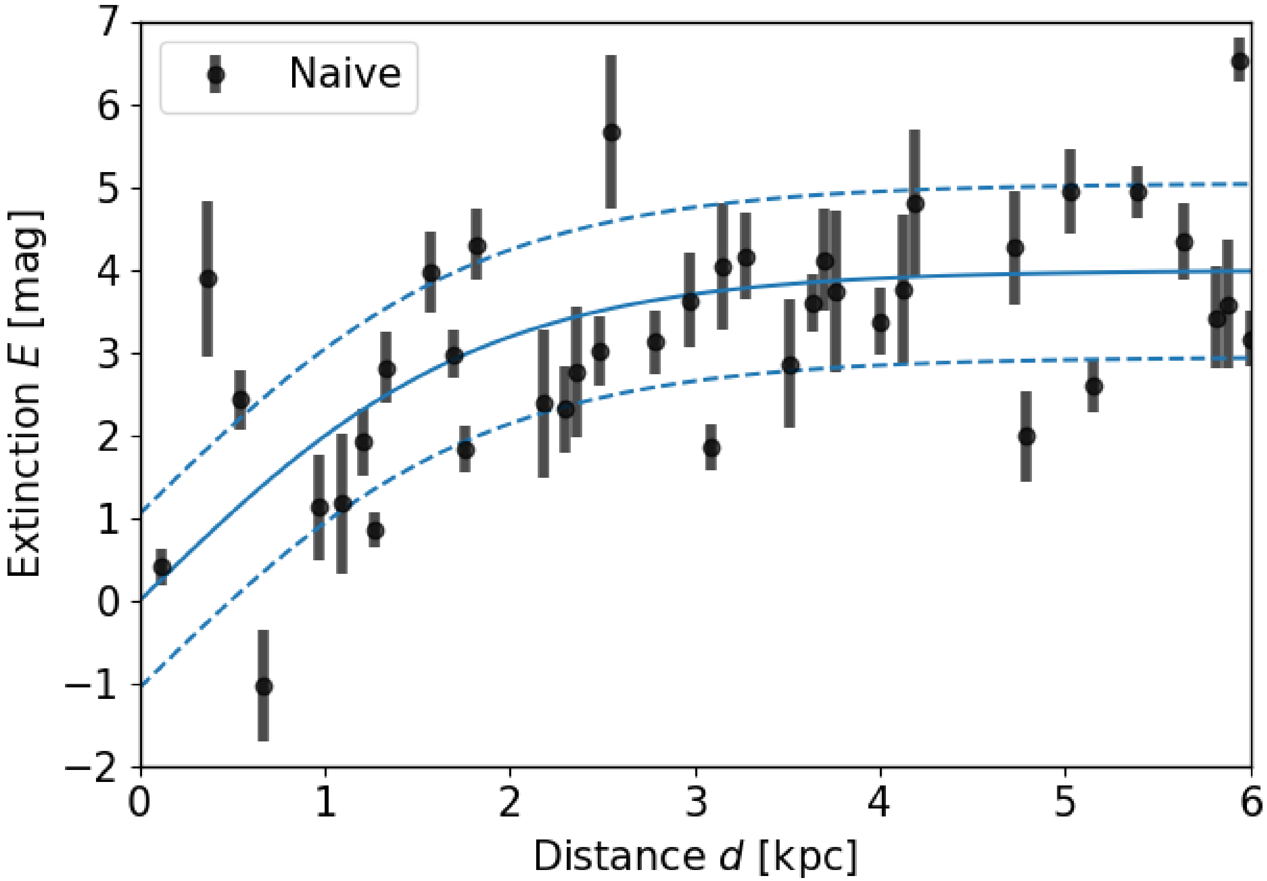
Level-1: Individual Object-observable →



Extinction of stars at various distances along a line of sight



- Each star has some some measurement with some uncertainty
  - $\mathbf{p(E_{t,j}|E_j) \sim Normal(E_j,\sigma_j)}$.

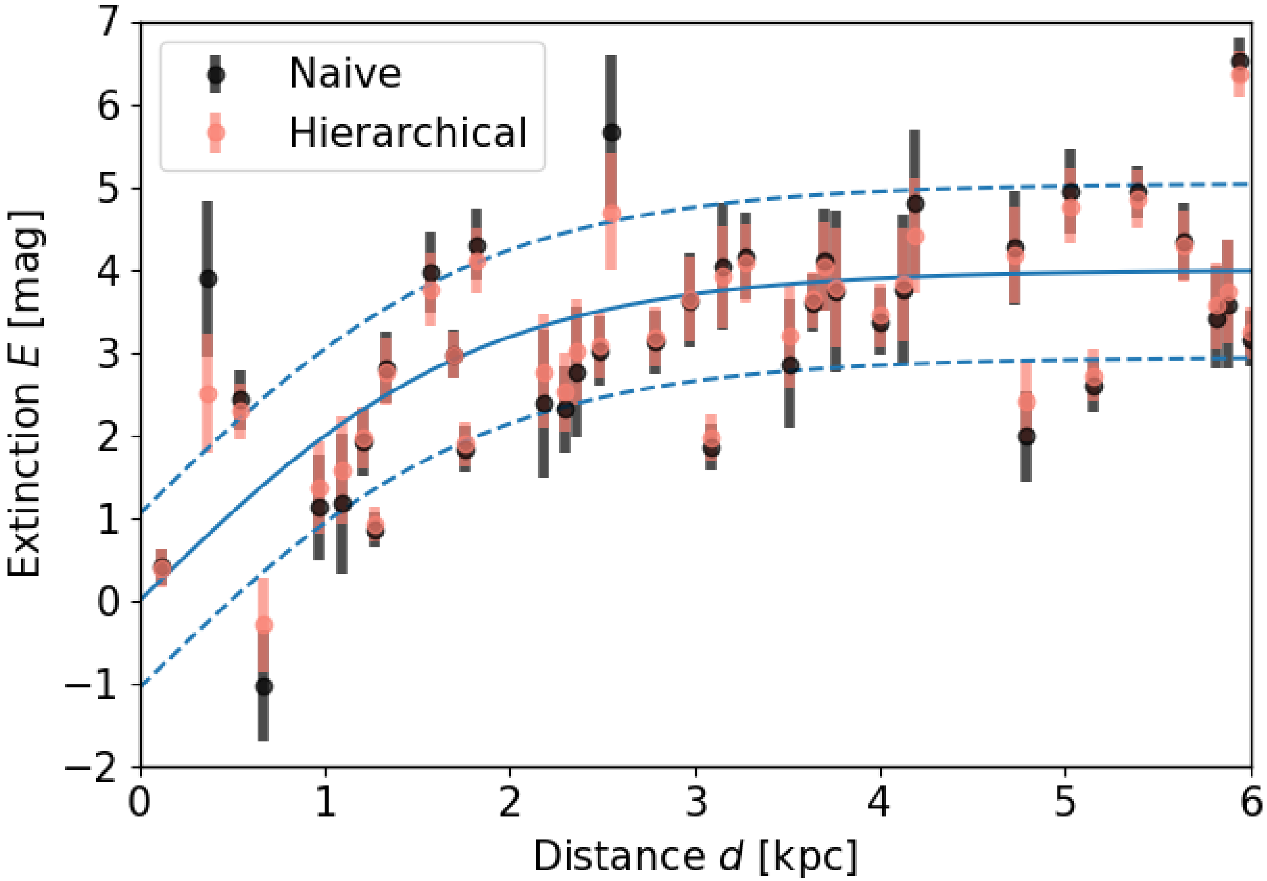


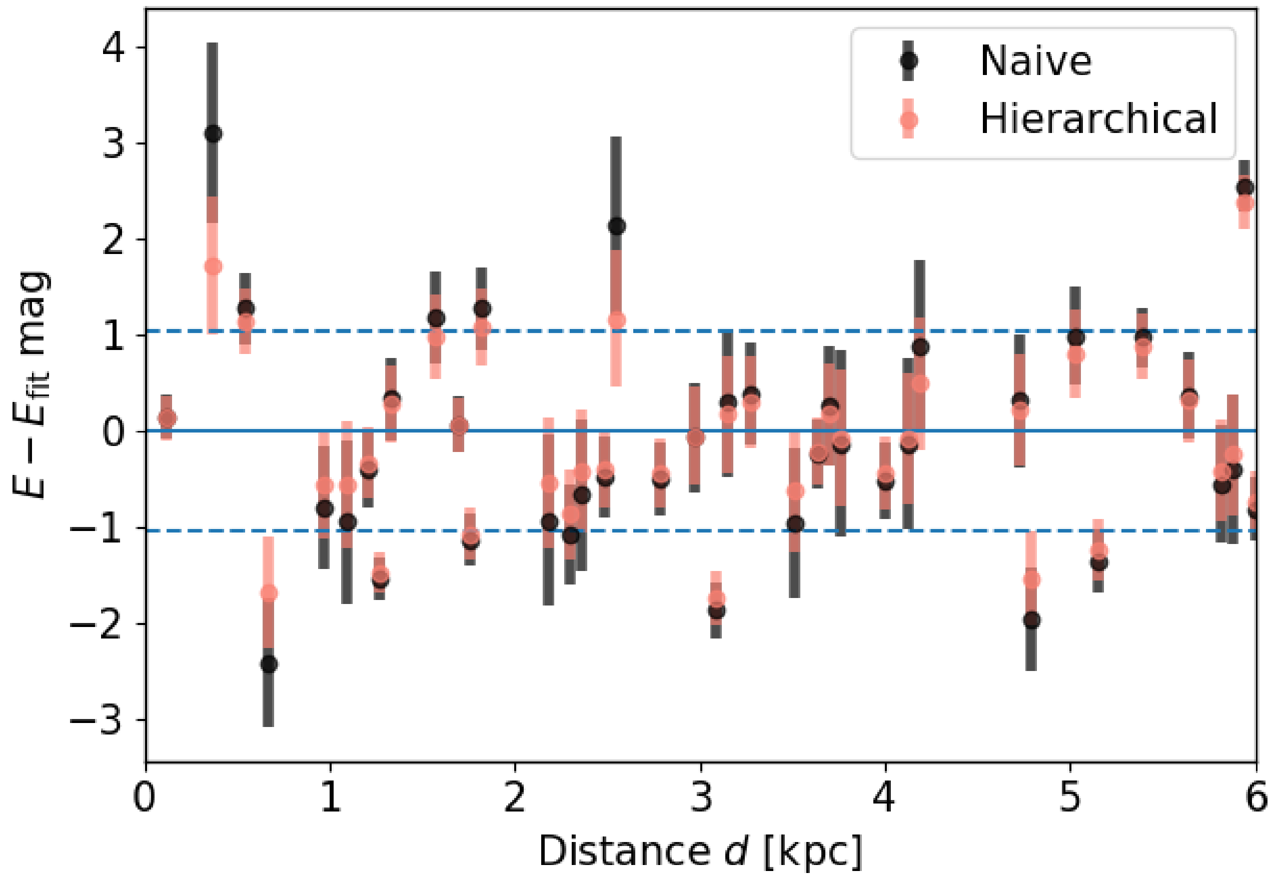
- What we want to know
  - Overall distance extinction relationship and its dispersion $\mathbf{(\alpha,E_{max},\sigma_E)}$.
  - Extinction of a star and its uncertainty $\mathbf{p(E_{t,j})}$.

BHM

- Some stars have very high uncertainty.
- There is more information in data from other stars.
  - $p(E_{t,j}|\alpha,E_{max},\sigma_E,E_j,\sigma_j) \sim p(E_{t,j}|\alpha,E_{max}\sigma_E)\ p(E_{t,j}\ |\ E,\sigma_j)$
  -
- But, population statistics depends on stars, they are interrelated.
- We get joint info about population of stars as well as for individual stars.
  - $p(\alpha,E_{max},\sigma_E, E_{t,j}|E_j,\sigma_j) \sim p(\alpha,E_{max},\sigma_E)\ \prod_j p(E_{t,j}|\ \alpha,E_{max}\sigma_E)\ p(E_{t,j}\ |\ E_j,\sigma_j)$

Shrinkage of error, shift towards mean




Naive
Hierarchical
$E - E_{\mathrm{fit}}$ mag
4
3
2
1
0
−1
−2
−3
0
1
2
3
4
5
6
Distance d [kpc]

Handling uncertainties

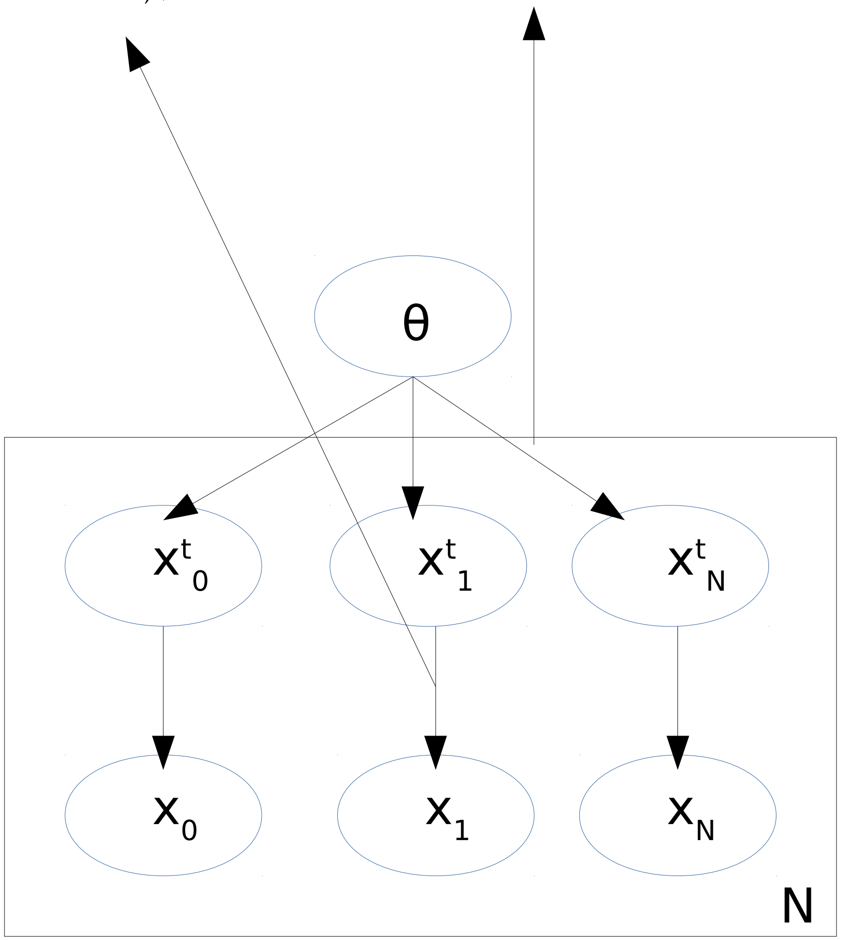
- $p(\theta, \{x^t_i\} \mid \{x_i\}, \{\sigma_{xi}\}) \sim p(\theta) \prod_i p(x^t_i \mid \theta) \; p(x_i \mid x^t_i , \sigma_{x,i})$

