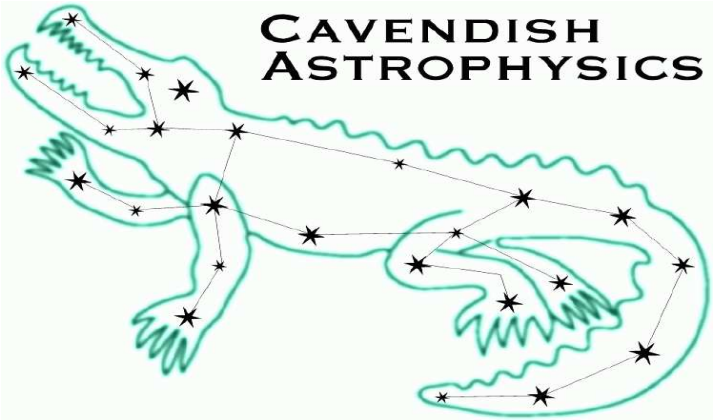
Multimodal Nested Sampling

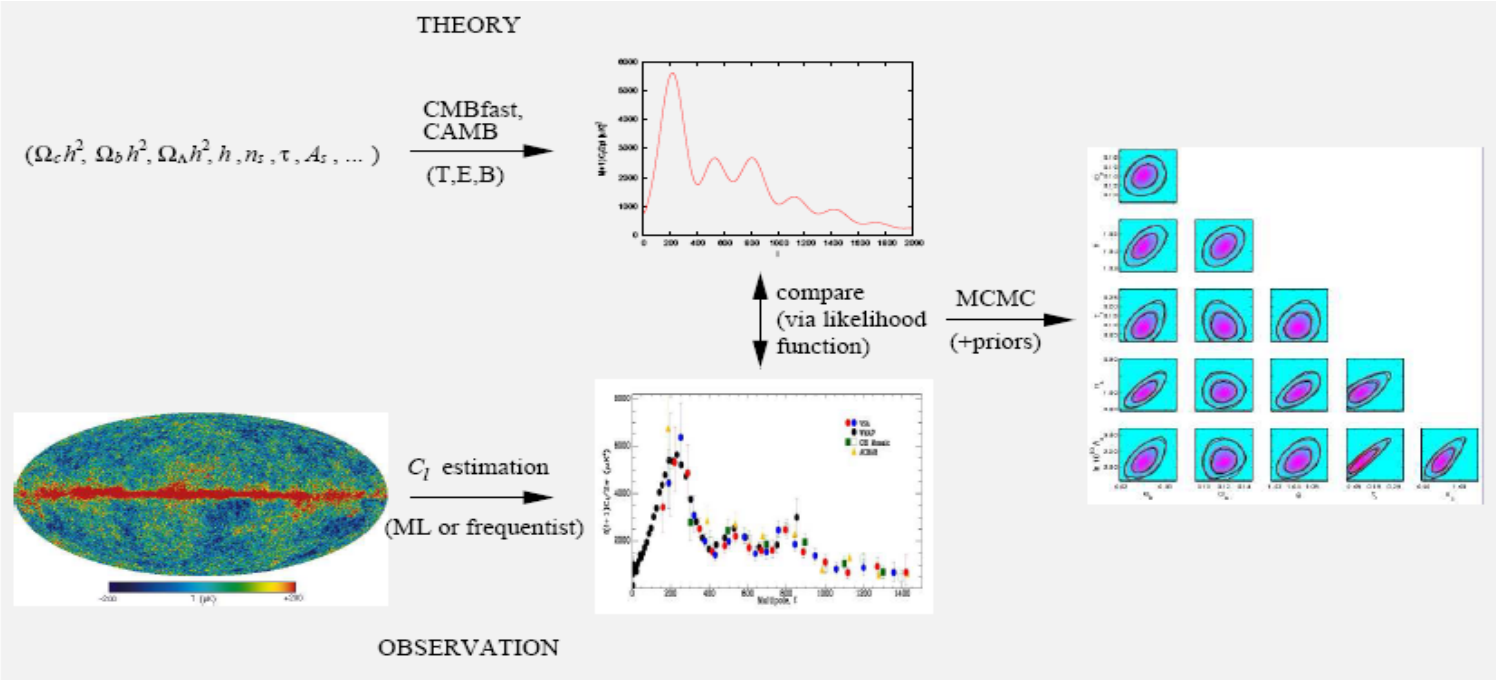


Farhan Feroz

Astrophysics Group, Cavendish Lab, Cambridge

Inverse Problems & Cosmology

- Most obvious example: standard CMB data analysis pipeline

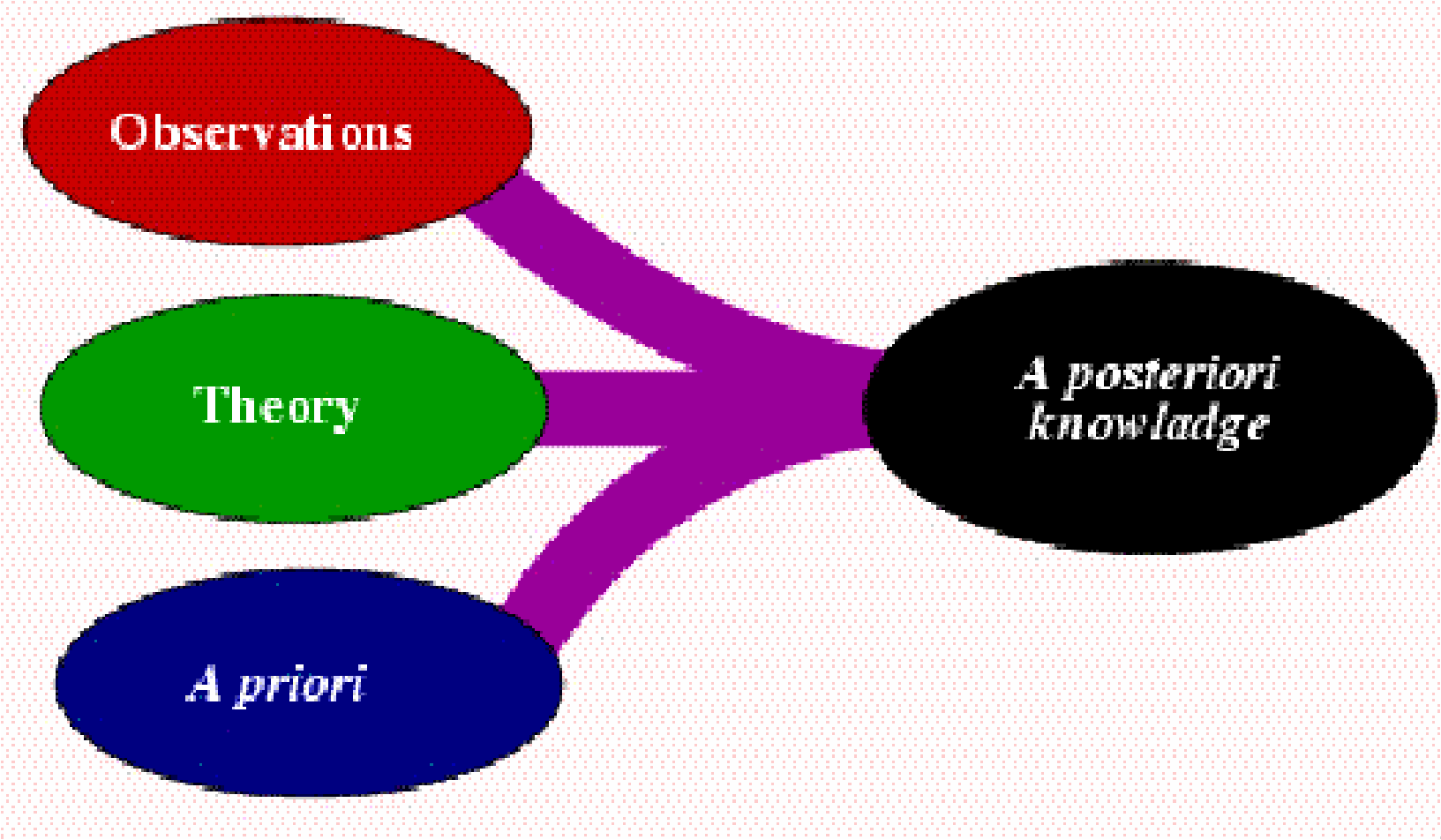


- But many others: object detection, signal enlargement, signal separation, …

Bayesian Inference

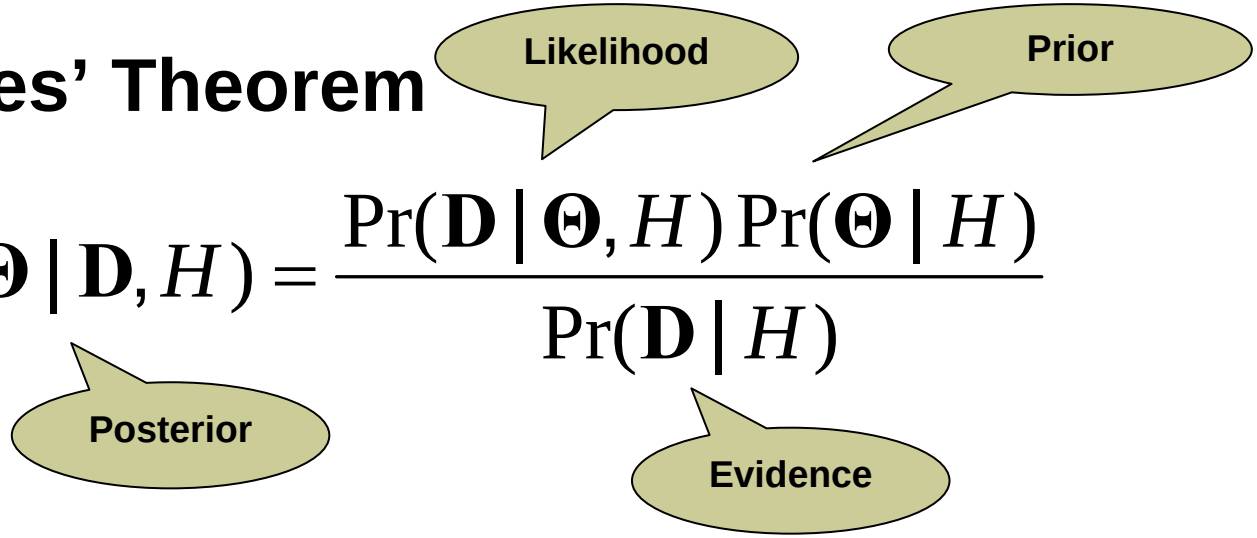
Definition: “an approach to statistics in which all forms of uncertainty are expressed in terms of probability” (Radford M. Neal)

