Sampling

Elena Sellentin

Sterrewacht Leiden

Universiteit Leiden

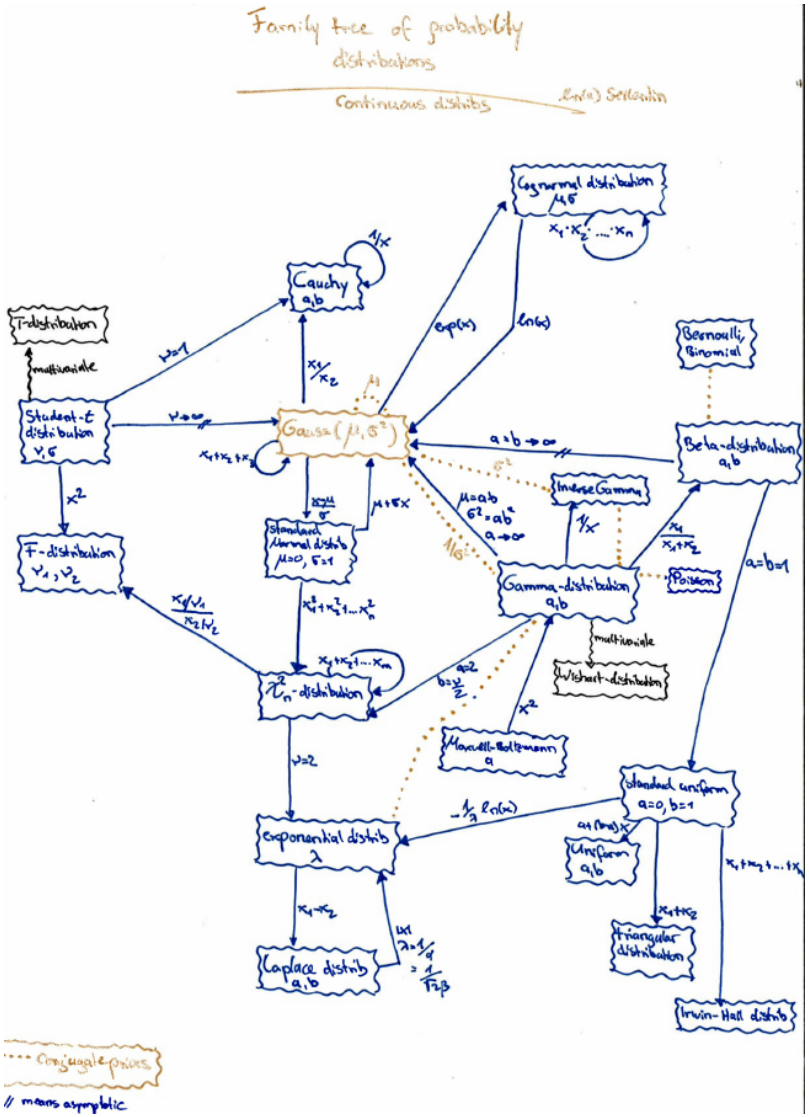
Inference

$$\mathcal{P}(\boldsymbol{\theta}, \mathcal{M}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta})}{\epsilon(\boldsymbol{x}|\mathcal{M})} \tag{1}$$

- $\mathcal{P}(\boldsymbol{\theta}, \mathcal{M}|\boldsymbol{x})$: the posterior.
- $L(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})$: the likelihood.
- $\pi(\boldsymbol{\theta})$: the priors.
- $\epsilon(\boldsymbol{x}|\mathcal{M})$: the evidence ('marginal likelihood').

→ Even for Gaussian data, the posterior can be difficult to obtain. Need sampling techniques.

Hacking ourselves into the posterior



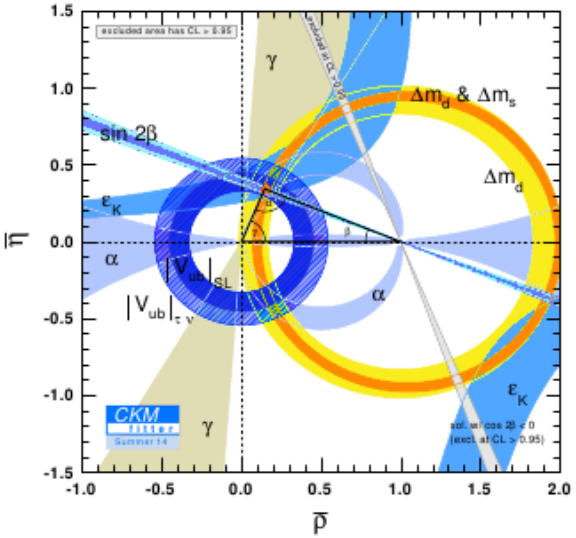
Family Tree of famous distributions. (11th Feb 2019.)