- $p(x_i \mid x^t_i , \sigma_{x,i}) \sim \mathrm{Normal}(x_i \mid x^t_i, \sigma_{yi})$

Level-0: Population

Level-1: Individual Object-intrinsic

Level-2: Individual Object-observable



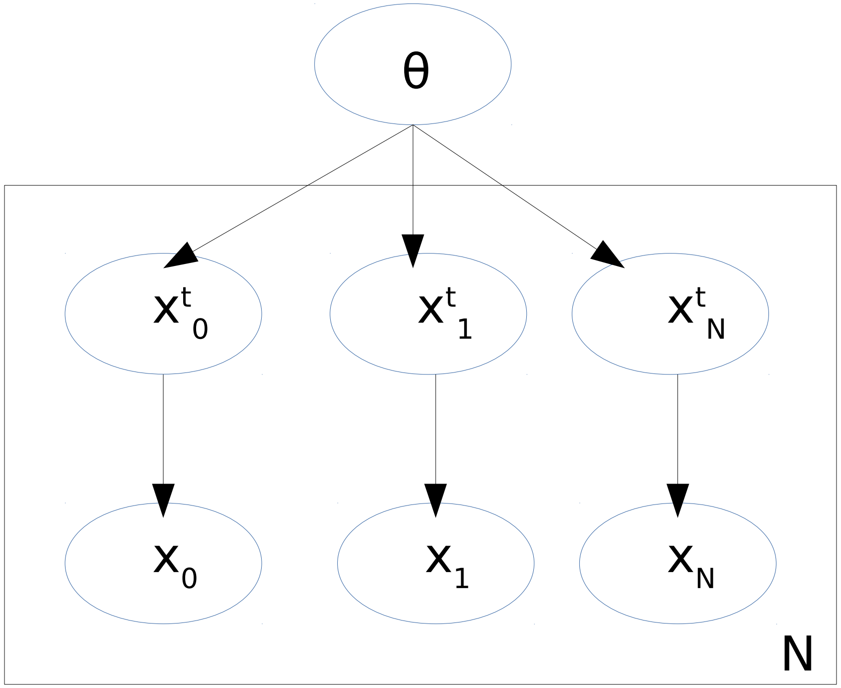
Missing variables: traditionally marginalization

- $p(\theta, \{x^t_i\} \mid \{x_i\}, \{\sigma_{xi}\}) \sim p(\theta) \prod_i p(x^t_i \mid \theta)\ p(x_i \mid x^t_i, \sigma_{x,i})$
- $p(x_i \mid x^t_i, \sigma_{x,i}) \sim \text{Normal}(x_i \mid x^t_i, \sigma_{yi})$
- Certain $\sigma_{xi} \to \infty$

Level-0: Population

Level-1: Individual Object-intrinsic

Level-2: Individual Object-observable



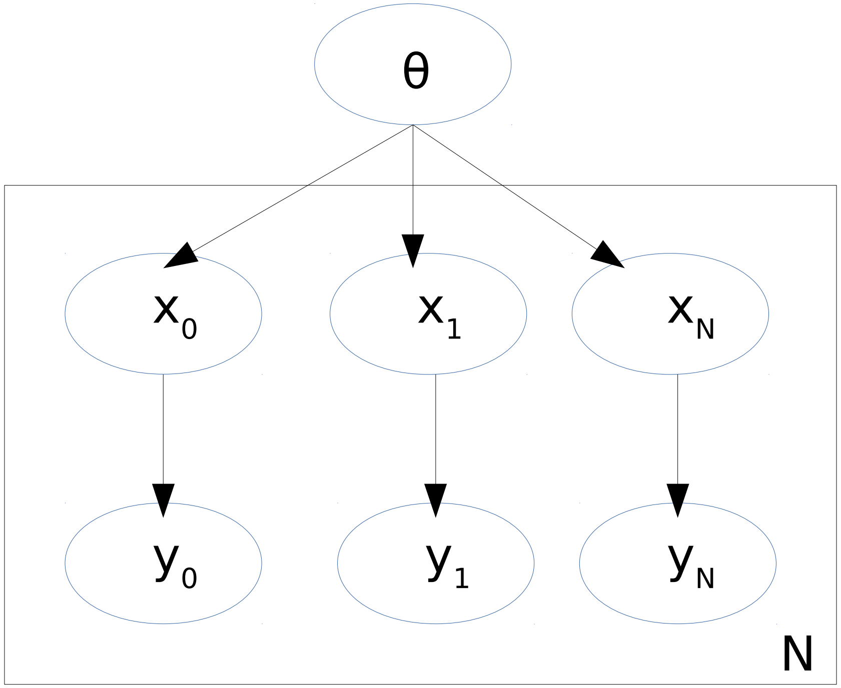
Hidden variables

- $\mathrm{p}(\theta, \{x_i\} \mid \{y_i\}, \{\sigma_{yi}\}) \sim \mathrm{p}(\theta) \prod_i \mathrm{p}(x_i \mid \theta)\ \mathrm{p}(y_i \mid x_i, \sigma_{yi})$
- A function $y(x)$ exists for mapping $x \rightarrow y$
- $\mathrm{p}(y_i \mid x_i, \sigma_{yi}) \sim \mathrm{Normal}(y_i \mid y(x_i), \sigma_{yi})$

Level-0: Population

Level-1: Individual Object-intrinsic

Level-1: Individual Object-observable

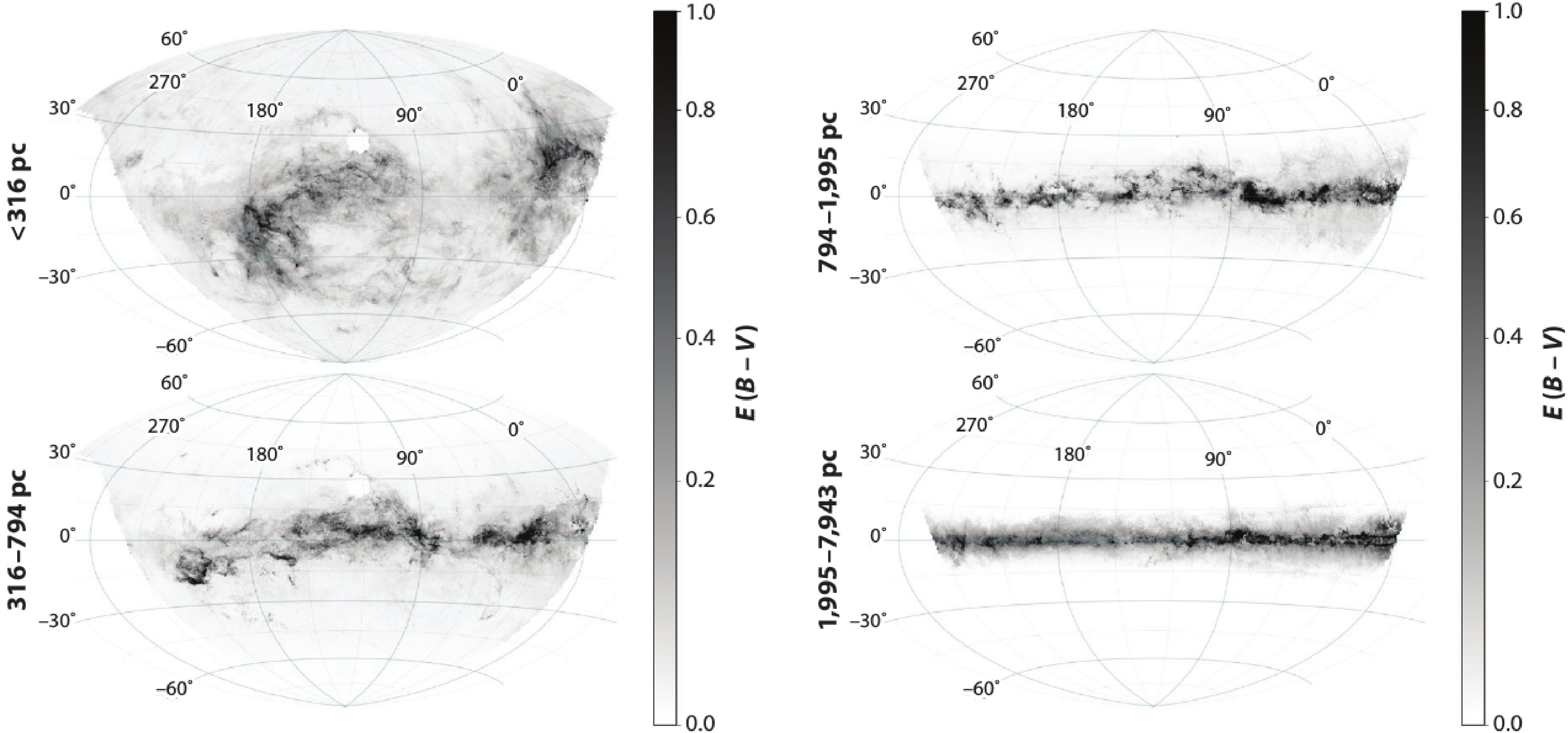


Intrinsic variables of a star.

- Intrinsic params: $x = ([M/H], \tau, m, s, l, b, E)$
- Obsevables: $y=(J, H, K, T_{eff}, \log g, [M/H], l, b)$
- Given x one can compute y using isochrones
- There exists a function $y(x)$ mapping x to y.