The Bayesian Way



Bayesian Inference

- **Bayes' Theorem**



- **Model Selection**

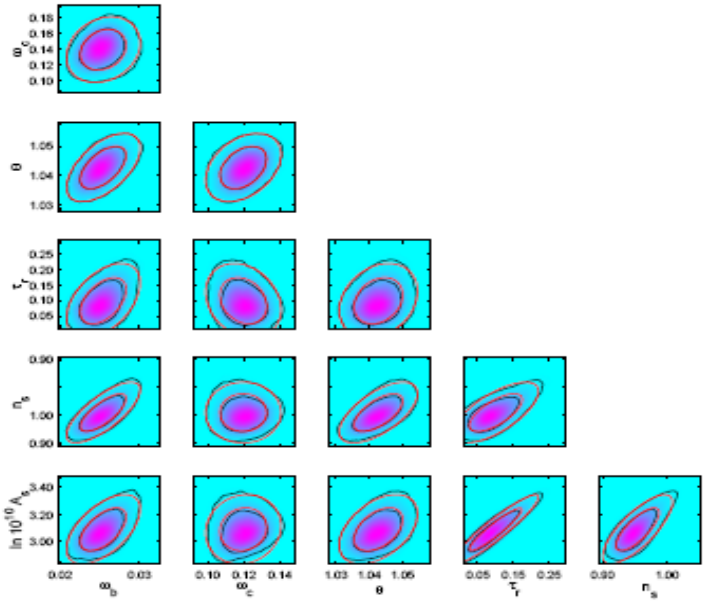
$$\frac{\Pr(H_1 \mid \mathbf{D})}{\Pr(H_0 \mid \mathbf{D})} = \frac{\Pr(\mathbf{D} \mid H_1)\Pr(H_1)}{\Pr(\mathbf{D} \mid H_0)\Pr(H_0)}$$

Bayesian Computation

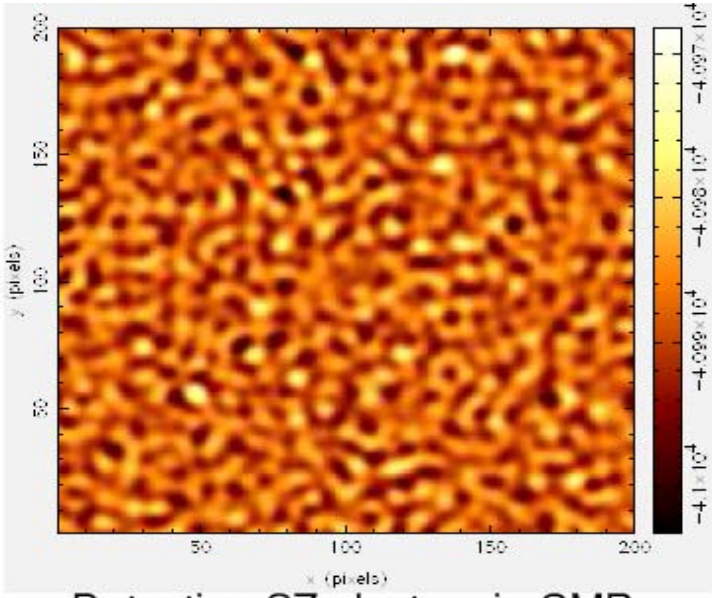
- Priors and posteriors are often complex distributions
- May not be easily represented as formulas
- Represent the distribution by drawing random samples from it
  - Visualize these samples by viewing them or low-dimensional projections of them
  - Make Monte Carlo estimates for their probabilities and expectations
- Sampling from the prior is often easy, sampling from the posterior, difficult

Some Cosmological Posteriors

- Some are nice, others are nasty



ΛCDM: $\boldsymbol{\theta} = (\omega_b, \omega_c, \theta, \tau, \ln A, n_s)$
using CMB+SDSS+HST data
(Trotta 2004)



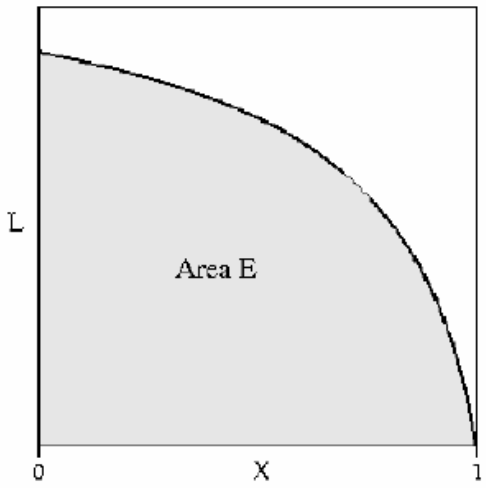
Detecting SZ clusters in CMB:
$\boldsymbol{\theta} = (X, Y, A, R)$
(Hobson & McLachlan 2003)

- Maximization (local or global) and covariance matrices → partial information → better to sample from the posterior using MCMC

Bayesian Evidence

- Evidence = $Z = \int L(\theta)\pi(\theta)d\theta$
- Evaluations of the *n*-dimensional integral presents great numerical challenge
- If dimension *n* of parameter space is small, calculate unnormalized posterior $\overline{P}(\theta) = L(\theta)\pi(\theta)$ over grid in parameter space → get evidence trivially
- For higher-dimensional problems, this approach rapidly becomes impossible
  - Need to find alternative methods
  - Gaussian approximation, Savage-Dickey ratio
- Evidence evaluation at least an order of magnitude more costly than parameter estimation.

Nested Sampling

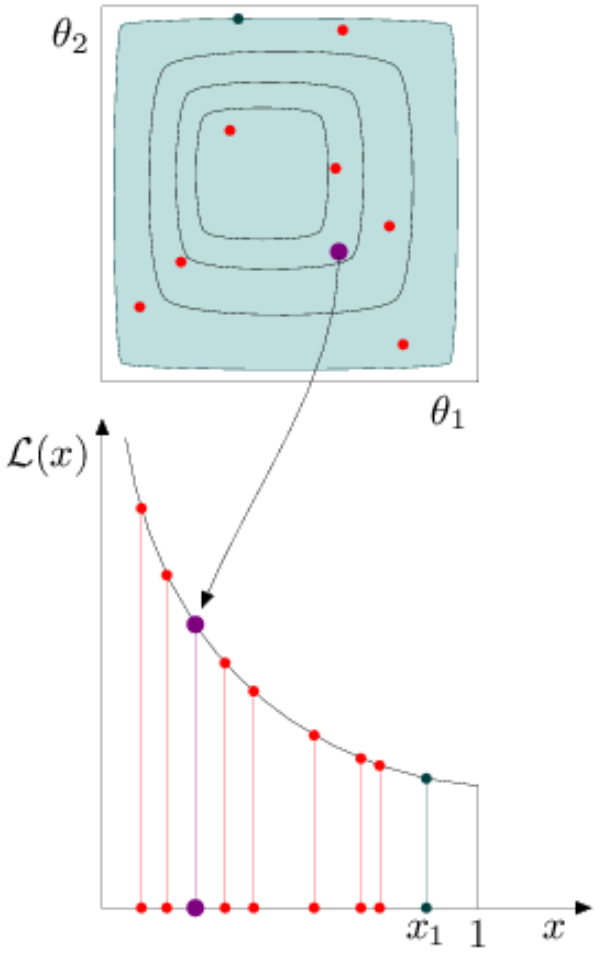


- Introduced by John Skilling in 2004.
- Monte Carlo technique for efficient evaluation of the Bayesian Evidence.
- Re-parameterize the integral with the prior mass X defined as, $dX = \pi(\theta) d^n\theta$, so that

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta) d^n\theta$$

- X defined such that it uniquely specifies the likelihood $Z = \int_0^1 L(X)\, dX$
- Suppose we can evaluate $L_j = L(X_j)$ where $0 < X_m < \ldots < X_2 < X_1 < 1$ then $Z = \sum_{j=1}^{m} L_j w_j$ where $w_j = (X_{j-1} - X_{j+1})/2$

Nested Sampling: Algorithm



1. Set $j = 0$; initially $X_0 = 1$, $Z = 0$
2. Sample N 'live' points uniformly inside the initial prior space ($X_0 = 1$) and calculate their likelihoods
3. Set $j = j + 1$
4. Find the point with the lowest L_i and remove it from the list of 'live' points
5. Increment the evidence as
   $Z = Z + L_i \, (X_{i-1} - X_{i+1}) / 2$
6. Reduce the prior volume $X_i / X_{i-1} = t_i$ where
   $P(t) = N \, t^{N-1}$
7. Replace the rejected point with a new point sampled from $\pi(\theta)$ with hard-edged region $L > L_i$
8. If $L_{\max} X_j < \alpha Z$ then set $Z = Z + \sum_{i=1}^{N} L(\theta_i) / N$
   stop else goto 3