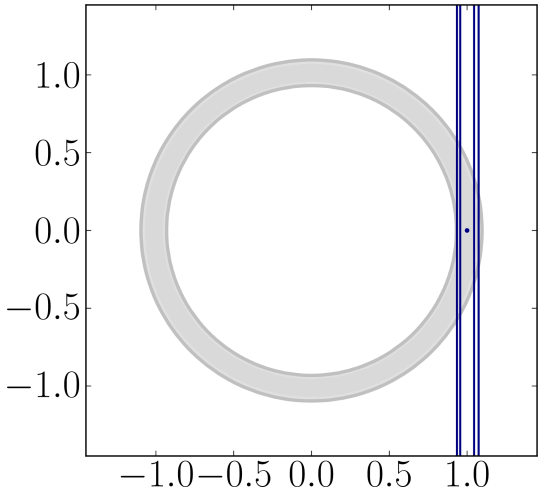
↓

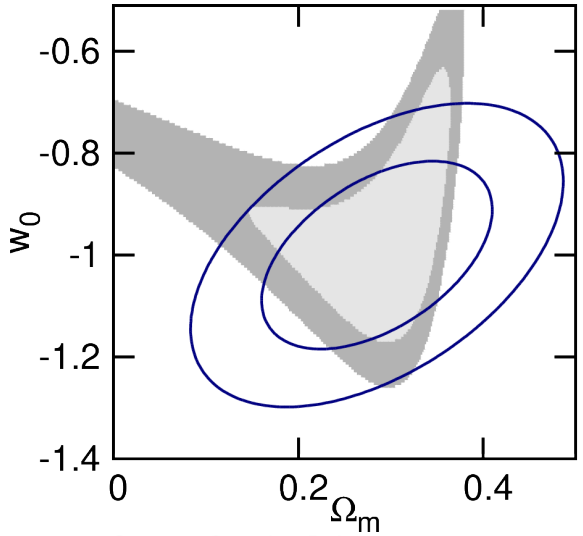
But what do we do? Our sought distribution $\mathcal{P}(\theta|\boldsymbol{x})$ is usually far from famous?

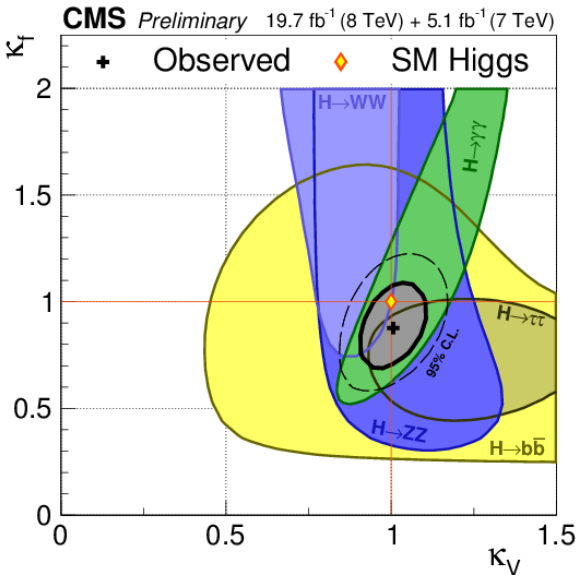
```
from random import gauss
#include < gsl/gsl_rng.h >
```