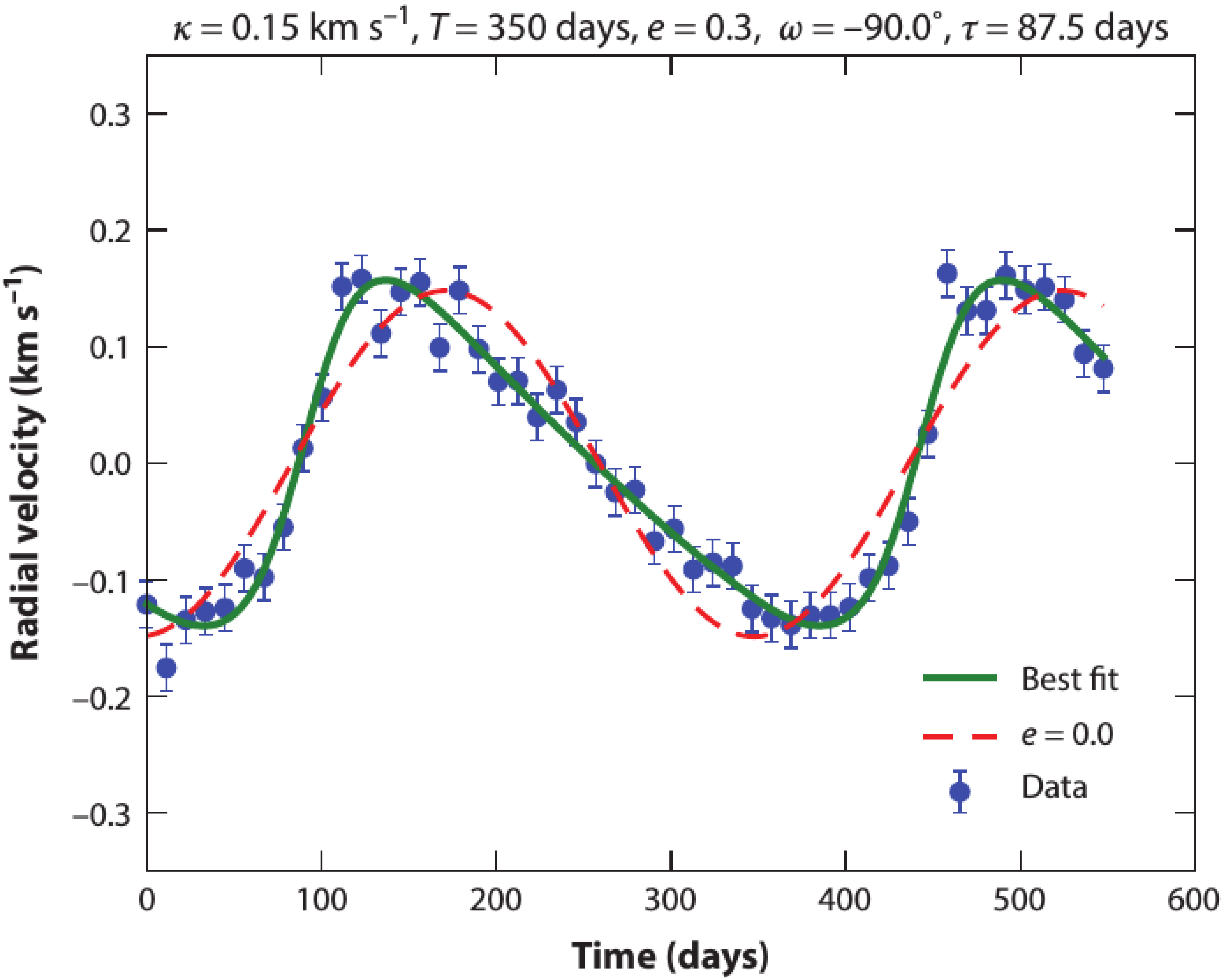
3d Extinction- $E_{B-V}(s)$

- Pan-STARRS 1 and 2MASS



Green et al. 2015

Exoplanets

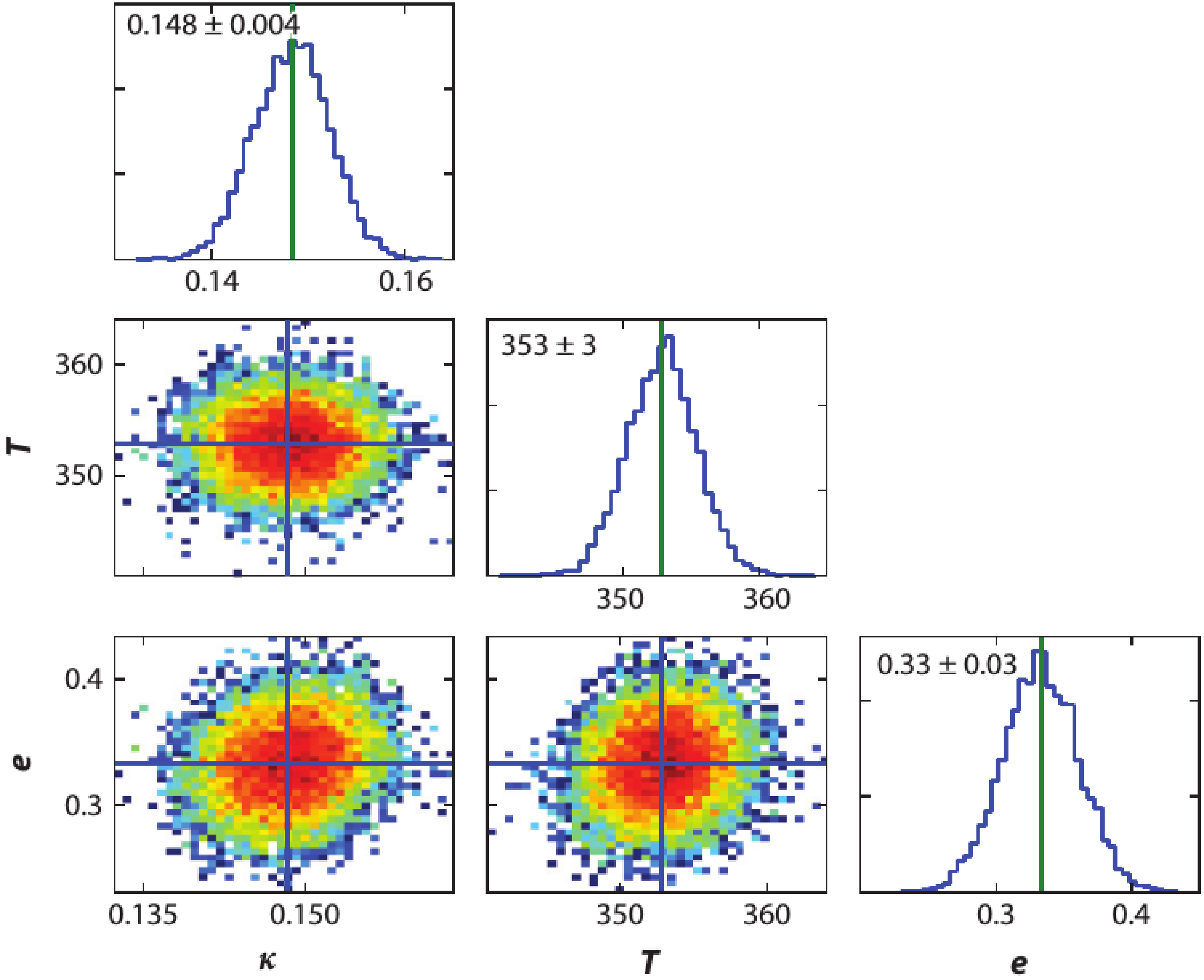


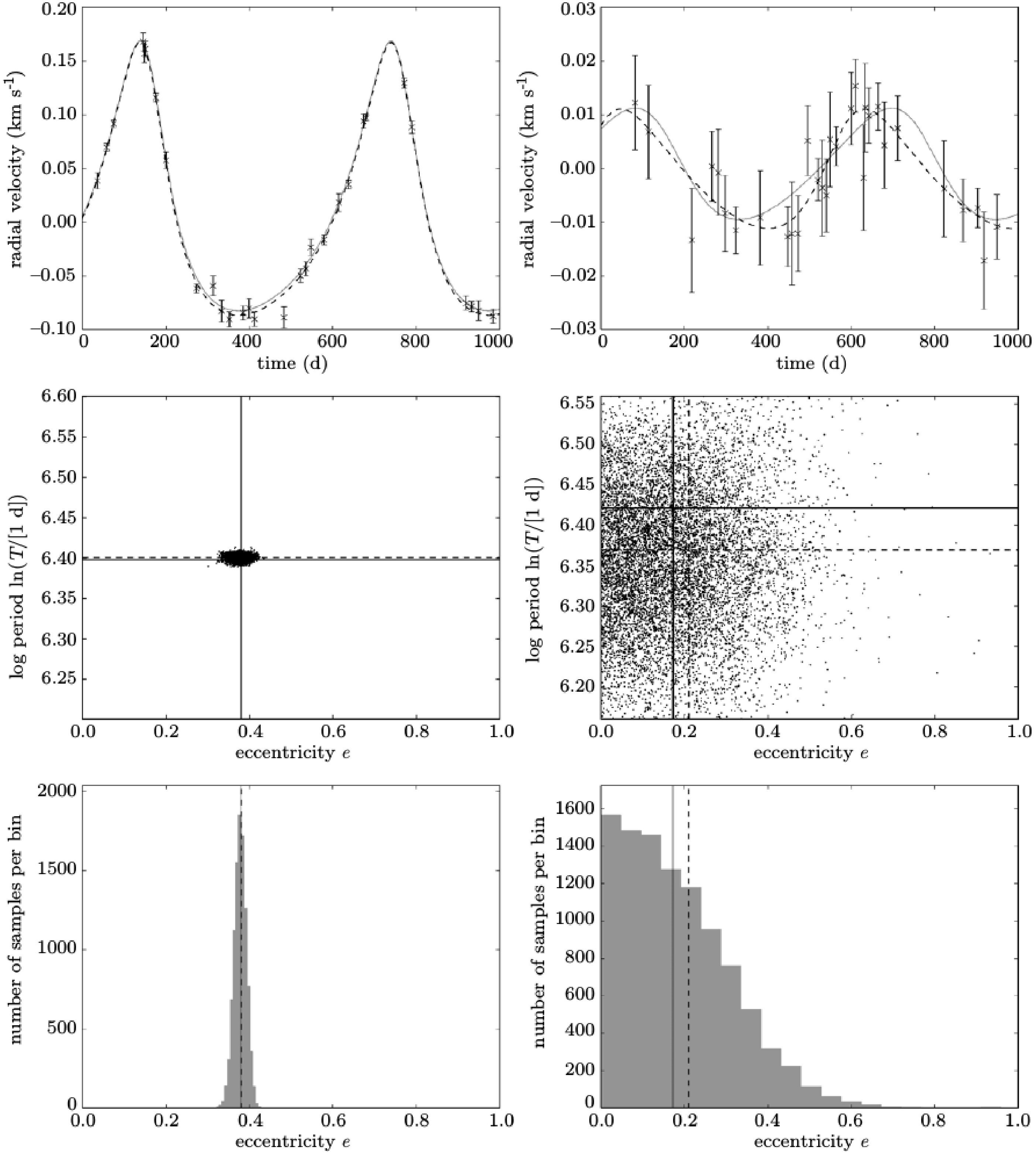
$\kappa = 0.15$ km s^{-1}, $T = 350$ days, $e = 0.3$, $\omega = -90.0°$, $\tau = 87.5$ days
Radial velocity (km s^{-1})
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
100
200
300
400
500
600
Time (days)
Best fit
$e = 0.0$
Data

- $x_i = (v_0, \kappa, T, e, \omega, \tau, S)$
- Mean velocity of center of mass v_0
- Semi-amplitude κ
- Time period T
- Eccentricity e
- Angle of pericenter from the ascending node ω
- Time of passage through the pericenter τ
- Intrinsic dispersion of a star S

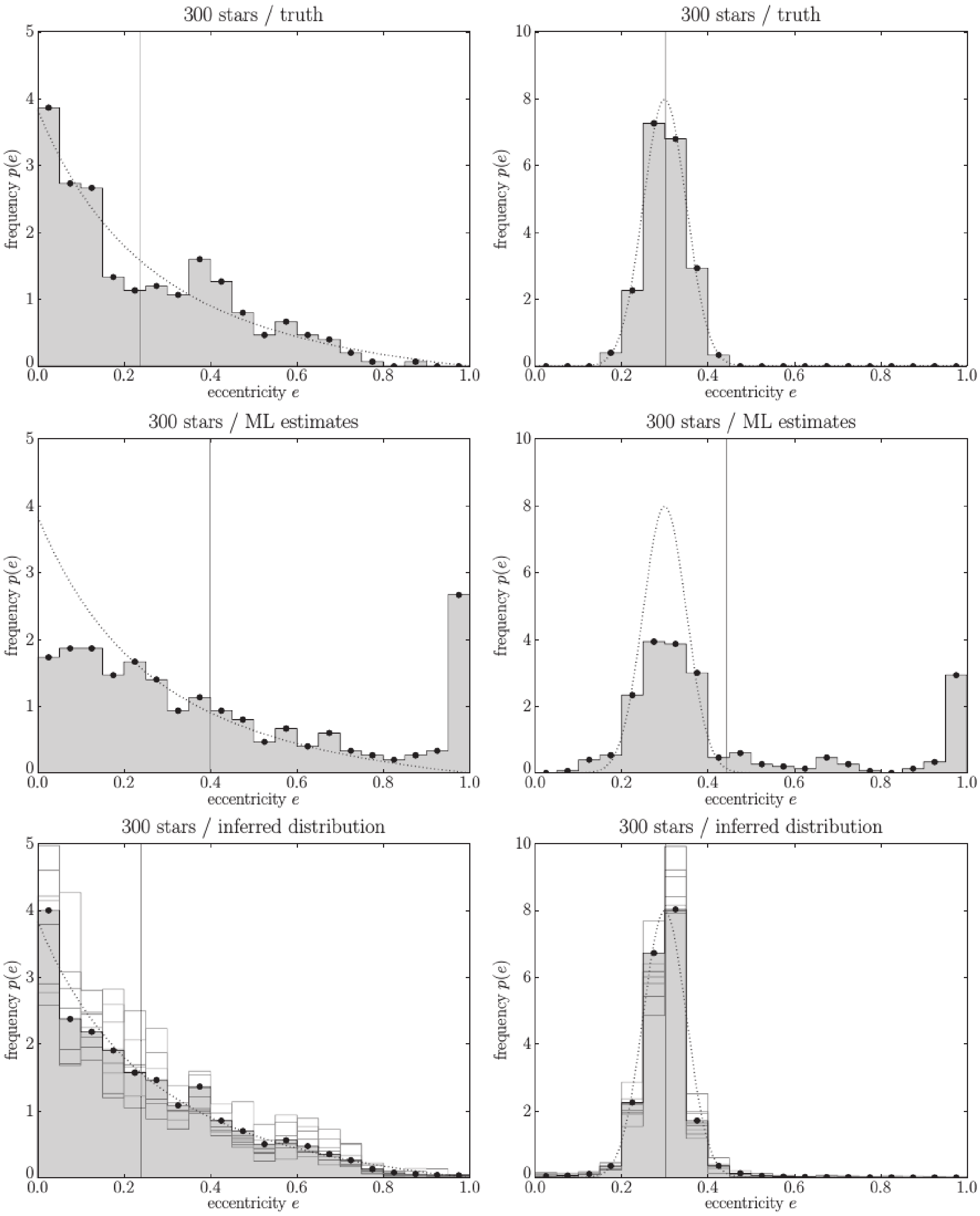
$$v(t) = \kappa\,[\cos(f+\omega) + e\cos\omega] + v_0, \quad \text{with } \kappa = \frac{(2\pi G)^{1/3} m \sin I}{T^{1/3}(M+m)^{2/3}\sqrt{1-e^2}}.$$

$$\tan(f/2) = \sqrt{\frac{1+e}{1-e}}\tan(u/2), \qquad u - e\sin u = \frac{2\pi}{T}(t-\tau).$$


0.148 ± 0.004
0.14
0.16
353 ± 3
360
350
T
350
360
0.33 ± 0.03
0.4
0.3
e
0.135
0.150
κ
350
360
T
0.3
0.4
e



- Hogg et al 2010



- Hogg et al 2010

How to solve BHM models

- Two step: Hogg et al. 2010
  - $p(\theta | \{y_i\}, \sigma_y) \sim p(\alpha) \prod_i \int dx_i \, p(y_i | x_i ,\sigma_{yi}) \, p(x_i | \theta)$

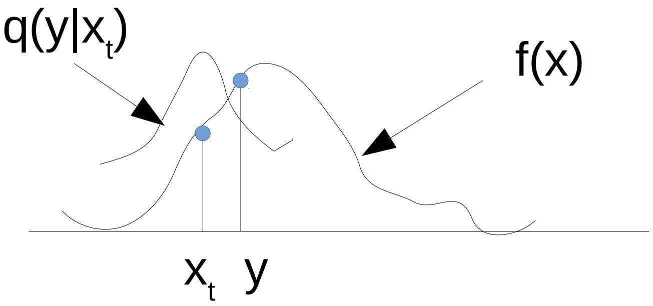
$$\sim p(\alpha) \prod_i \int dx_i \, \underset{\text{sample } x_{ik}}{\underline{p(y_i | x_i , \sigma_{yi}) \, p(x_i)}} \, [p(x_i | \theta)/p(x_i)]$$

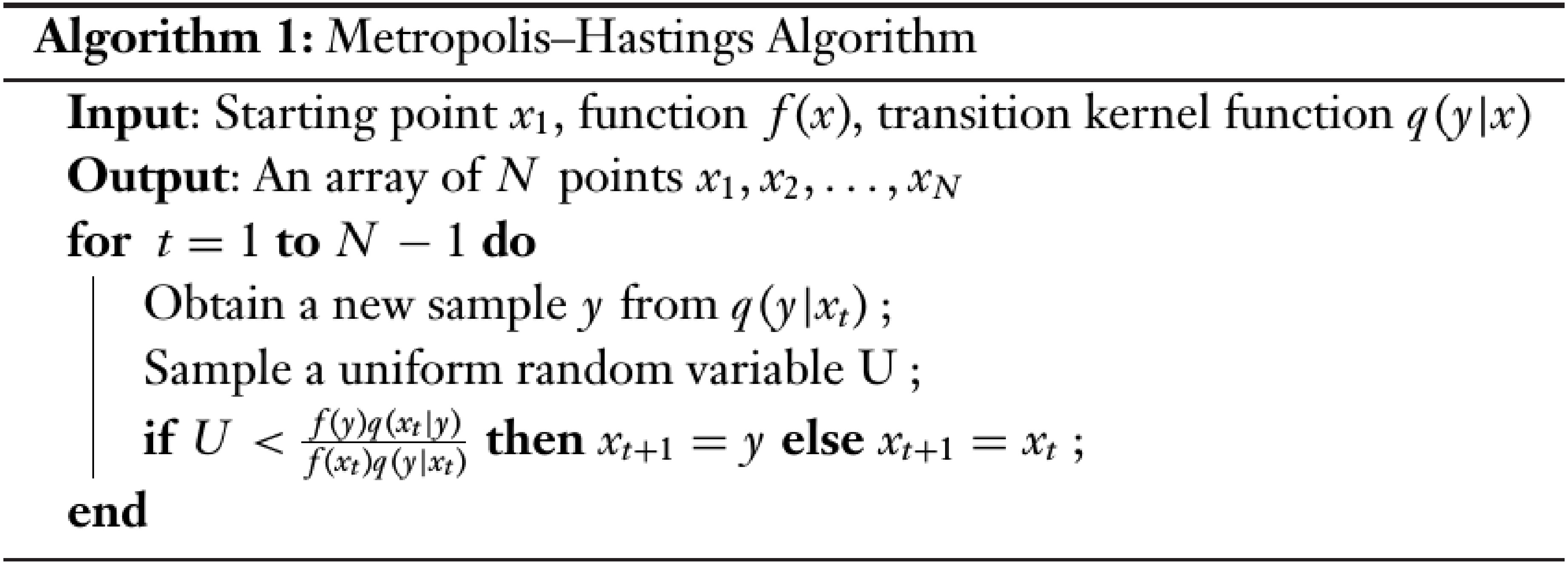
$$\sim p(\alpha) \prod_i (1/K) \sum_k [p(x_{ik} | \theta) / p(x_{ik})]$$