Error Estimation

- Bulk of posterior around $X \approx e^{-H}$ where H is the information

$$H = \int \log(dP/dX)dX \quad \text{where } dP = LdX/Z$$

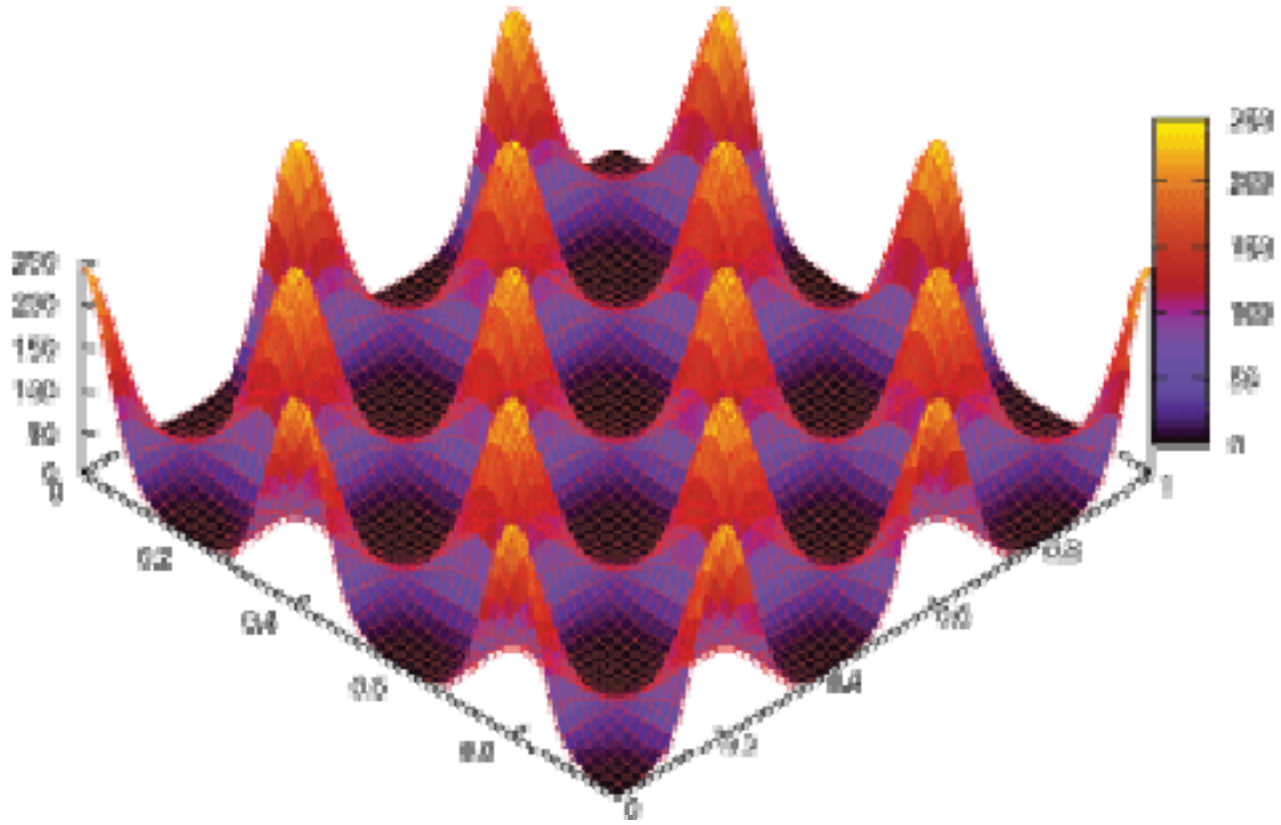
- Since $\log X_i = (i \pm \sqrt{i})/N$, we expect the procedure to take $NH \pm \sqrt{NH}$ steps to shrink down the bulk of posterior

- Dominant uncertainty in Z is due to the Poisson variability in the number of steps, $NH \pm \sqrt{NH}$, required to reach the bulk of posterior

- $\log X_i$ and $\log Z$ are subject to standard deviation uncertainty of $\sqrt{H/N}$

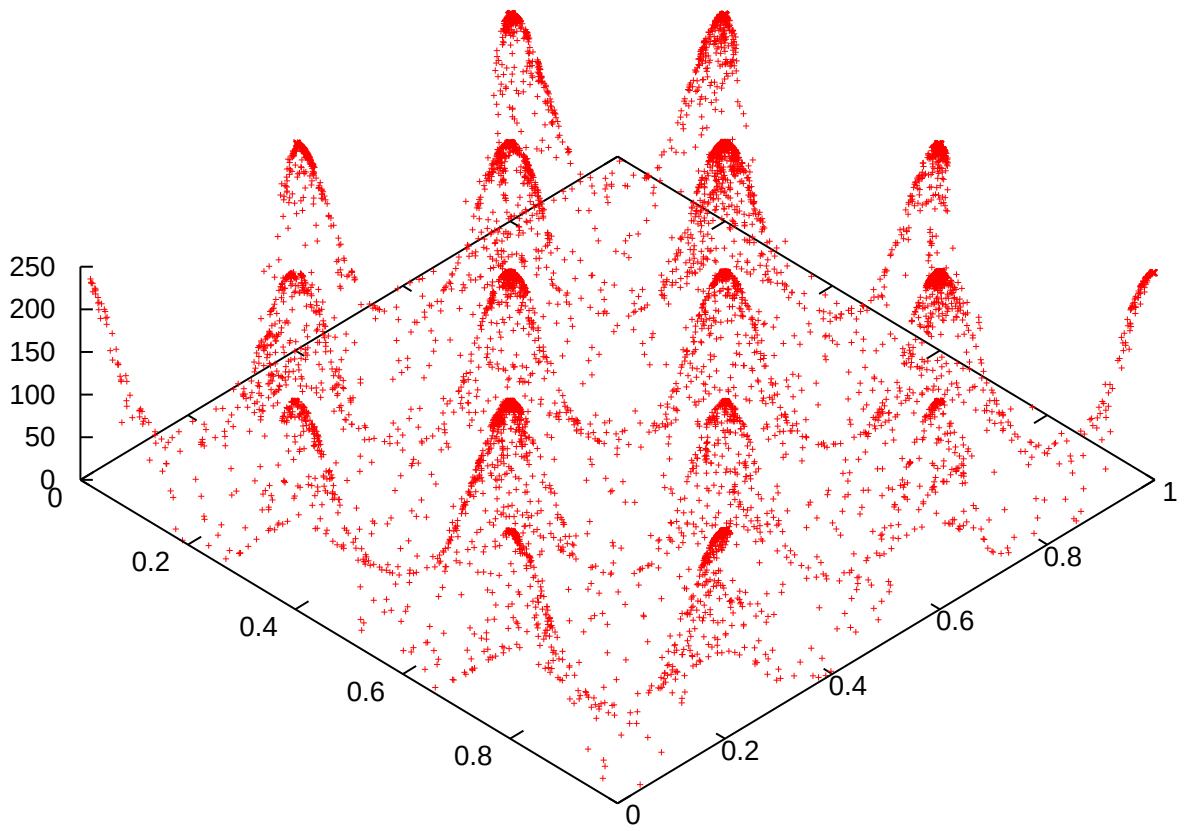
$$\therefore \log Z = \log \sum_i \left[L_i \frac{(X_{i-1} - X_{i+1})}{2} \right] \pm \sqrt{\frac{H}{N}}$$

Nested Sampling: Demonstration



Egg-Box Posterior

Nested Sampling: Demonstration



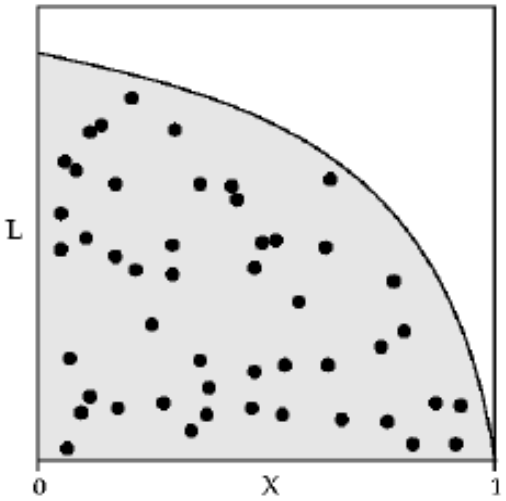
Egg-Box Posterior

Nested Sampling

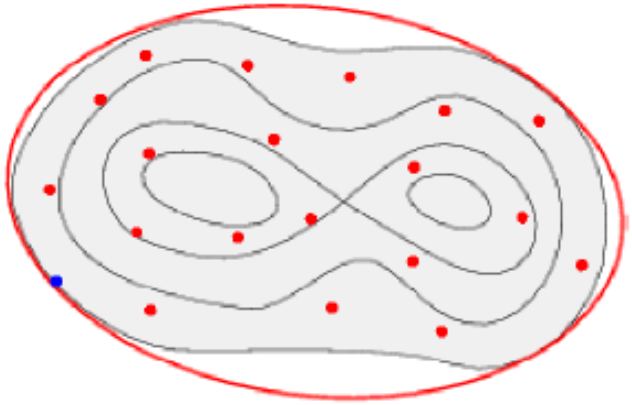
- Advantages:
  - Typically requires around 100 times fewer samples than thermodynamic integration for evidence calculation
  - Does not get stuck at phase changes
  - Parallelization possible if efficiency is known

- Bonus: posterior samples easily obtained as by-product Take full sequence of rejected points, θ_i, & weigh i^{th} sample by $p_i = L_i w_i / Z$

- Problem: must sample efficiently from prior within complicated, hard-edged likelihood constraint. MCMC can be inefficient

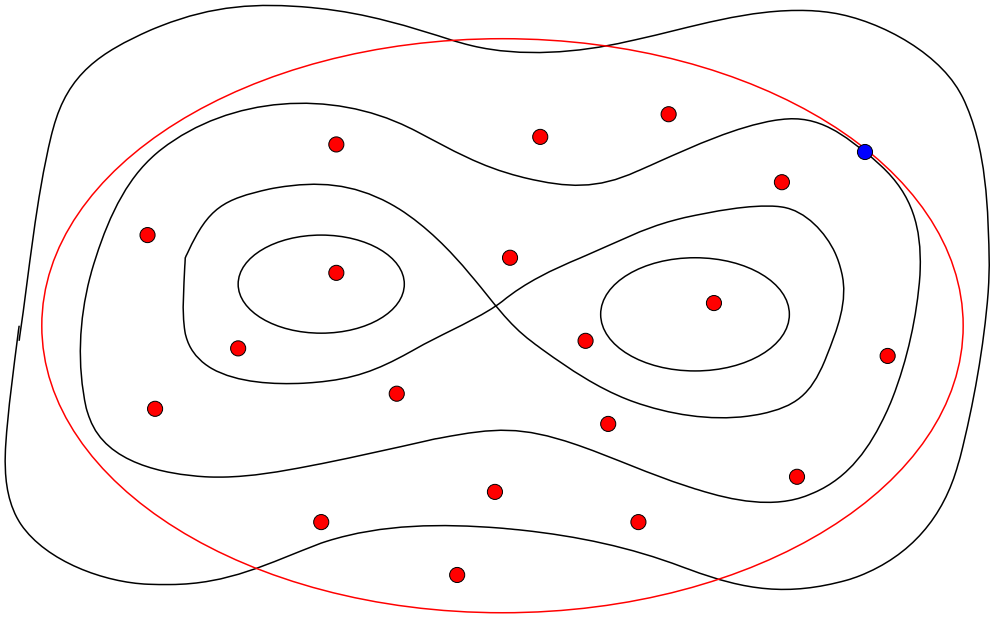


Ellipsoidal Nested Sampling

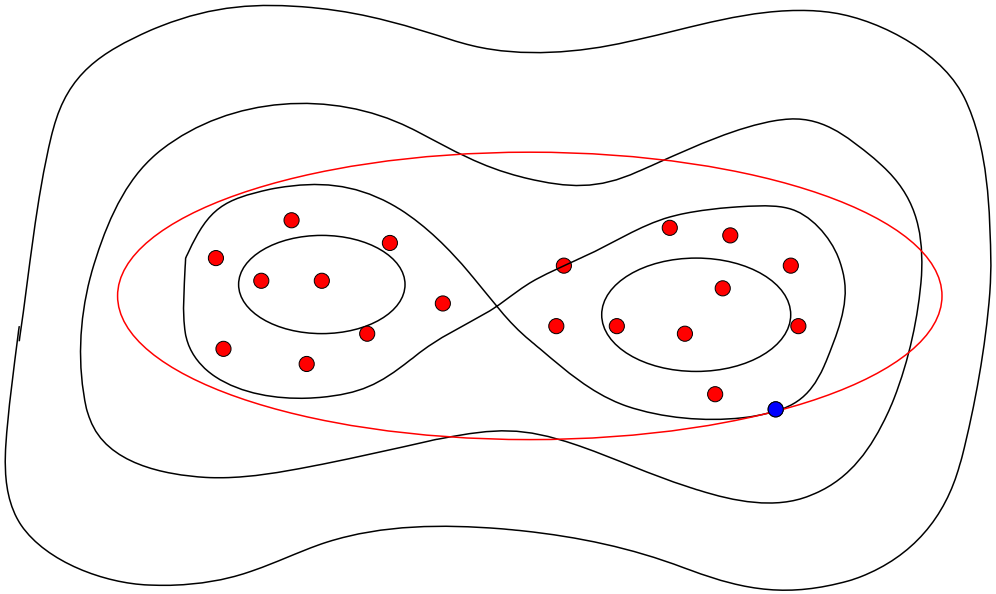


- Mukherjee et al. (2006) introduced ellipsoidal bound for the remaining prior volume with hard constraint, $L > L_i$, at each iteration
- Construct an *n*-dimensional ellipsoid using the covariance matrix of the current live points
- Enlarge this ellipsoid by some enlargement factor (*f*)
- Easily extendable to multi-modal problems through clustering