Importance Sampling

- MWG:
  - $p(\theta, \{x_i\} | \{y_i\} , \{\sigma_{yi}\}) \sim p(\theta) \prod_i p(y_i | x_i , \sigma_{yi}) \, p(x_i | \theta)$

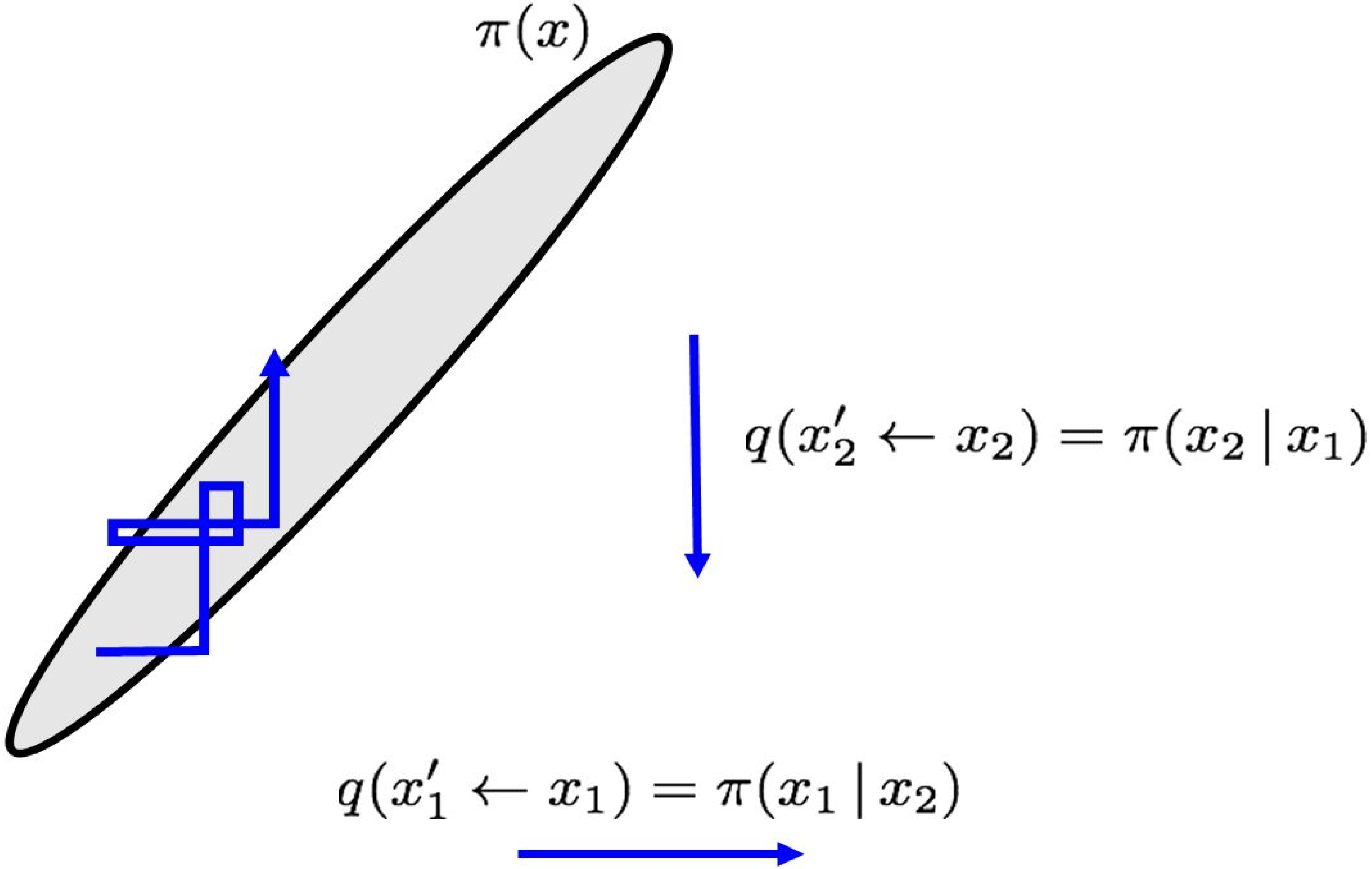
MH algorithm





```
Algorithm 1: Metropolis–Hastings Algorithm
Input: Starting point $x_1$, function $f(x)$, transition kernel function $q(y|x)$
Output: An array of $N$ points $x_1, x_2, \ldots, x_N$
for $t = 1$ to $N - 1$ do
    Obtain a new sample $y$ from $q(y|x_t)$ ;
    Sample a uniform random variable U ;
    if $U < \frac{f(y)q(x_t|y)}{f(x_t)q(y|x_t)}$ then $x_{t+1} = y$ else $x_{t+1} = x_t$ ;
end
```

Gibbs Sampling



$\pi(x)$

$$q(x_2' \leftarrow x_2) = \pi(x_2 \,|\, x_1)$$

$$q(x_1' \leftarrow x_1) = \pi(x_1 \,|\, x_2)$$

Image: Ryan Adams

Metropolis Within Gibbs

- Gibbs sampler requires sampling from conditional distribution.

- Replace this with a MH step.

- Rather than updating all at one time, one can do it one dimension at a time.

- A complicated distribution can be broken up into sequence of smaller or easier to samplings is the main strength of this.

BMCMC- a python package

- pip install bmcmc
- https://github.com/sanjibs/bmcmc
- Ability to solve hierarchical Bayesian models.
- Documentation:
  - http://bmcmc.readthedocs.io/en/latest/

```
class gauss2(bmcmc.Model):
    def set_descr(self):
        # setup descriptor
        self.descr['mu']     =['l0',0.0,1.0,r'$\mu$'    ,-500,500.0]
        self.descr['sigma']  =['l0',1.0,1.0,r'$\sigma$',1e-10,1e3]
        self.descr['xt']     =['l1',0.0,1.0,r'$x_t$'    ,-500.0,500.0]

    def set_args(self):
        # setup data points
        np.random.seed(11)
        # generate true coordinates of data points
        self.args['x']=np.random.normal(loc=self.eargs['mu'],scale=self.eargs['sigma'],size=self.eargs['dsize'])
        # add observational uncertainty to each data point
        self.args['sigma_x']=np.zeros(self.args['x'].size,dtype=np.float64)+0.5
        self.args['x']=np.random.normal(loc=self.args['x'],scale=self.args['sigma_x'],size=self.eargs['dsize'])

    def lnfunc(self,args):
        # log posterior
        temp1=scipy.stats.norm.logpdf(args['xt'],loc=args['mu'],scale=args['sigma'])
        temp2=scipy.stats.norm.logpdf(args['x'],loc=args['xt'],scale=args['sigma_x'])
        return temp1+temp2
```

Create an object and run the sampler.

```
>>> mymodel=gauss2(eargs={'dsize':100,'mu':1.0,'sigma':2.0})
>>> mymodel.sample(['mu','sigma','xt'],50000,ptime=1000)
```

Why do we need model selection?

- Models are designed to explain and understand the data.
- In general, we do not know the true model, we build models to fit the observed data and keep improving them by adding new features.
- More than one competing model or theory
- Parameter fitting, the number of parameters not known.

Why do we need model selection criterion?

- As we increase the number of parameters the model will fit the data better and better.

- 10 data points from function y=sin(2πx)+ε (green)
- fitted by polynomials (red) of degree M.
- What will happen if we add a new point?

Oscillating function like Asin(nx) can be made to pass through all data points. It has only two Parameters.

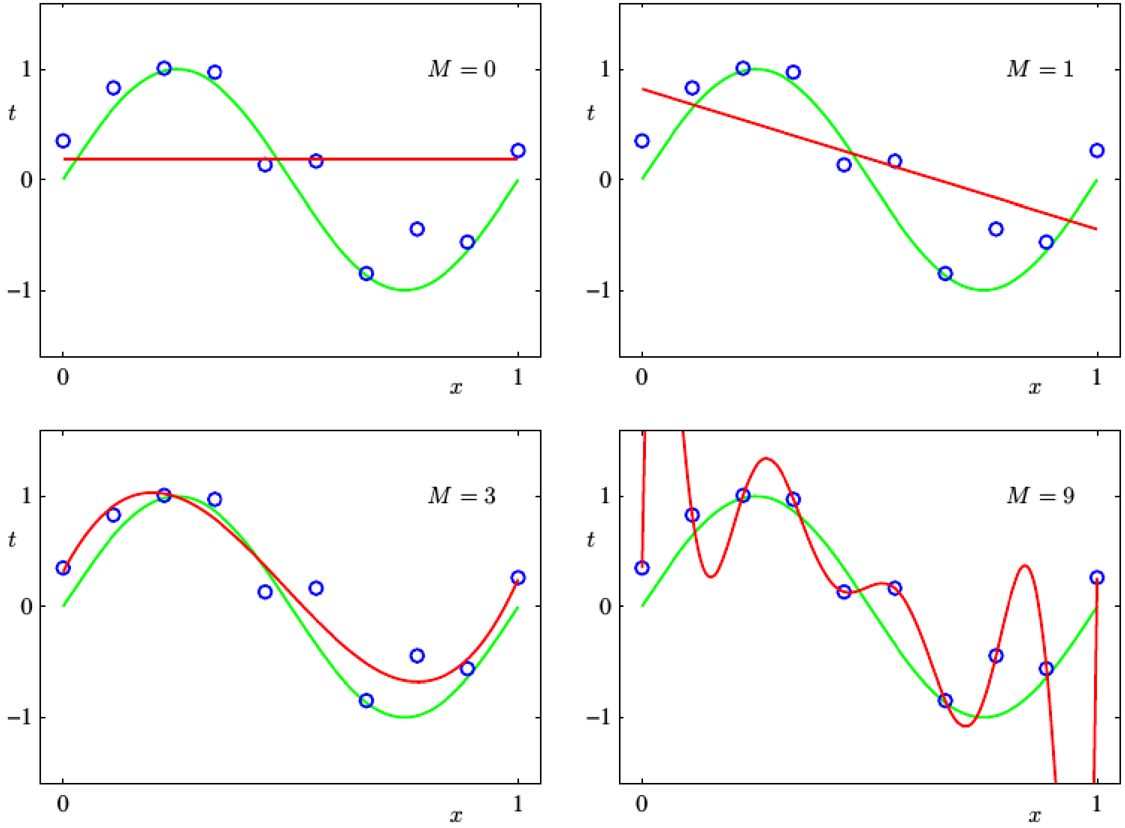
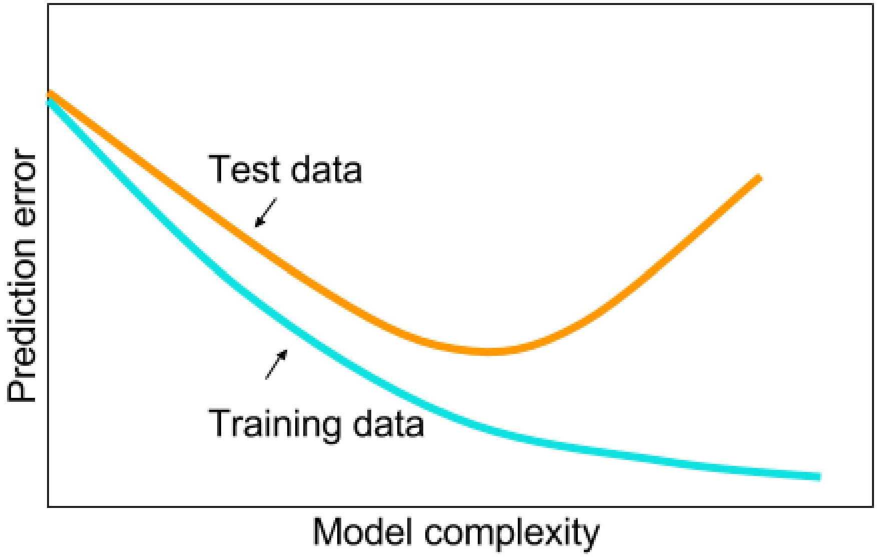
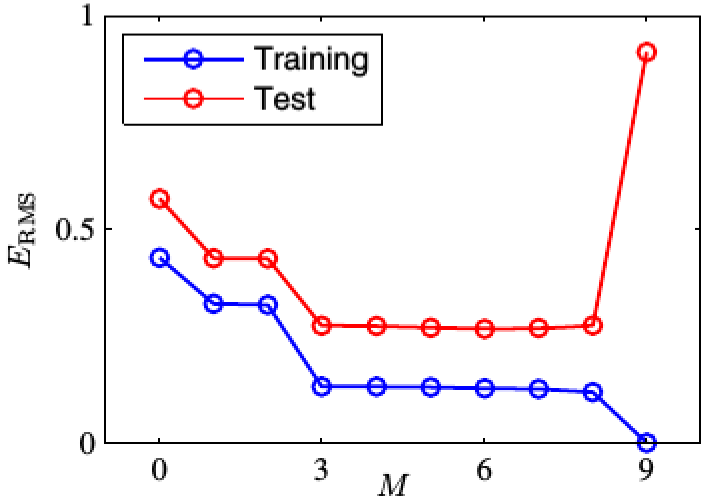


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

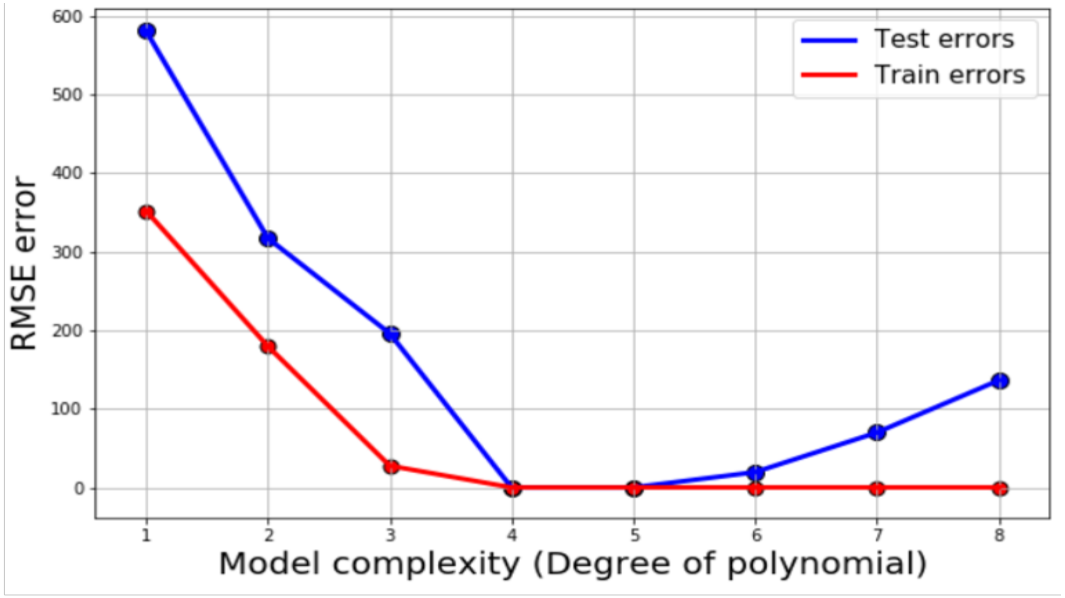
Bishop book

- Polynomial model parameters $\theta=\theta(Y_{train})$
- $E(\theta,Y_{test})=(1/N)\sum_i\{y(x_i;\theta) - y_i\}^2$
- For M<3, E is very high, as model too simple/inflexible/rigid.
- For 3<M<8, not much change, power series expansion of sin(x) contains terms of all orders.
- For M=9, E_{train} =0, 10 dof for 10 data points, however E_{test} very high.





Bishop book



Why do we need model selection criterion?

- As we increase the number of parameters the model will fit the data better and better.

- Given a new data it will perform badly.
  - overfitting
  - We do not want to overfit the models.
    - Cross validation

- Bayesian model comparison has the built in Occams factor that penalizes more complex models. However, it is not easy to do.

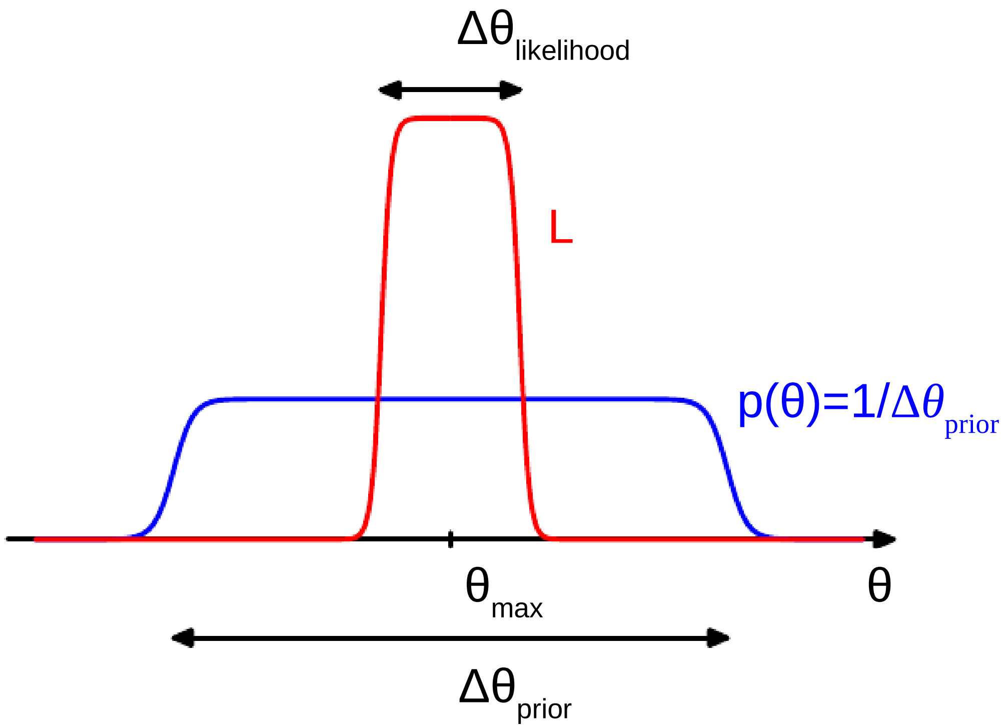
Cross validation

- How well will the model work on future data set.
- Observed Data set → Training set + Test/Validation set
- One test set:
  - (70,30), (50,50),Unreliable, wastes too much data
- Exhaustive:
  - Leave-p-out cv (LPOCV): C(n,p), C(100,30)~3 10^{25}
  - Leave-one-out cv (LOOCV): n
  - Costly
- K-fold cross validation:
  - Split into K subsamples, use as validation one of them, repeat k times.
  - Cheaper.
  - Use k=10, not too expensive does not waste too much data.

Bayesian model comparison (Bayes Factor)

Bayes factor of model M_2 wrt M_1 . Model M_2 has θ as a free parameter, while model M_1 has a fixed value of θ_0 for it.

$$\begin{aligned} B_{21} &= p(D|M_2) / p(D|M_1) \\ &= \int p(\mathrm{D}|\theta)\, p(\theta)\mathrm{d}\theta \,/\, p(\mathrm{D}|\theta_0) \\ &= [L(\theta_{\mathrm{max}})\, \Delta\theta_{\mathrm{likelihood}} / \Delta\theta_{\mathrm{prior}}] / L(\theta_0) \\ &= [L(\theta_{\mathrm{max}}) / L(\theta_0)]\, [\Delta\theta_{\mathrm{likelihood}} / \Delta\theta_{\mathrm{prior}}] \end{aligned}$$

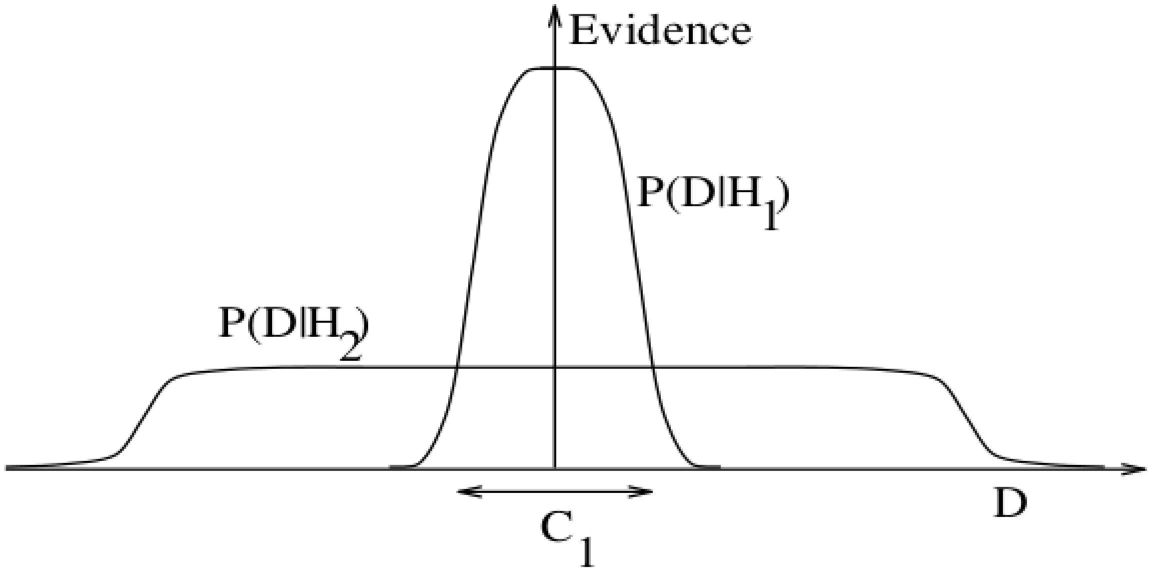


$\Delta\theta_{\mathrm{likelihood}}$
L
$p(\theta)=1/\Delta\theta_{\mathrm{prior}}$
θ_{max}
θ
$\Delta\theta_{\mathrm{prior}}$

Bishop book

Bayesian model comparison (Bayes Factor)

- $B_{21} = p(D'|H_2)/p(D'|H_1)$
- A simple model H_1 only makes a limited range of predictions.
- A complex model H_2 (more free prams) is able to predict a large range of data sets.
- Note, $\int p(D|H_1)\ \mathrm{d}\ \mathrm{D} = 1$
- Hence, for observed data D', $p(D'|H_2) < p(D'|H_1)$



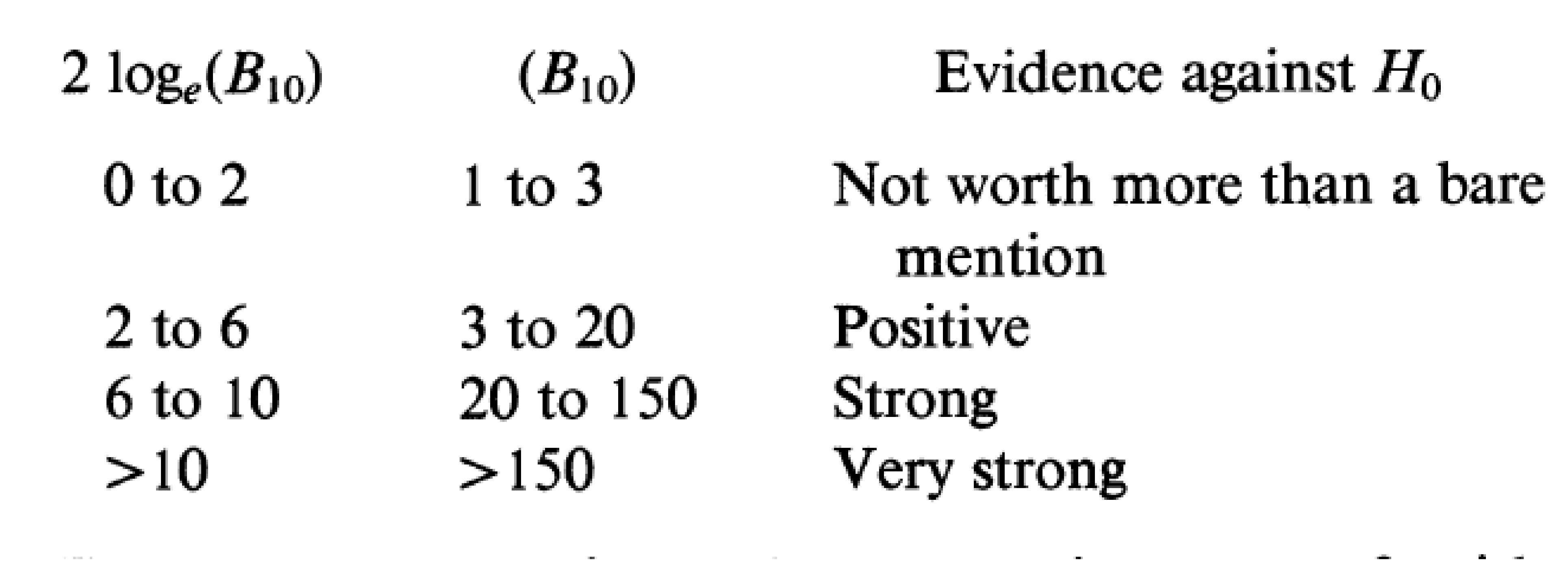
MacKay book

Horizontal axes: space of all possible data sets

Bayes factor: caveats

- Bayesian Model selection comparison is complicated.
  - $B_{21}=p(D|M_2)/p(D|M_1)$
  - $=[L(\theta_{\mathrm{max}}) \,/\, L(\theta_0)]\ [\Delta\theta_{\mathrm{likelihood}}/\Delta\theta_{\mathrm{prior}}]$
  - For parameter estimation range of priors is not an issue but for model selection it is.
    - In most cases we do not have a reasonable sense of range of priors.
      - What is the prior for the coefficients of a polynomial?
  - Computing Bayes factor is computationally challenging.
    - $p(D|M)=\int p(D|\theta,M)p(\theta|M)\, d\theta$
    - Likelihood is peaked and confined to a narrow region but has long tails whose contribution cannot be neglected.