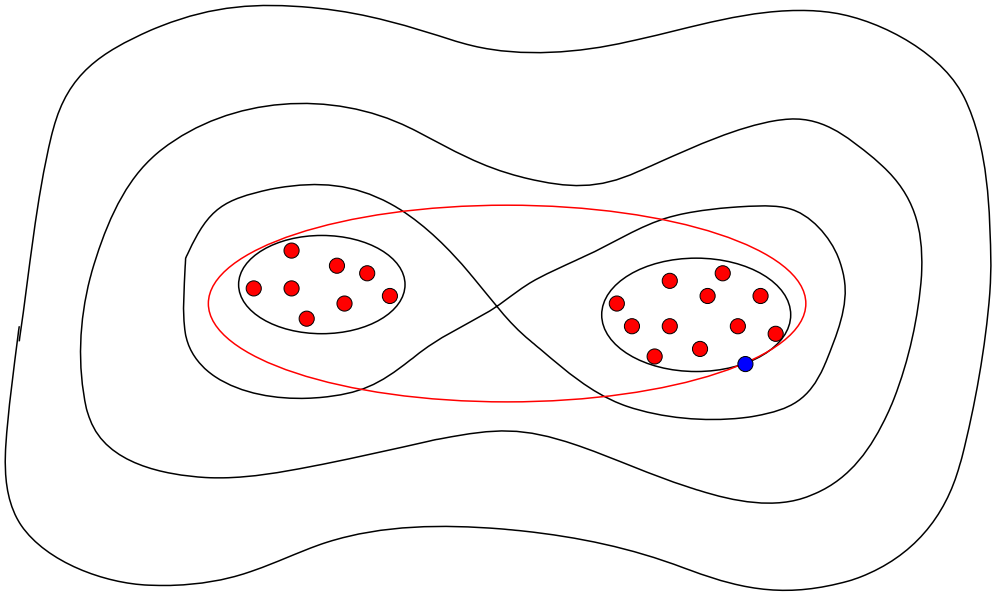
…Ellipsoidal Nested Sampling



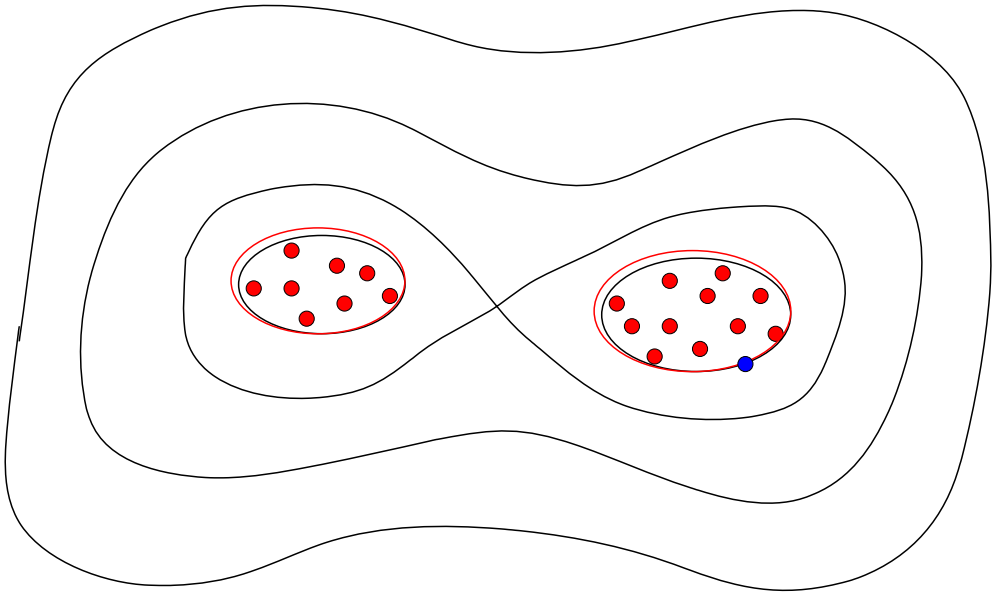
…Ellipsoidal Nested Sampling



…Ellipsoidal Nested Sampling - Problems



…Ellipsoidal Nested Sampling - Solution



Simultaneous Nested Sampling

- Introduced by Feroz & Hobson (2007) (arXiv:0704:3704)
- Improvements over recursive ellipsoidal nested sampling
  - Non-recursive so requires fewer likelihood evaluations in multimodal problems
  - Identify the number of clusters using X-means
  - Can use ellipsoidal, Metropolis or any other sampling method to sample from the hard constraint
  - Evaluation of ‘local’ as well as ‘global’ evidence values

Identification of Clusters

- Infer appropriate number of clusters from the current live point set using X-means (Pelleg et al. 2000)
- X-means: partition into the number of clusters that optimizes the Bayesian Information Criteria (BIC)
- X-means performs well overall but has some inconsistencies

Evaluation of 'Local' Evidences

- In simultaneous nested sampling, if a cluster is non-intersecting with its sibling and non-ancestor clusters, it is added to the list of 'isolated' clusters

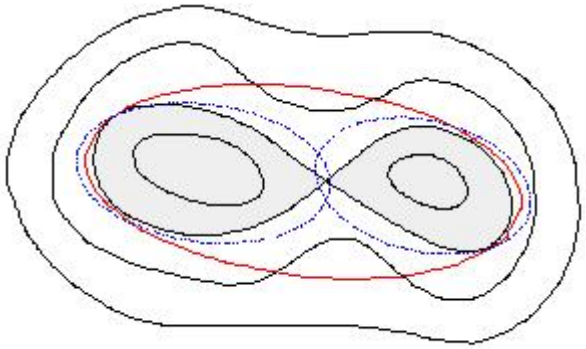
- Sum the evidence contributions from the rejected points inside this 'isolated' cluster to the local evidence of the corresponding mode

- Underestimated local evidence of the modes that are sufficiently close
  - Store information about clusters of the past few iterations
  - Match the 'isolated' clusters with the ones at the past iterations and increment its local evidence if the rejected points in those iteration fall into its matched clusters

Sampling from Overlapping Ellipsoids

- k clusters at iteration i with $n_1, n_2, \ldots, n_k$ points and $V_1, V_2, \ldots, V_k$ volumes of the corresponding (enlarged) ellipsoids

- Choose an ellipsoid with probability $p_k = V_k / V_{tot}$, where $V_{tot} = \sum_{j=1}^{k} V_j$

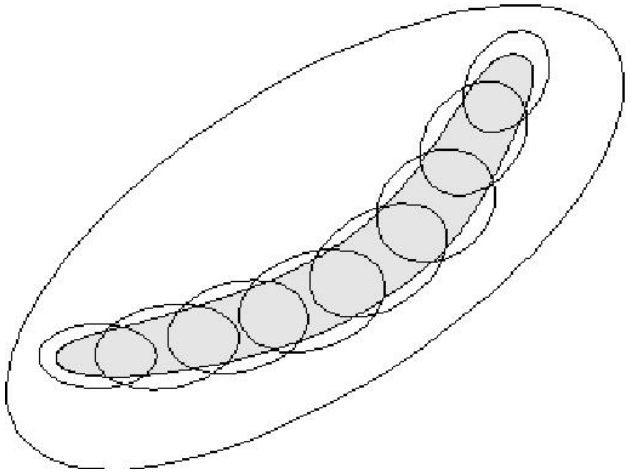


- Sample from the chosen ellipsoid with the hard constraint $L > L_i$

- Find the number n, of ellipsoids the chosen sample lies and accept the sample with probability $1 / n$

Dealing with Degeneracies

- One ellipsoid is a very bad approximation to a banana shaped likelihood region
- Sub-cluster every cluster found by X-means
- Minimum number of points in each sub-cluster being (D + 1) with D being the dimensionality of the problem
- Expand these sub-clusters by sharing points with neighboring sub-clusters

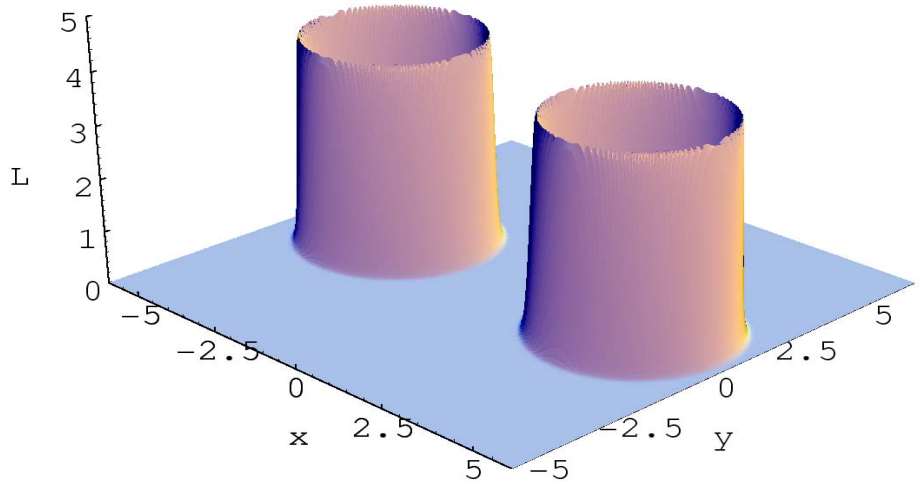


- Sample from them using the strategy outlined in previous section

Metropolis Nested Sampling (MNS)

- Replace ellipsoidal sampling in simultaneous ellipsoidal nested sampling by Metropolis-Hastings method
- Proposal distribution: Isotropic Gaussian with fixed width, σ, during a nested sampling iteration
- At each iteration, pick one of N live points randomly as the starting position for random walk
- Take n_s (=20) steps from the starting point with each new sample, x', being accepted if $L(x') > L_i$.
- Adjust σ after every nested sampling iteration to maintain the acceptance rate around 50%

Example: Gaussian Shells



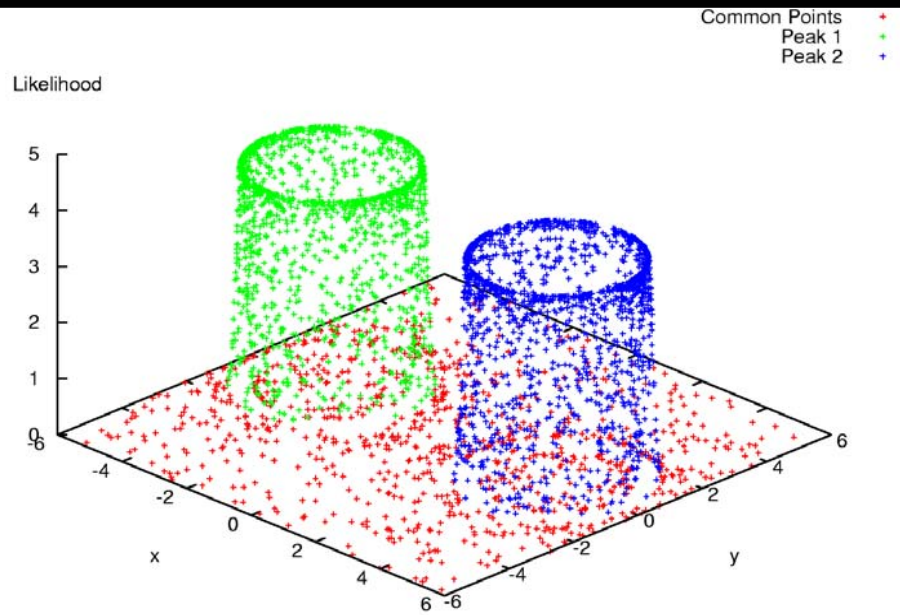
- Posterior defined as

$$L(\mathbf{x}) = circ(\mathbf{x}; c_1, r_1, w_1) + circ(\mathbf{x}; c_2, r_2, w_2), \text{ where}$$

$$circ(\mathbf{x}; c, r, w) = \frac{1}{\sqrt{2\pi w^2}} \exp\left[-\frac{\left(|\mathbf{x} - c| - r\right)^2}{2w^2} \right].$$

- Typical of degeneracies in many beyond-the-Standard-Model parameter space scans in Particle Physics

Gaussian Shells in 2D: Results



- $w_1 = w_2 = 0.1$, $r_1 = r_2 = 2$, $c_1 = (-3.5,0.0)$, $c_2 = (3.5,0.0)$
- Analytical Results: $\log Z = -1.75$, $\log Z_1 = -2.44$, $\log Z_2 = -2.44$
- Ellipsoidal & Metropolis Nested Sampling with $N_{like} \sim 20{,}000$
  $\log Z = -1.78 \pm 0.08$, $\log Z_1 = -2.49 \pm 0.09$, $\log Z_2 = -2.47 \pm 0.09$
- Bank sampler (modified Metropolis-Hastings, arXiv:0705.0486) required $N_{like} \sim 1 \times 10^6$, for parameter estimation and no evidence evaluation

Gaussian Shells upto 100D: Results

	Analytical		Metropolis Nested Sampling			
dim	log Z	local log Z^*	log Z	local log Z_1	local log Z_2	N_{like}
10	-14.6	-15.3	-14.6 ± 0.2	-15.4 ± 0.2	-15.3 ± 0.2	127,463
30	-60.1	-60.8	-60.1 ± 0.5	-60.4 ± 0.5	-61.3 ± 0.5	489,416
50	-112.4	-113.1	-112.2 ± 0.5	-112.9 ± 0.5	-113.0 ± 0.5	857,937
70	-168.2	-168.9	-167.5 ± 0.6	-167.7 ± 0.6	-170.7 ± 0.7	1,328,012
100	-255.6	-255.3	-254.2 ± 0.8	-254.4 ± 0.8	-256.7 ± 0.8	2,091,314

*analytically local log Z_1 = local log Z_2 = local log Z

Application: Astronomical Object Detection

- Main Problems:
  - Parameter estimation
  - Model comparison
  - Quantification of detection

Quantifying Cluster Detection

- $$R = \frac{\Pr(H_1 \mid \mathrm{D})}{\Pr(H_0 \mid \mathrm{D})} = \frac{\Pr(\mathrm{D} \mid H_1)\Pr(H_1)}{\Pr(\mathrm{D} \mid H_0)\Pr(H_0)} = \frac{Z_1 \Pr(H_1)}{Z_0 \Pr(H_0)}$$

- H_0 = “there is no cluster with its center lying in the region S”
- H_1 = “there is one cluster with its center lying in the region S”

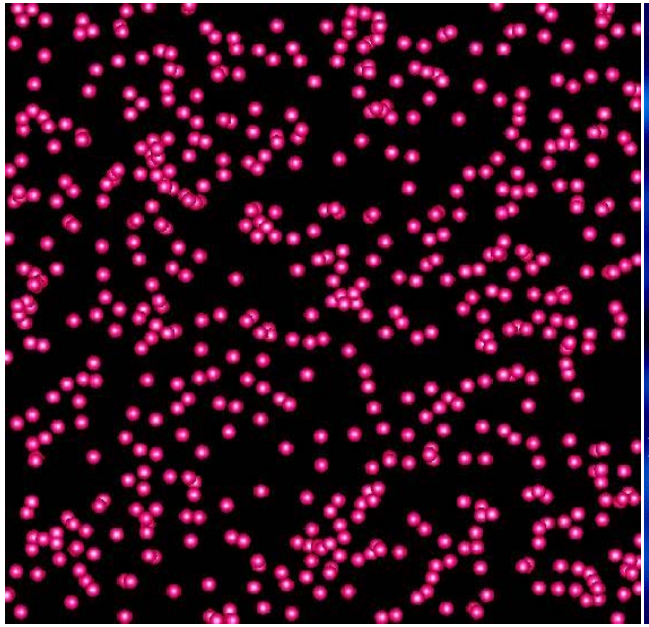
- $$Z_0 = \frac{1}{|S|}\int_S L_0 dX = L_0$$