Bayes factor interpretation
Kass & Raftery 1995

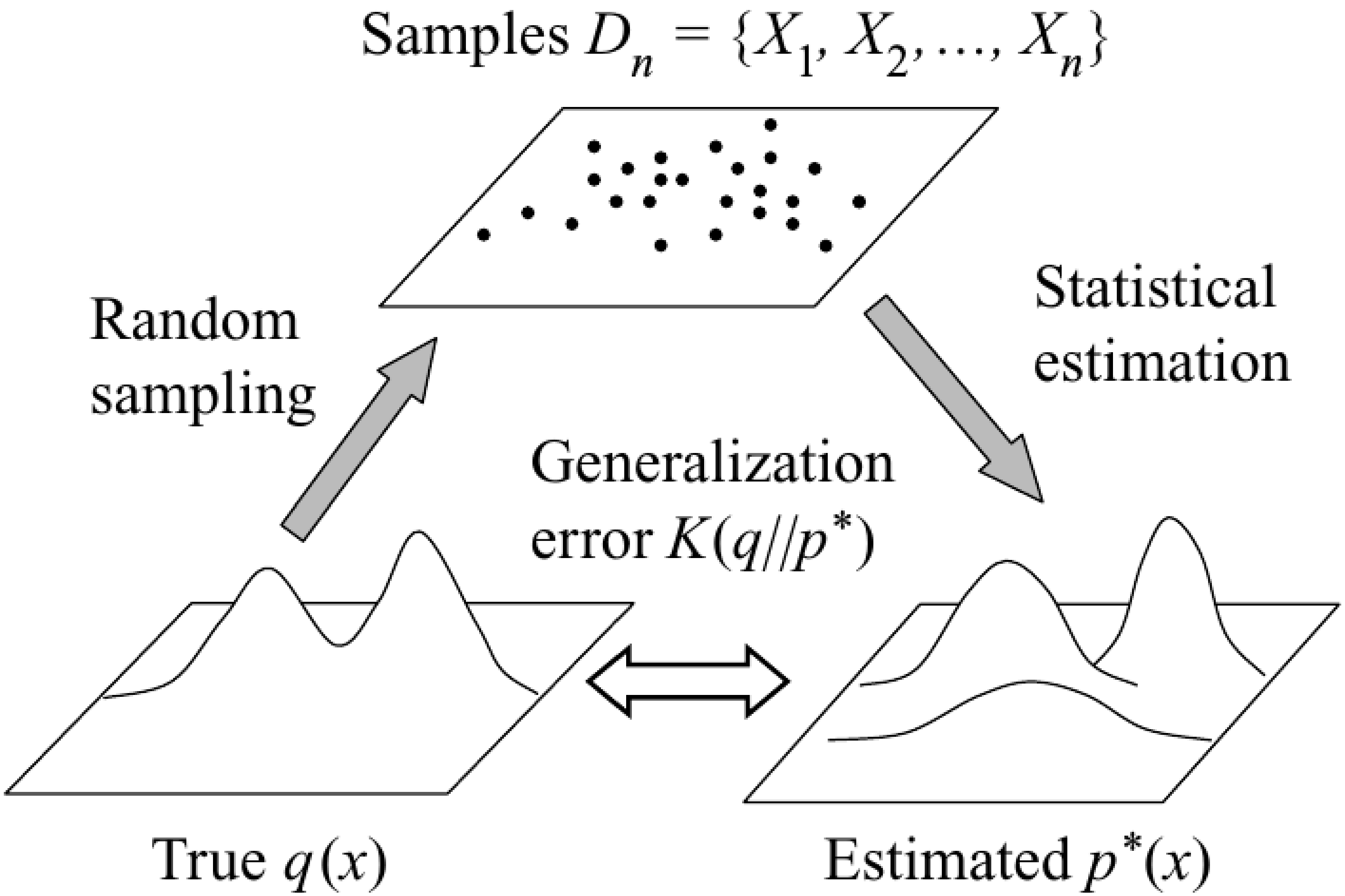


$2 \log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

Information criteria

- $Y=\{y_1, y_2, \dots, y_n\}$
  - $\ln p(Y|\bar{\theta}) = \ln [\prod_i p(y_i|\bar{\theta})] = \sum_i \ln p(y_i|\bar{\theta})$
- AIC: Akaike 1974
  - **$-\ln p(Y|\bar{\theta})+d$**
  - Oct 2014, 14000 cites, 73rd most cited
  - Frequentist. Based on information theory.
- BIC: Schwarz 1978
  - **$-\ln p(Y|\bar{\theta})+d\,(\ln n)/2$**
  - Based on Bayesian model comparison
  - An approximation of Bayesian evidence p(D|M).
  - Roughly equivalent to model selection based on Bayes Factor.
  - Does not require prior, making it useful when priors are difficult to compute.
- Of the form: -(Goodness of fit)+(penalty for model complexity)

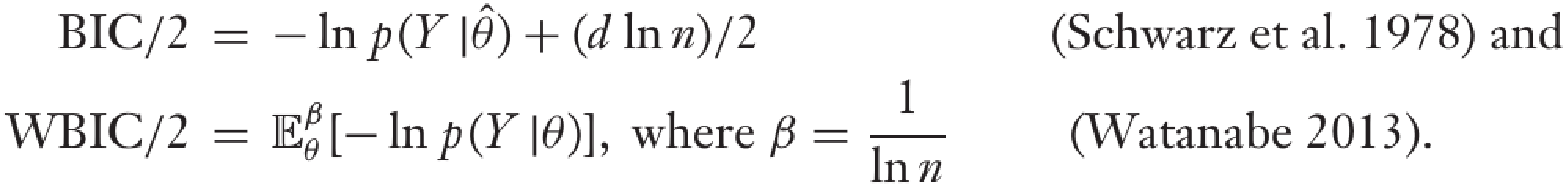
Statistical learning


Samples $D_n = \{X_1, X_2, \ldots, X_n\}$
Random sampling
Statistical estimation
Generalization error $K(q//p^*)$
True $q(x)$
Estimated $p^*(x)$

Statistical learning

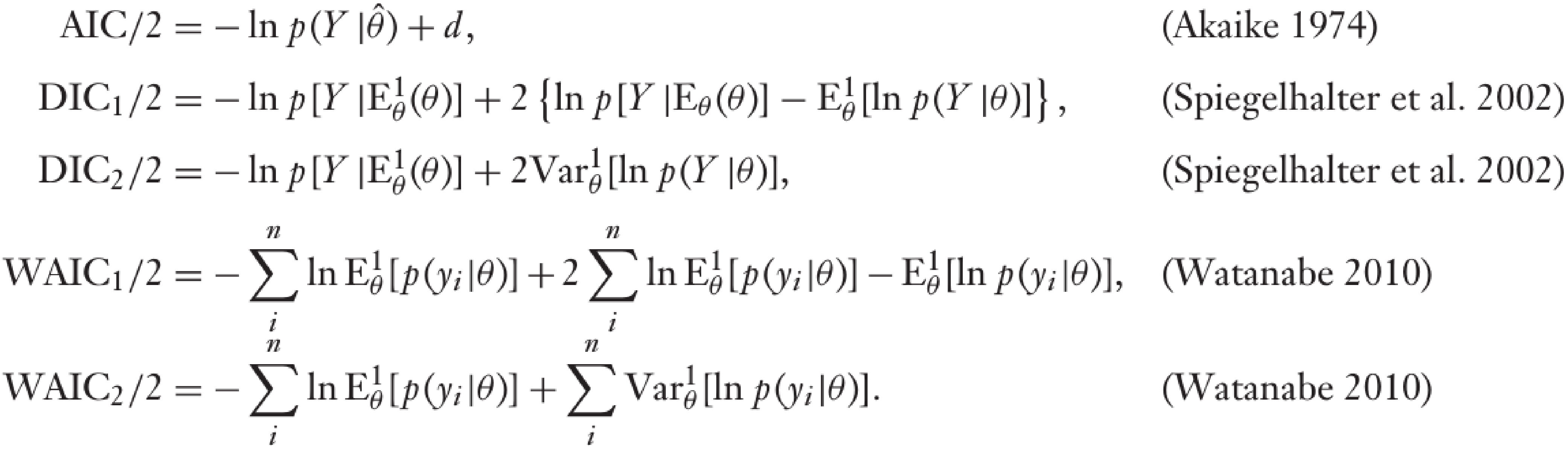
Watanabe 2009 Book (Algebraic Geometry and Statistical Leraning Theory)

Other information criteria.



$$\mathrm{BIC}/2 \;=\; -\ln p(Y\,|\hat{\theta}) + (d\ln n)/2 \qquad \text{(Schwarz et al. 1978) and}$$

$$\mathrm{WBIC}/2 \;=\; \mathbb{E}_{\theta}^{\beta}[-\ln p(Y\,|\theta)],\ \text{where } \beta = \frac{1}{\ln n} \qquad \text{(Watanabe 2013).}$$



$$\mathrm{AIC}/2 = -\ln p(Y\,|\hat{\theta}) + d, \qquad \text{(Akaike 1974)}$$

$$\mathrm{DIC}_1/2 = -\ln p[Y\,|\mathrm{E}_{\theta}^{1}(\theta)] + 2\left\{\ln p[Y\,|\mathrm{E}_{\theta}(\theta)] - \mathrm{E}_{\theta}^{1}[\ln p(Y\,|\theta)]\right\}, \qquad \text{(Spiegelhalter et al. 2002)}$$

$$\mathrm{DIC}_2/2 = -\ln p[Y\,|\mathrm{E}_{\theta}^{1}(\theta)] + 2\mathrm{Var}_{\theta}^{1}[\ln p(Y\,|\theta)], \qquad \text{(Spiegelhalter et al. 2002)}$$

$$\mathrm{WAIC}_1/2 = -\sum_{i}^{n}\ln \mathrm{E}_{\theta}^{1}[p(y_i|\theta)] + 2\sum_{i}^{n}\ln \mathrm{E}_{\theta}^{1}[p(y_i|\theta)] - \mathrm{E}_{\theta}^{1}[\ln p(y_i|\theta)], \qquad \text{(Watanabe 2010)}$$

$$\mathrm{WAIC}_2/2 = -\sum_{i}^{n}\ln \mathrm{E}_{\theta}^{1}[p(y_i|\theta)] + \sum_{i}^{n}\mathrm{Var}_{\theta}^{1}[\ln p(y_i|\theta)]. \qquad \text{(Watanabe 2010)}$$

WAIC and WBIC

- Works for Singular statistical models.
- Makes use of predictive density.
- In the asymptotic limit of large sample size both AIC and WAIC
  - Equivalent to LOOCV
  - Equivalent to expected KL divergence of predicted distribution from the true distribution.

BIC vs AIC

- First term giving likelihood of data (goodness of fit) is same.
- But penalty term for model complexity is more severe in BIC
  - For n>=8, d (ln n) /2> d
- BIC favors smaller models than AIC.
- Differences will be more pronounced for large n.

BIC vs AIC/WAIC

- Asymptotically consistent:
  - If the candidate list of models contains the true model, the method will asymptotically select the true model with probability one.
- Asymptotically efficient:
  - The method will asymptotically select the model that minimizes the mean squared error of prediction.
- AIC is asymptotically efficient yet not consistent
- BIC is asymptotically consistent yet not efficient.

Burnham & Anderson 2002, Lecture by Cavanaugh 2012, Spiegelhalter 2014

BIC vs AIC: practical perspective

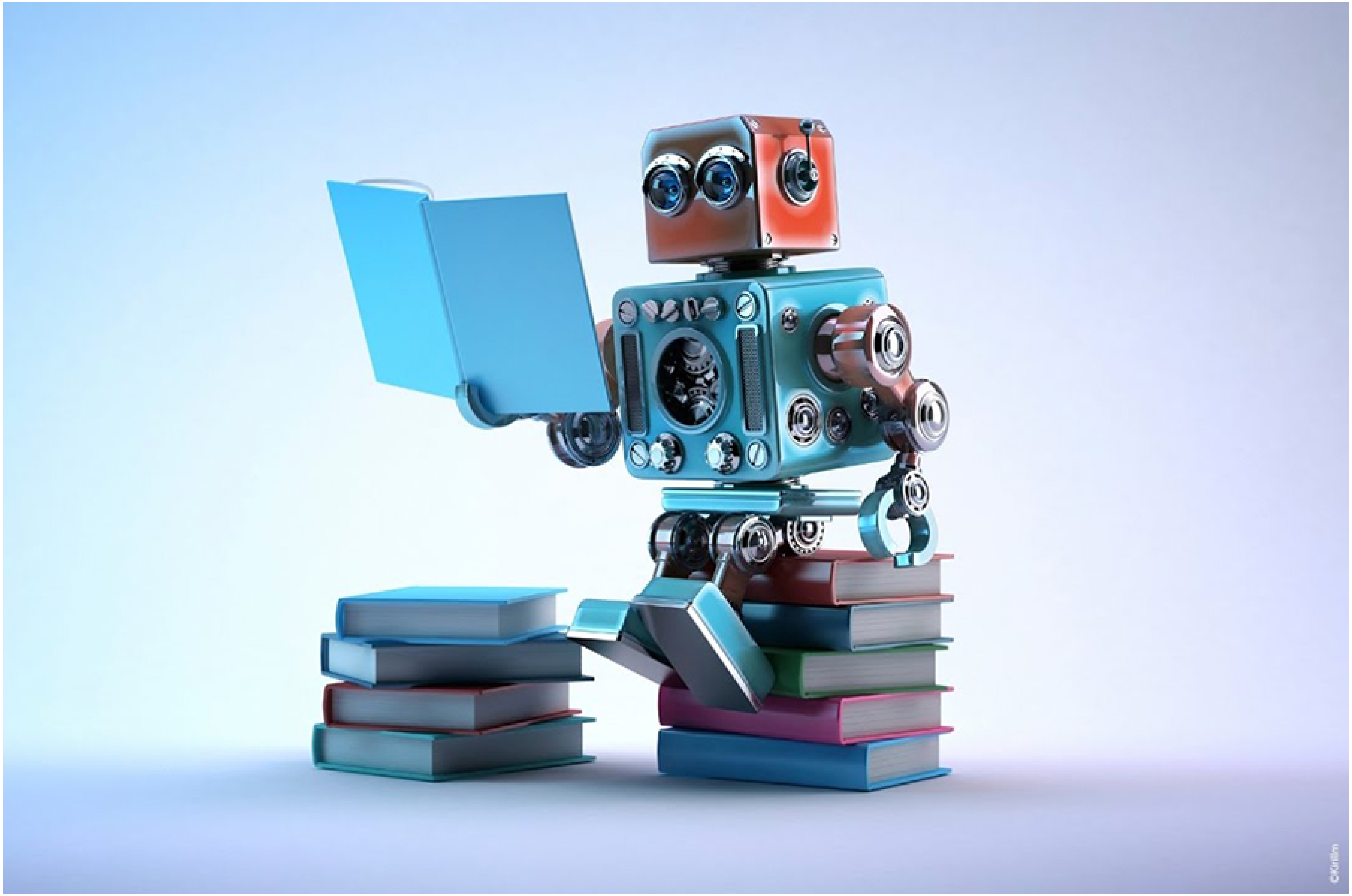
- AIC: primary goal of modelling is predictive
  - Build model to fit future data effectively
- BIC: primary goal of modelling is descriptive
  - Build a model with most meaningful factors influencing outcome based on an aseessment of relative importance.
- As the sample size grows, predictive accuracy improves as subtel effects are admitted to the model. AIC will increasingly favor the inclusion of such effects; BIC will not.

Source: Lecture Cavanaugh 2012

Which to choose.

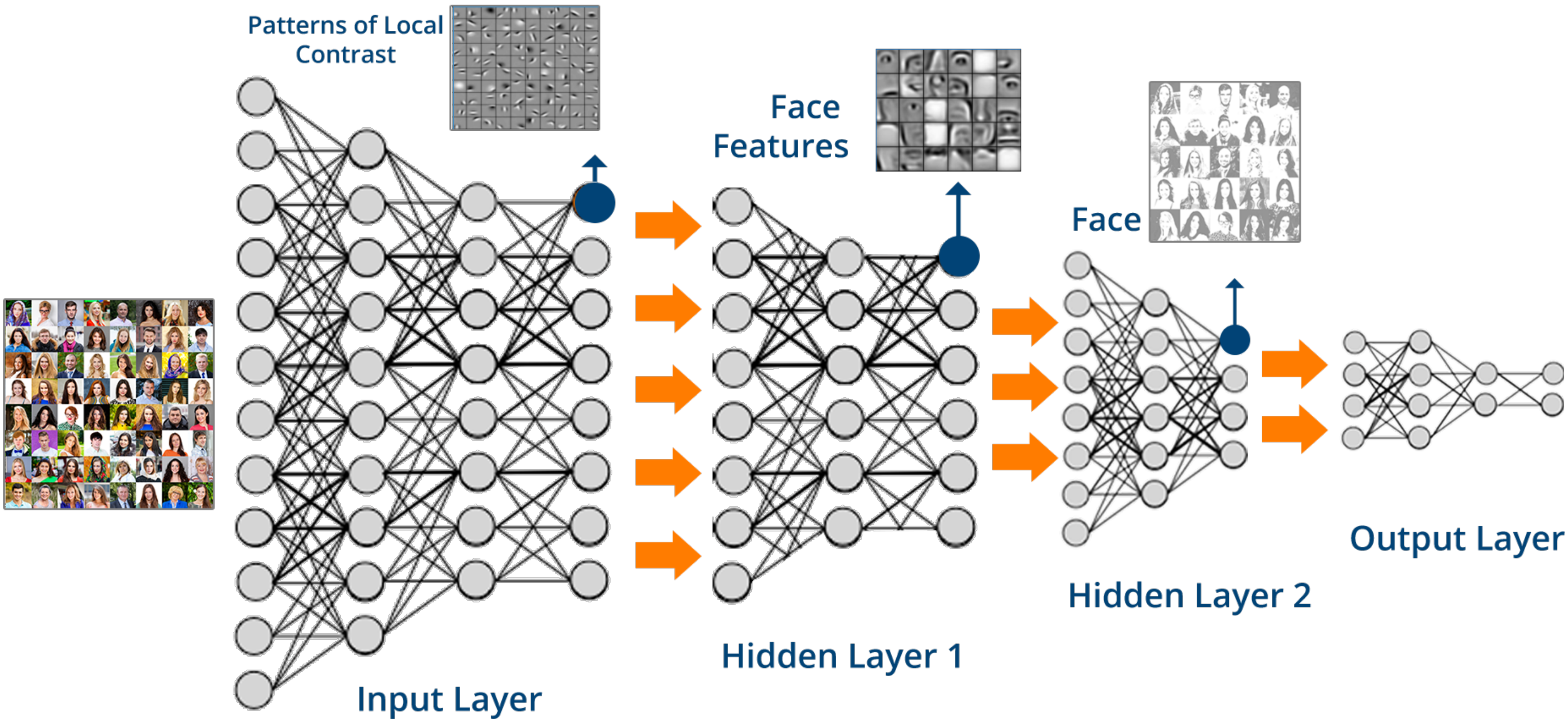
- Both Bayesian and predictive have their strength and weaknesses.
- If the model is physical and the choice of priors is well justified, then Bayes factor are the best suited.
  - BIC can also be used, if priors an issue.
- If model is explanatory and empirical, which means predictive accuracy for future data is desired, choose WAIC.

Future



Machine Learning

Image:www.iamwire.com



Deep Learning

Image: https://www.edureka.co

Bayesian statistics a glue connecting different fields.

- *My or your model fitting problem is also everyone elses problem.*

- Growth in data science, inference.
  - Predictive analysis of great use for industry.
  - Confluence of industry and science. (facebook, google).
  - **autodiff, pytorch, tensorflow**

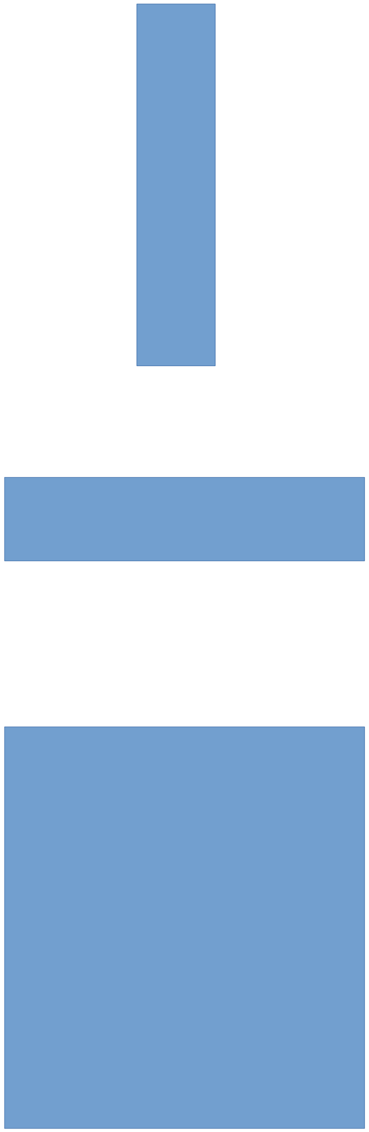
- Development of good optimizers.

- Platforms for probabilistic inference.
  - **Stan, Edward, PyMC3**

Future

- Big Data
  - Tall (N), Wide (d),
  - Model: Complexity (d), Hierarchies

- MCMC too slow
  - MLE, optimization
  - Speed up traditional MCMC for tall data.
  - Hamiltonian Monte Carlo
  - Variational Bayes

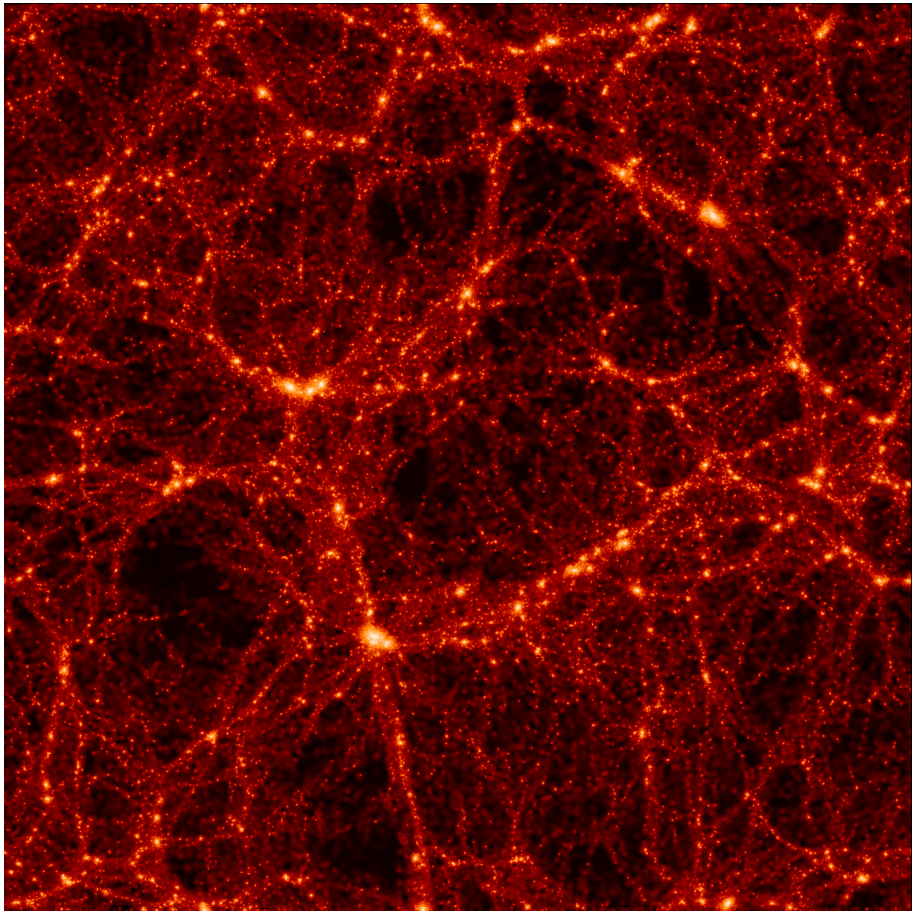


Bayesian nonparametrics (BNP)

- Useful for big data.
- Properties of big data
  - Feature space is large → complex models
  - Difficult to find suitable model.

Big data analogy

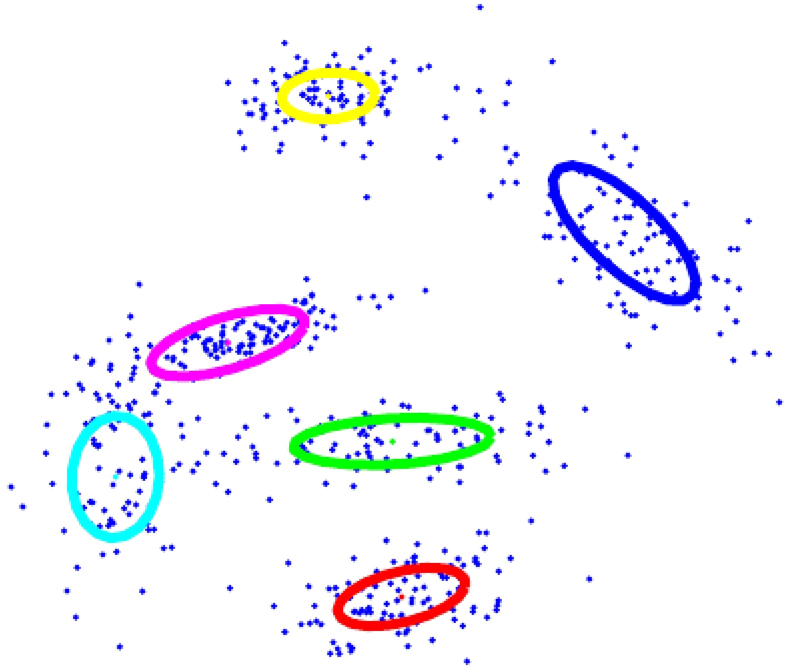
- More the data, more substructures and more hierachy of substructures.



- A flexible model whose complexity can grow with data size.
  - Polynomials with degree being free
  - Gaussian mixture model with number of clusters free

BNP

- $p(x \mid \theta) = \sum \alpha_i \, \text{Normal}(x \mid \mu_i, \sigma_i^2)$, $i=\{1,\ldots,K\}$
- Put a prior on $p(K)$
- Can do this without Bayesian model comparison.
- **Dirichlet Process** mixture models (Neal 2000).
  - A prior on $p(\alpha)_{,} K \rightarrow \infty$



Pseudo marginal MCMC for big data

- Speeding up MCMC for big data.
  - Subsample the data and compute the likelihood
  - $f'(x,y), y$ set of rows to use
  - $f'(x,y) = \exp[\sum_i \log f(x_i)]$, for each i in y
  - Likelihood becomes stochastic.

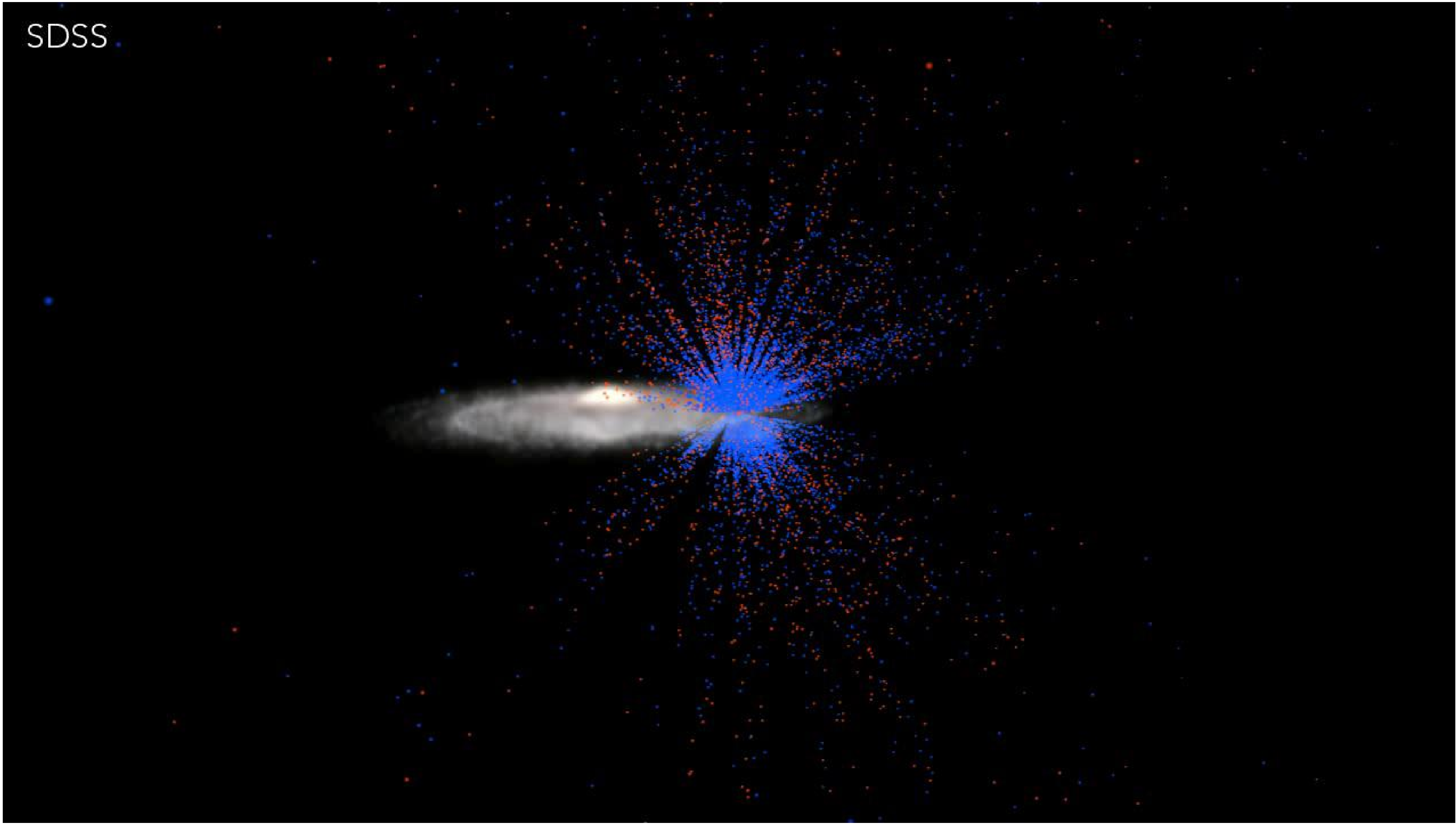
- Other cases of stochastic likelihood.
  - Marginalization problems
    - $p(\theta | x) = \int p(x | \theta, \alpha) d\alpha = \int \int p(x, z | \theta, \alpha, z) d\alpha\, dz$
  - Doubly intractable integrals

Doubly intractable integrals

- Singly intractable integral.
  - $\mathrm{p}(\theta|y) = \mathrm{p}(y\,|\,\theta)\,\mathrm{p}(\theta)/\mathrm{p}(y)$
  - The normalization constant $\mathrm{p}(y)$ (Evidence) is not known.
  - But we do not need to know it, to compute expectations.
  - We only need to sample from it.
  - $\mathrm{E}[\,f\,] = \int f(\theta)\,\mathrm{p}(\theta|y)\,\mathrm{d}\theta = 1/N \sum f(\theta_i)$

- What if $\mathrm{p}(y\,|\,\theta) = f(y;\theta)/Z(\theta)$?
  - Now expectation is doubly intractable integral.