- For clusters distributed according to Poisson distribution

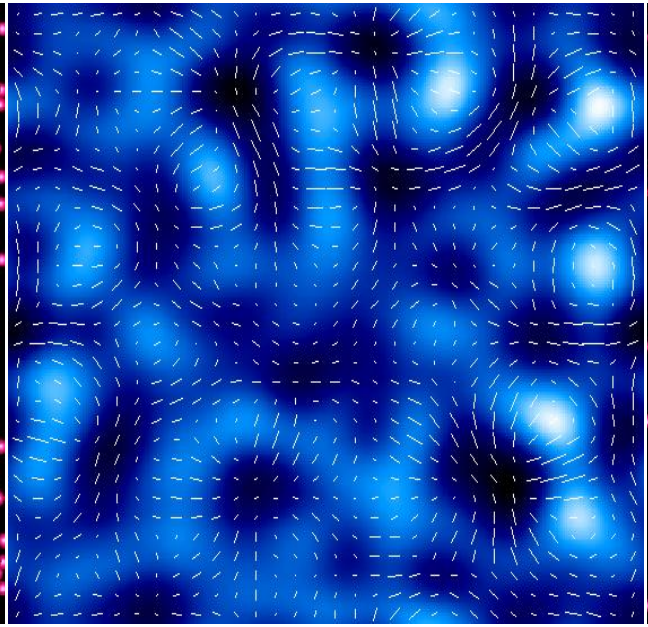
$$\frac{\Pr(H_1)}{\Pr(H_0)} = \mu_s$$

$$\therefore R = \frac{Z_1 \mu_s}{L_0}$$

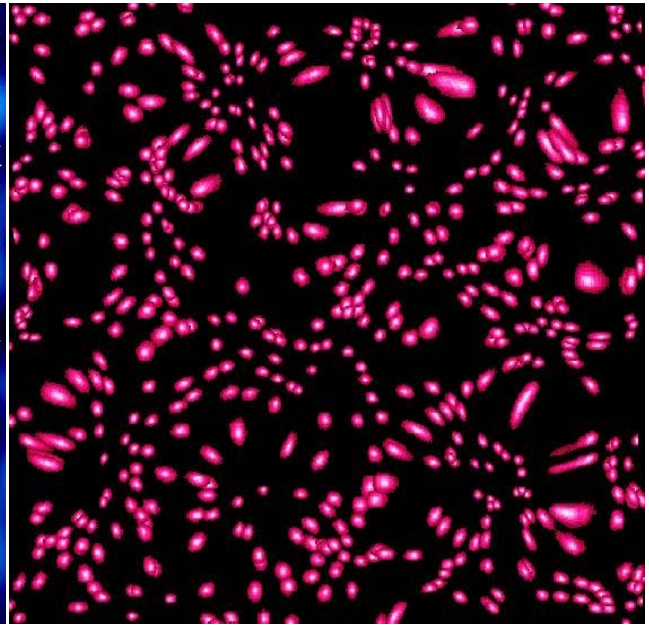
Weak Gravitational Lensing



unlensed galaxies

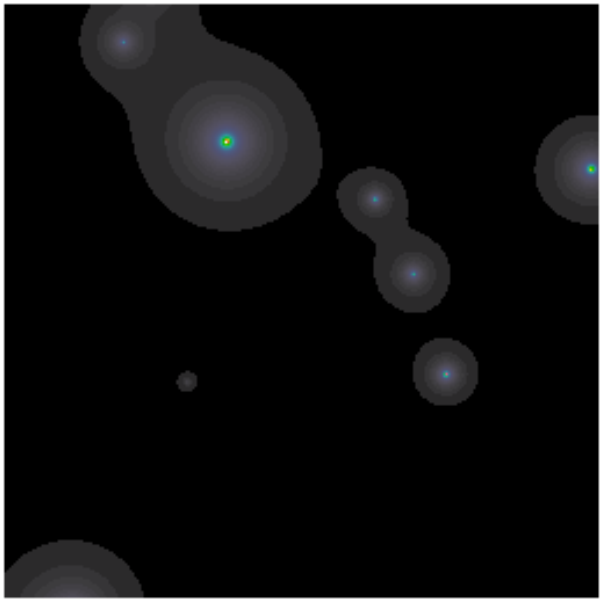


projected mass with shear map overlaid

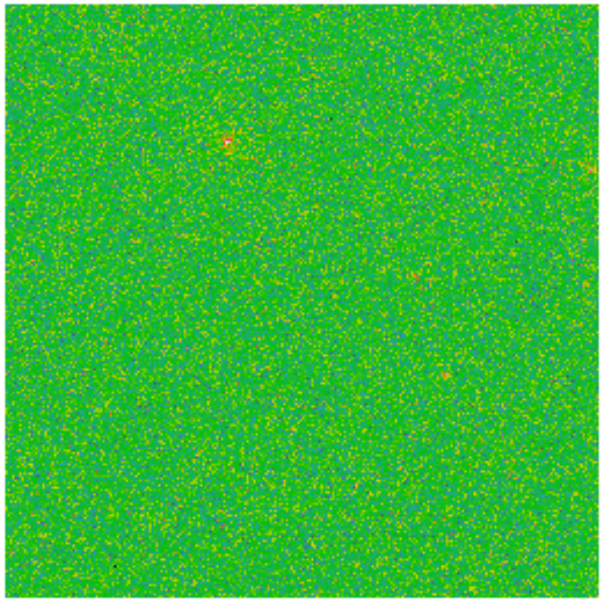


lensed galaxies

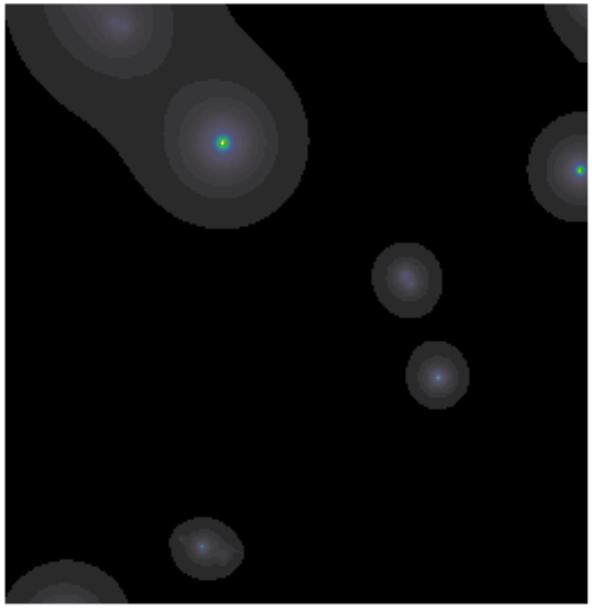
Wide Field Weak Gravitational Lensing



true convergence map



noisy convergence map

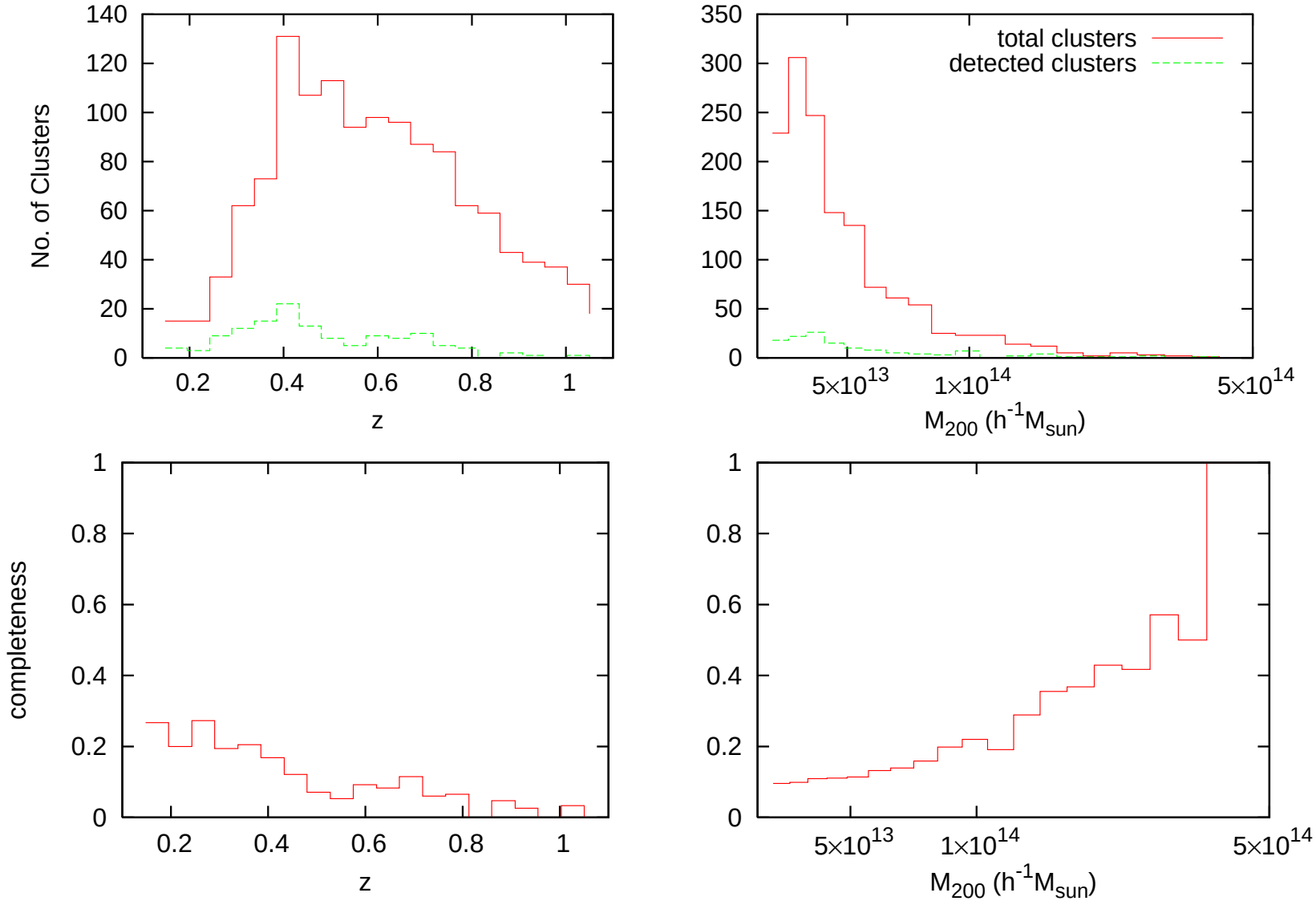


inferred convergence map

- 0.5 X 0.5 degree^2, 100 gal per arcmin^2 & σ = 0.3
- Concordance ΛCDM Cosmology with cluster mass & redshifts drawn from Press-Schechter mass function

Wide Field Lensing: Application to N-Body Simulations

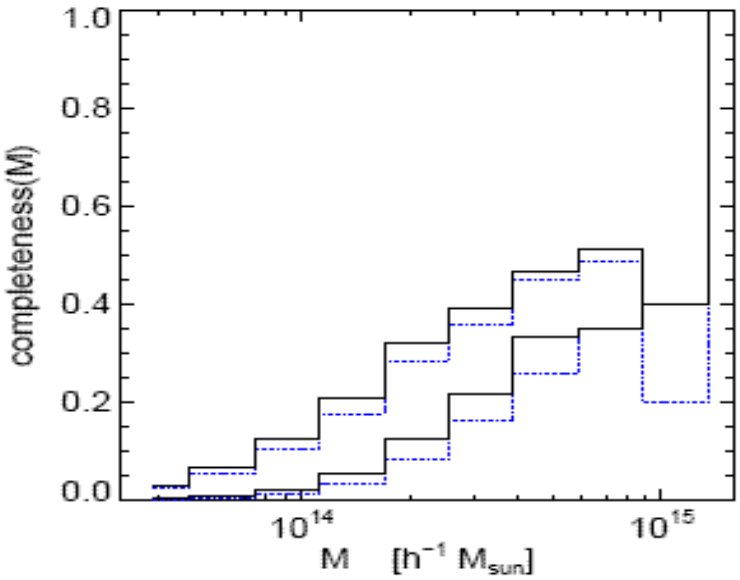
- Produced by Martin White, 2005
- Covering 3 X 3 degree^2
- Concordance ΛCDM Cosmology
- 65 galaxies per arcmin^2
- $\sigma = 0.3$
- 1350 halos with $M_{200} > 10^{13.5}\ h^{-1}\ M_{sun}$



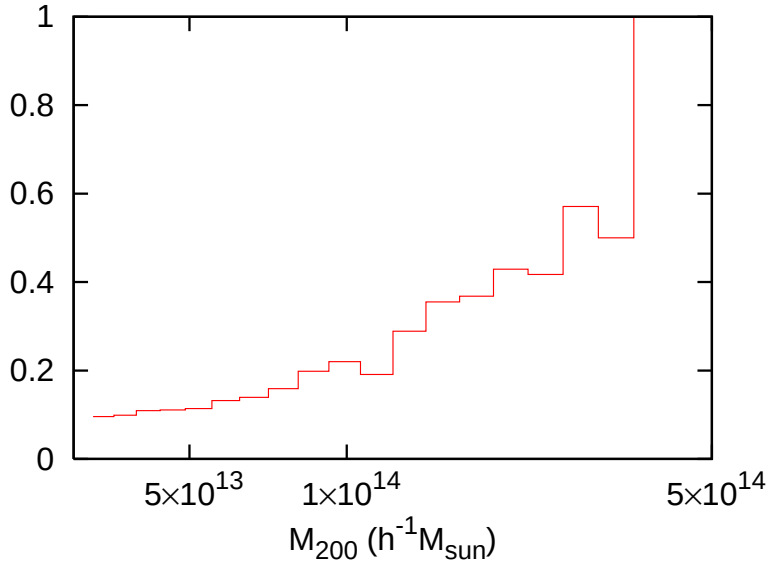
- 146 positive detections of which 131 are true

(In)completeness of Weak Lensing

Hennawi & Spergel, 2007



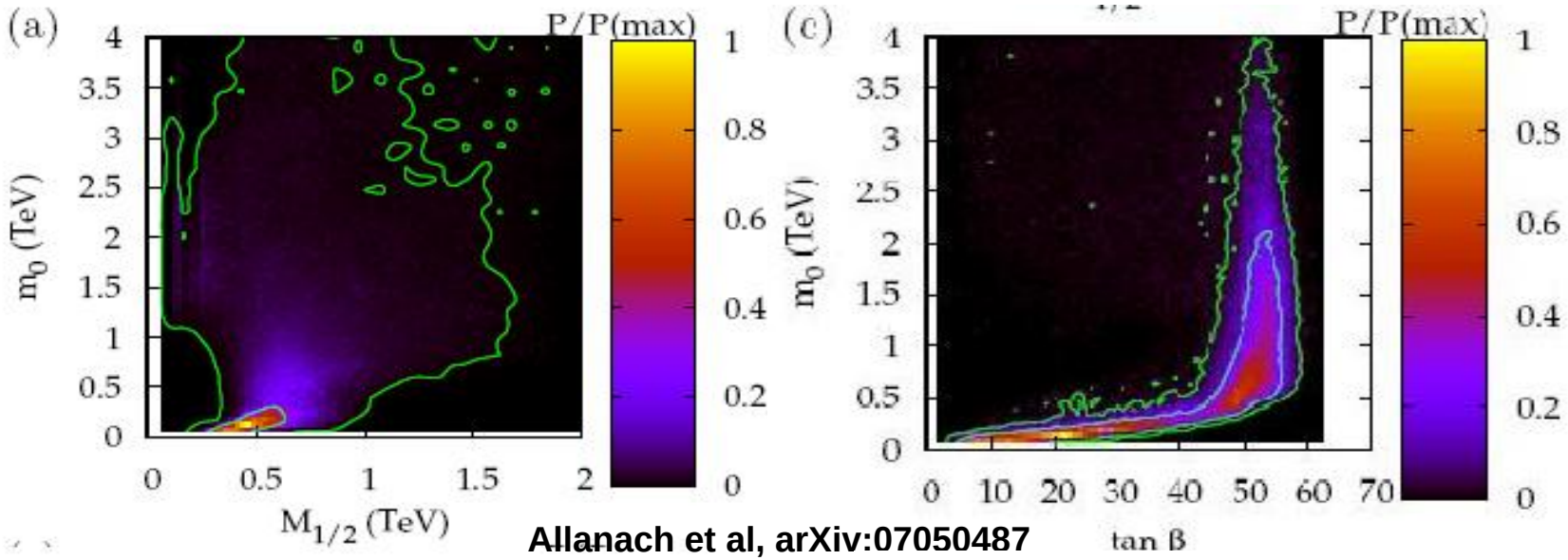
Feroz et al. in preparation



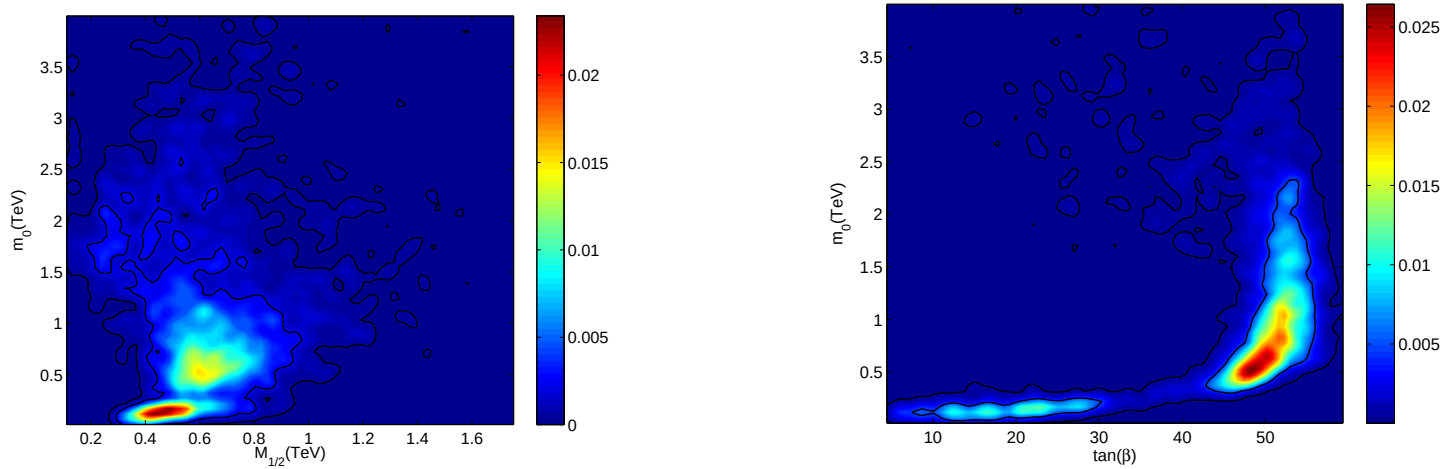
Bayesian Analysis of mSUGRA

- Most popular realization of MSSM with universal boundary conditions
- 5 mSUGRA parameters $(M_{1/2}, m_0, \tan\beta, A_0, \mathrm{sgn}(\mu))$ + Standard Model parameters
- Allanach et al. performed the bank sampler analysis → parameter constraints
- Bayesian evidence based model comparison vital for analyzing models of SUSY breaking at low energies using LHC data

Bayesian Analysis of mSUGRA: Results



Allanach et al, arXiv:07050487



Feroz et al. in preparation

Conclusions

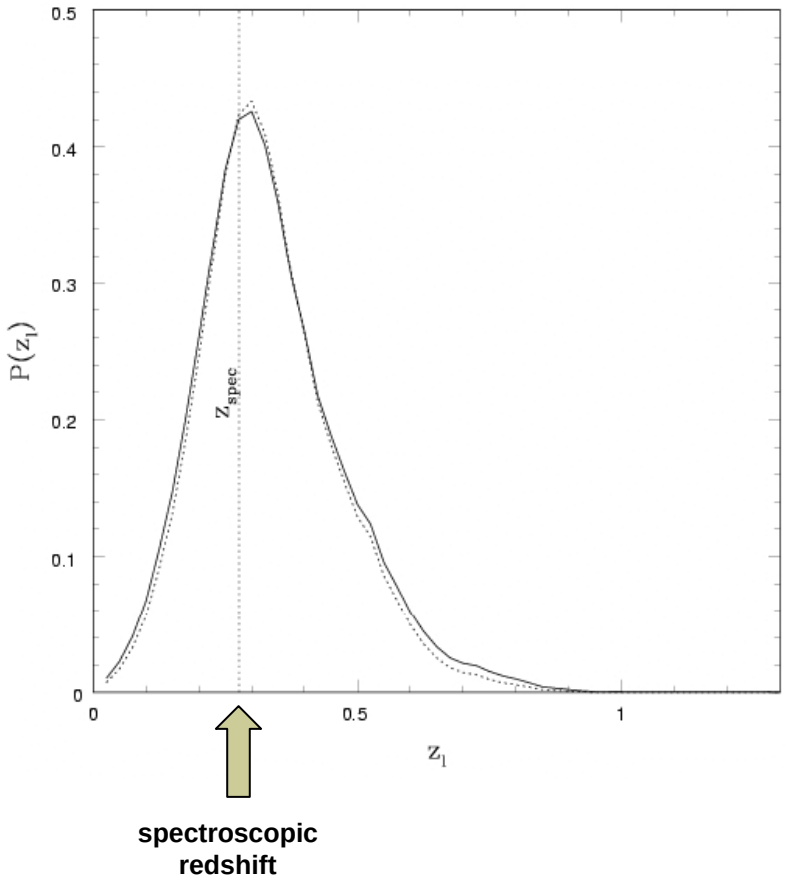
- Bayesian framework provides unified approach with 2 levels of inference
  - parameter estimation and confidence limits by maximising or exploring posterior
  - model selection by integrating posterior to obtain evidence

- Nested sampling efficient in both evidence evaluation and parameter estimation
  - main issue is sampling from prior within hard likelihood constraint
  - MCMC and ellipsoidal bound methods promising
  - clustering allows sampling from multimodal/degenerate posteriors

- Many cosmological and particle physics applications – so try it for yourself!