- $p(x \mid \theta, S) = \rho(x \mid \theta)\, S(x) \,/\, \int \rho(x \mid \theta)\, S(x)\, dx$
- Fitting stellar halo density for stars in two cones (SDSS).



Handling stochastic likelihoods

- Monte Carlo Metropolis-Hastings
- If $U < f(x') / f(x)$**:**

    xl.append(**x’**)

  Else:

    xl.append(**x**)
- What if the function f is stochastic?

Pseudo Marginal MCMC

- Andrieu and Roberts (2009), Beaumont 2003

  Sample auxillary variable $\mathbf{y}_n$

```
If $U < f'(x_n, y_n)/f'(x, y)$:
    xl.append(x_n)
    yl.append(y_n)
Else:
    xl.append(x)
    yl.append(y)
```

- Does sample $f(x)$ provided $f'(x, y)$ is unbiased.
  - $E_y[f'(x, y)]=f(x)$
- If $\mathrm{Var}[\log f'(x_n, y_n)-\log f'(x, y)] > 1$, will get stuck.

Approximate MCMC

Murray 2006, Liang 2011, Sharma 2014, Sharma 2017

- Sample $\mathbf{y_n}$
- If $U < f'(x_n, y_n)/f'(x, y_n)$:

```
    xl.append(x_n)
    yl.append(y_n)
Else:
    xl.append(x)
    yl.append(y_n)
```

-
- Does not sample $f(x)$, rather $f_{\text{approx}}(x)$
- More stable, does not get stuck.

Pseudo marginal MCMC for big data

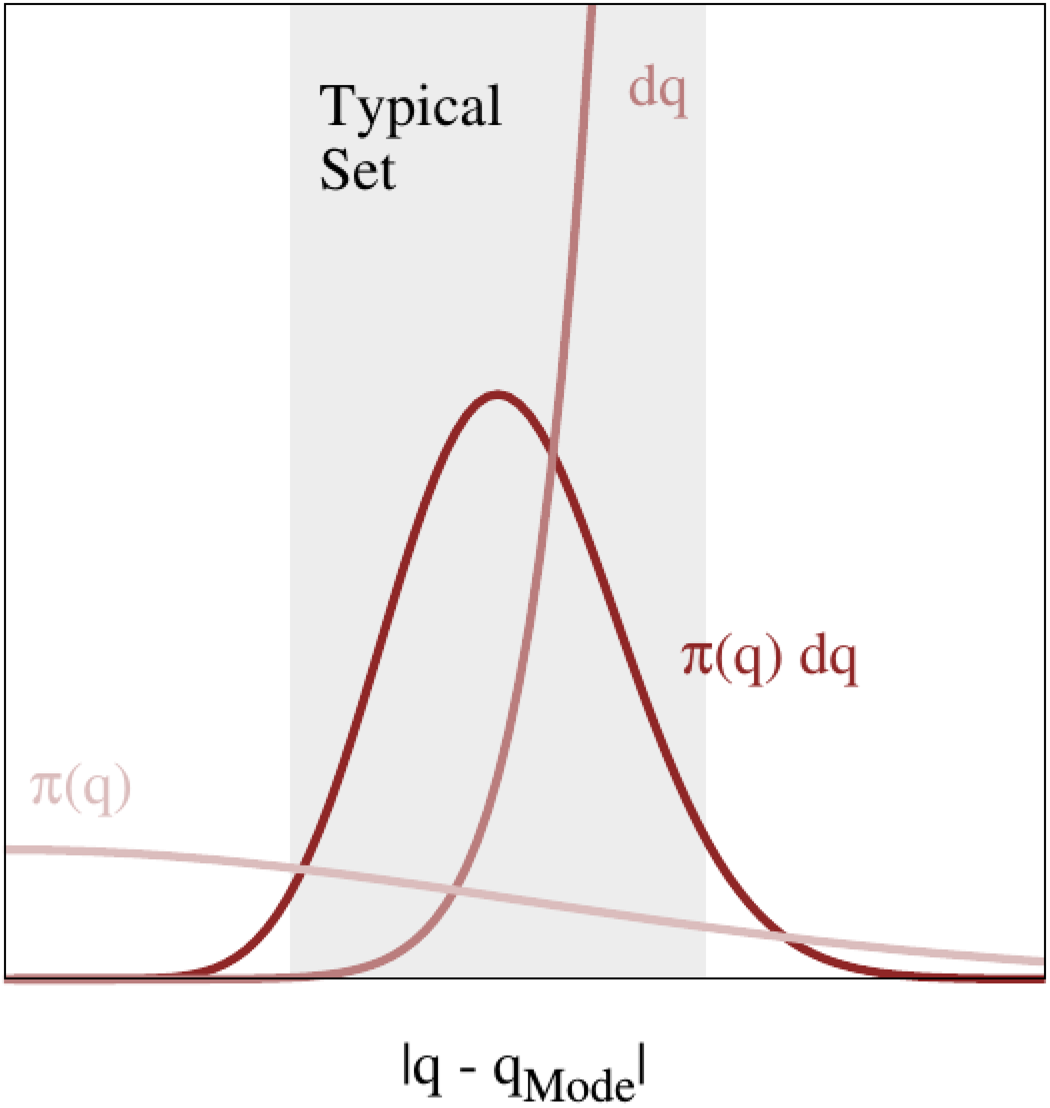
- Speeding up MCMC for big data.

- Subsample the data and compute the likelihood
  - $f'(x, y), y$ set of rows to use
  - $f'(x, y) = \exp[\sum_i \log f(x_i)]$, for each i in y

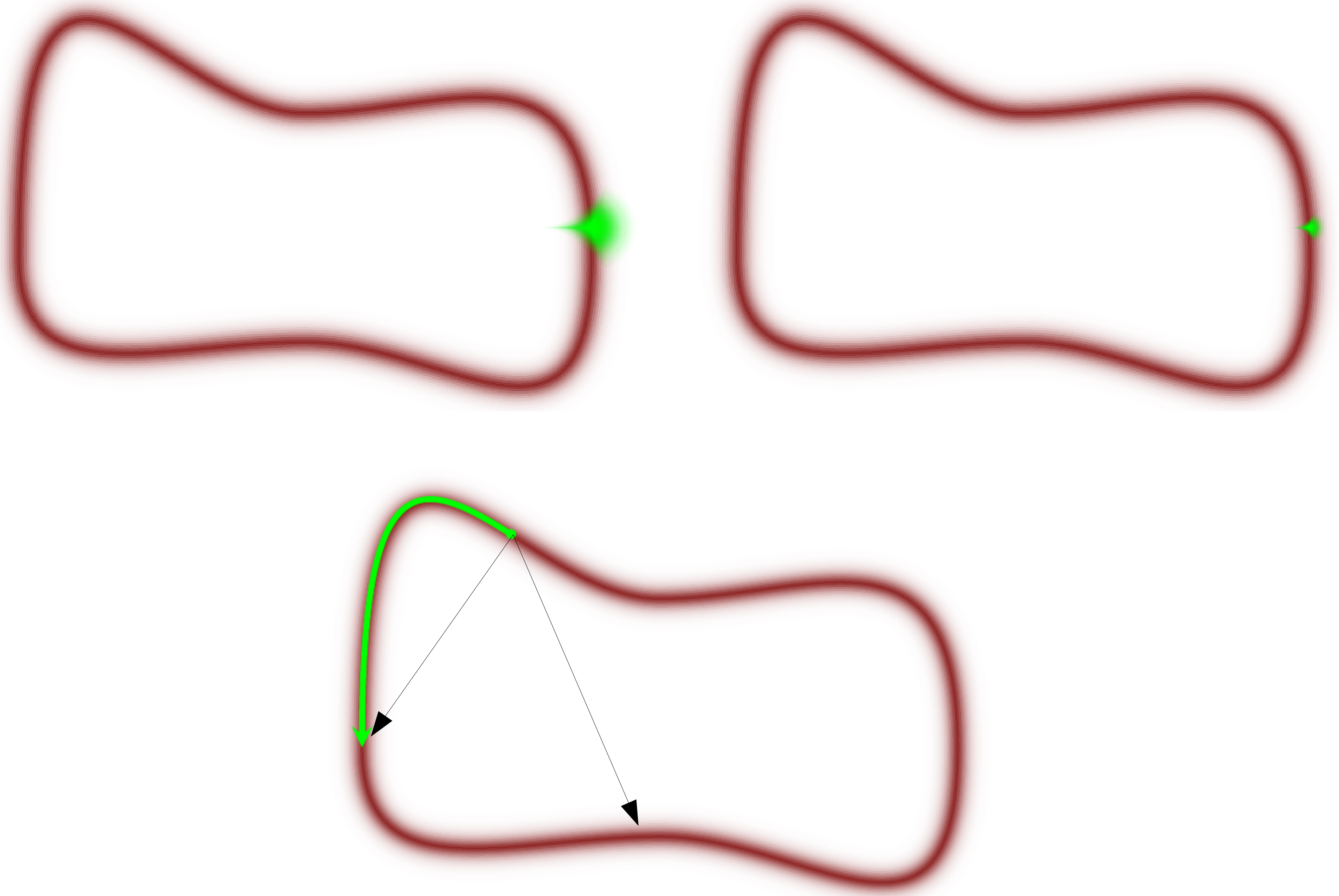
- Unbiased for $\log(f(x))$ but this does not give an unbiased estimator of $f(x)$.
- Bardent 2014, Korattikara 2014, Maclaurin & Adams 2014, Quiroz (2016,2017).

Hamiltonian Monte Carlo
(Duane 1987, Neal 1995)

- When d is large the typical set is confined to a thin shell.



Betancourt 2017



Jump to unexplored areas (like punching through a wormhole).

Betancourt 2017

Hamiltonian Monte Carlo

- $H = U(\theta) + K(u) = -\log p(\theta | x) + u^2/2$
- For $i = 0, M$:

```
    Sample new momentum- $u_i \sim N(0, 1)$
    Advance-    $(\theta', u') = \text{Leapfrog}(\theta_i, u_i)$
    if $U < \text{Min}(1, p(\theta', u')/p(\theta_i, u_i))$ :
        $(\theta_{i+1}, u_{i+1}) = (\theta', u')$
    else:
        $(\theta_{i+1}, u_{i+1}) = (\theta_i, u_i)$
```

HMC: caveats

- Need Gradients
  - Magic of Automatic differentiation
  - Driven by rapid advances in machine learning
- Tuning of stepsize :
  - The No-U-Turn-Sampler (NUTS)
    - Hoffman & Gelman (2014)
- Solves the high d problem.
- What about large N problem?
  - Subsampling HMC,
  - Possible to do says Dang et al. (2017)
  - However, Betancourt 2015 says that it is difficult to do so, fundamental incompatibility.

Variational Bayes

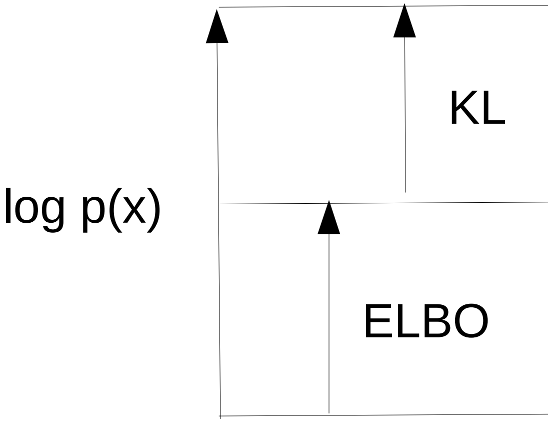
- Posterior:
- $p(\theta|x) = p(x|\theta)\, p(\theta) / p(x)$
- Approximate posterior by
- $q(\theta|\lambda)$

- KL: Kullback-Leibler divergence

$KL(q||p) = \int q(\theta|\lambda) \log [q(\theta|\lambda) / p(\theta|x)]$

$\lambda^* = \arg\min KL(\lambda)$

Note $p(x)$ is hard to compute



- ELBO: The Evidence Lower Bound

$ELBO(\lambda) = \int q(\theta|\lambda) \log p(\theta, x) / q(\theta|\lambda)$

$\log p(x) = KL(\lambda) + ELBO(\lambda)$

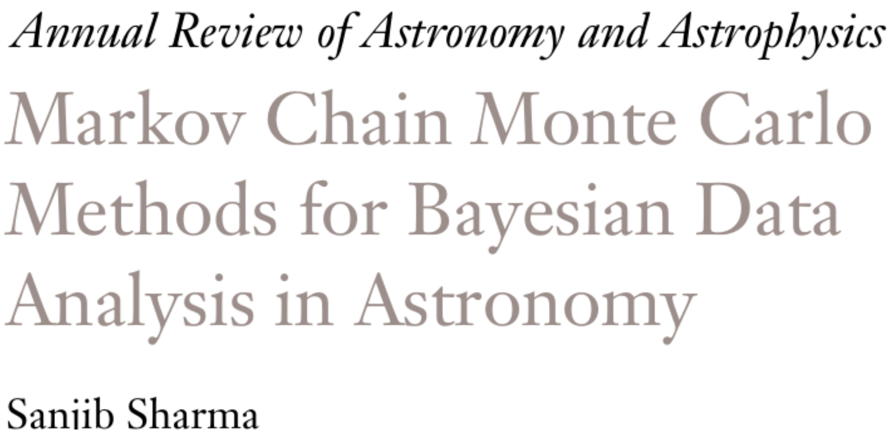
$\lambda^* = \arg\max ELBO(\lambda)$

Variational Bayes

- Reduces from sampling to an optimization problem.
- ADVI:
  - Automatic differentiation, Variational inference
  - Leveraging advances in ML
  - Stan, Edward, (Kucukelbir 2017)
  - *“Black box inference”* just like we had for MCMC
- Works both for large N and d.

Summary

- Hierarchical Bayesian models allow you to tackle a wide range of problems in astronomy.
- Large N: Bayesian nonparametric modelling.
- Large dim d -Hamiltonian Monte Carlo
- Large N, large d- Variational Bayes.
- For more info and Monte Carlo based algorithms to solve Bayesian inference problems see, Sharma 2017



Annual Review of Astronomy and Astrophysics

Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy

Sanjib Sharma