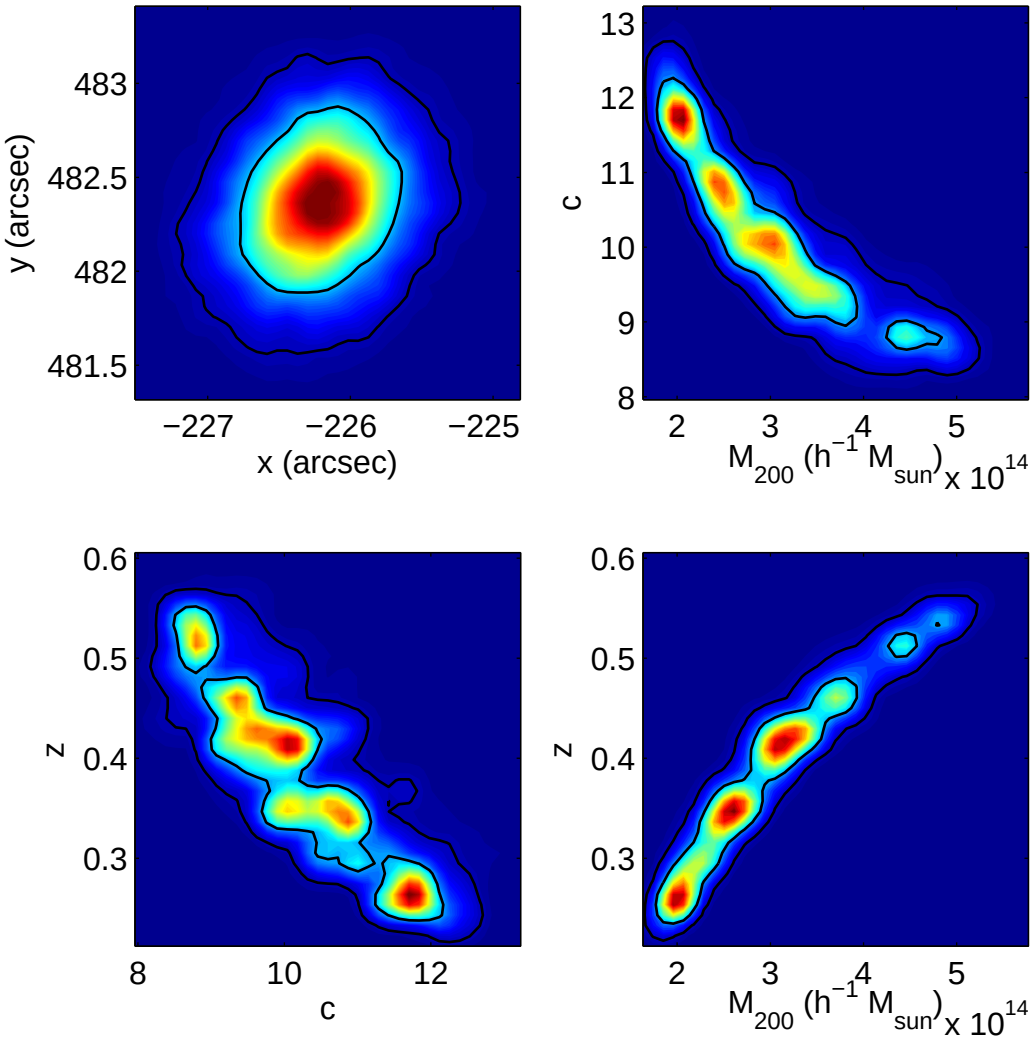
Cluster Tomography

Wittman et al., 2001

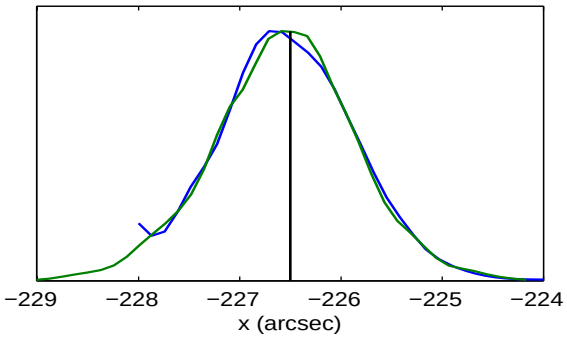


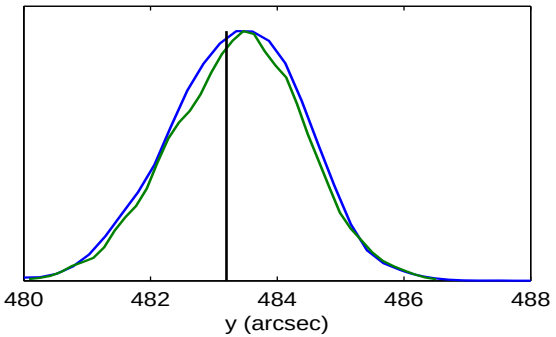
- Assume a mass profile & fit for shear as a function of source photometric redshift
- How reliable is this technique?

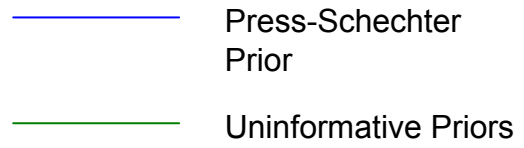
Weak Lensing: Parameter Constraints

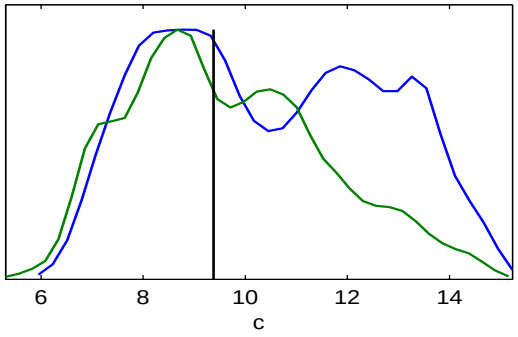


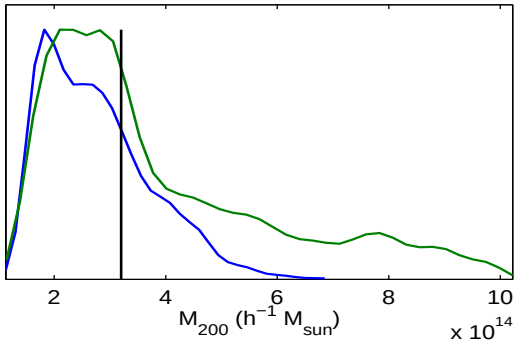
Weak Lensing: Parameter Constraints

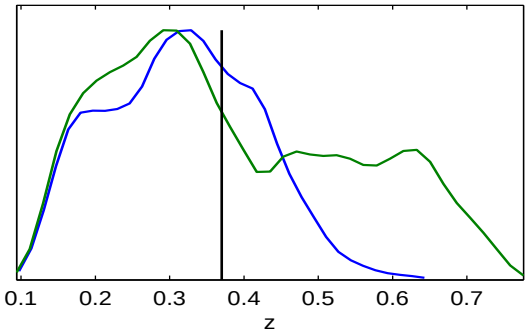




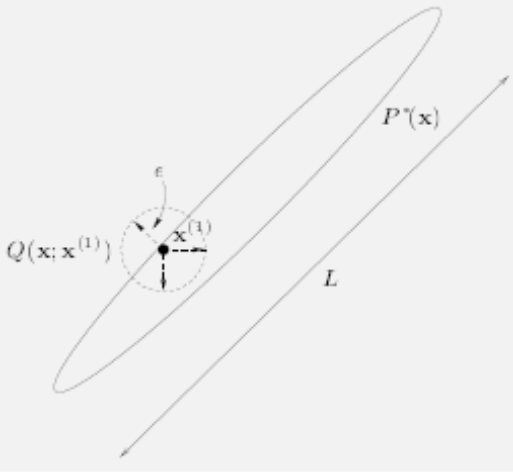








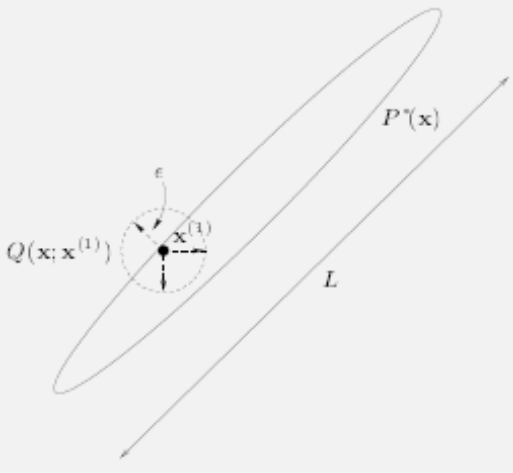
Metropolis Hastings Algorithm



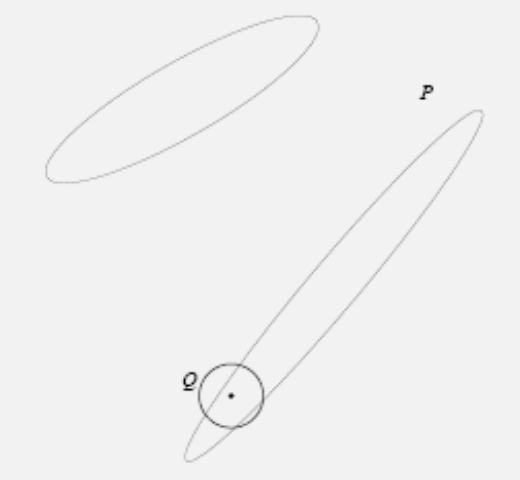
- Metropolis-Hastings algorithm to sample from $P(\theta)$
  - Start at an arbitrary point θ_0
  - At each step, draw a trial point, θ', from the proposal distribution $Q(\theta' \mid \theta_0)$
  - Calculate ratio $r = P(\theta') \, Q(\theta_n \mid \theta') \, / \, P(\theta_n) \, Q(\theta' \mid \theta_n)$
  - accept $\theta_{n+1} = \theta'$ with probability $max(1,r)$ else set $\theta_{n+1} = \theta_n$

- After initial burn-in period, any (positive) proposal $Q \longrightarrow$ convergence to $P(\theta)$
- Common choice of Q, multivariate Gaussian centred on θ_n but many others

Metropolis Hastings Algorithm – Some Problems



- Choice of proposal Q strongly affects convergence rate and sampling efficiency
  - large proposal width ε → trial points rarely accepted
  - small proposal width ε → chain explores $P(\theta)$ by a random walk → very slow
- If largest scale of $P(\theta)$ is L, typical diffusion time $t \sim (L/\varepsilon)^2$
- If smallest scale of $P(\theta)$ is l, need $\varepsilon \sim l$, diffusion time $t \sim (L/l)^2$



- Particularly bad for multimodal distributions
  - Transitions between distant modes very rare
  - No one choice of proposal width ε works
  - Standard convergence tests will suggest convergence, but actually only true in a subset of modes

Thermodynamic Integration

- MCMC sampling (with annealing) from full posterior requires no assumptions regarding hypotheses or priors

- Basic method is thermodynamic integration: define $Z(\lambda) = \int L^{\lambda}(\theta)\pi(\theta)d\theta$ so the required evidence value is $Z(1)$

- Begin MCMC sampling from $L^{\lambda}(\theta)\pi(\theta)d\theta$, starting with $\lambda = 0$ then slowly raising the value according to some annealing schedule until $\lambda = 1$.

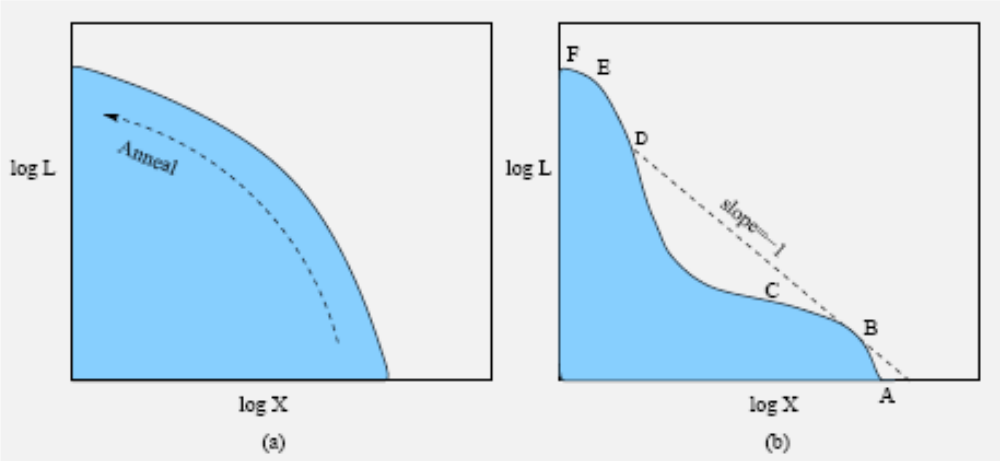
- Use the N_s samples corresponding to any particular value of λ to obtain an estimate of the quantity $\langle \log L \rangle_{\lambda}$

- But $\langle \log L \rangle_{\lambda} = \frac{1}{Z}\frac{dZ}{d\lambda} = \frac{d \log Z}{d\lambda}$, *so*

$$\log Z(1) = \log Z(0) + \int_0^1 \langle \log L \rangle_{\lambda} d\lambda \approx \sum_{j=1}^{N_j} \langle \log L \rangle_{\lambda_j} \Delta\lambda_j$$

…Thermodynamic Integration

- Problems:
  - Evidence value stochastic, need multiple runs to estimate the error on the evidence
  - Accurate evidence evaluation requires slow annealing
  - Common schedules (linear, geometric) can get stuck in local maxima
  - Can not navigate through phase changes



•Let $dX = \pi d\theta$: prior mass

•As $\lambda : 0 \rightarrow 1$, annealing should track along the curve

•But $d \log L / d \log X = -1/\lambda$ so annealing schedule can not navigate through convex regions (phase changes)