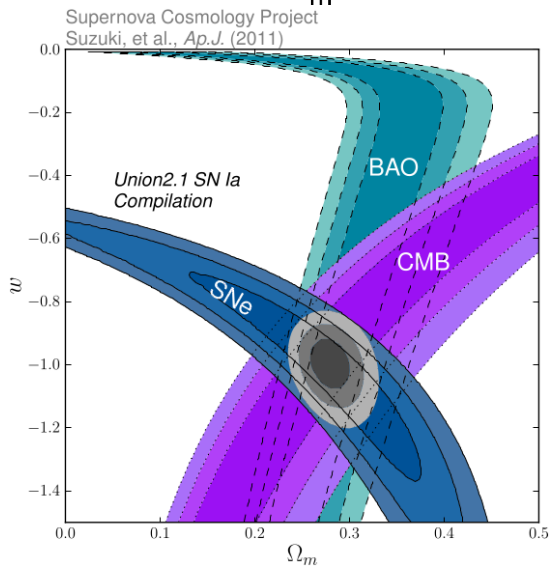
Examples


Δm_d & Δm_s
Δm_d
sin 2β
ε_K
α
|V_ub|_SL
|V_ub|_τν
γ
CKM fitter
Summer 14
ρ̄
η̄

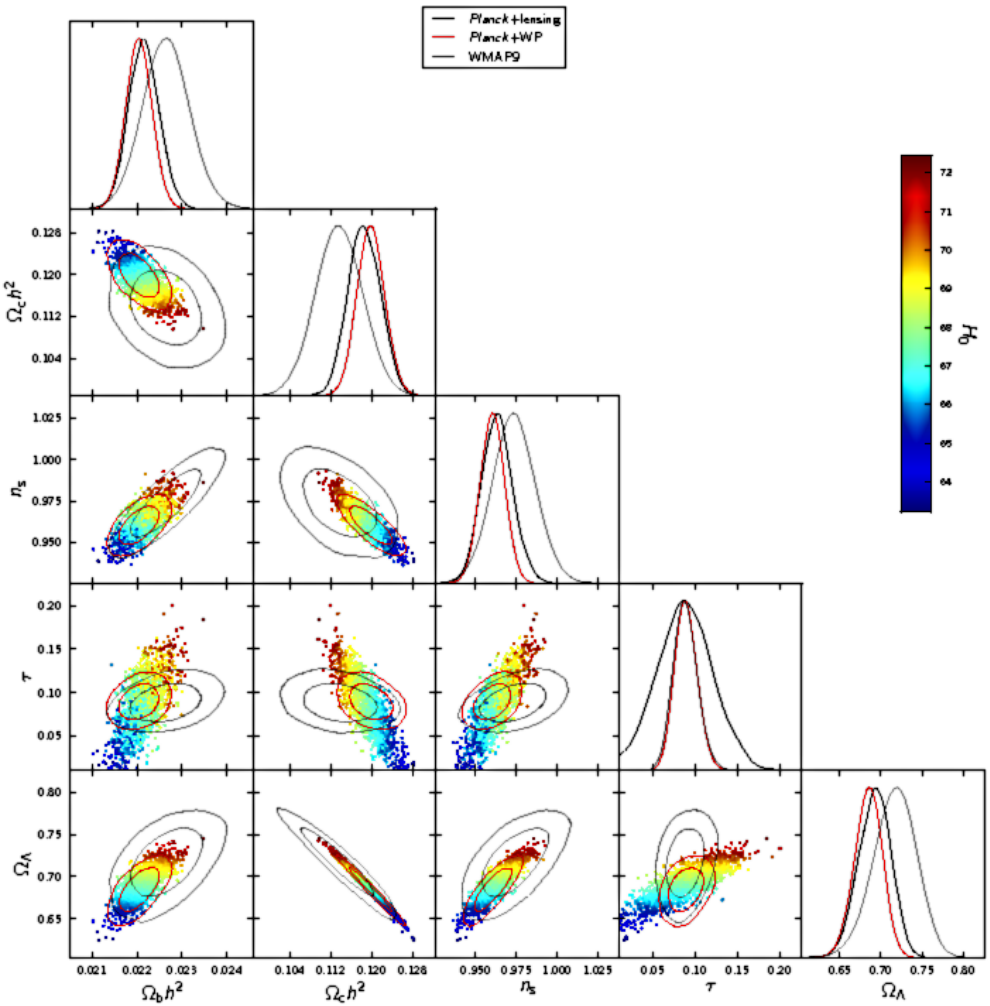



w0
Ωm


CMS Preliminary 19.7 fb-1 (8 TeV) + 5.1 fb-1 (7 TeV)
Observed
SM Higgs
H→WW
H→γγ
H→ττ
H→ZZ
H→bb̄
95% CL
κf
κV

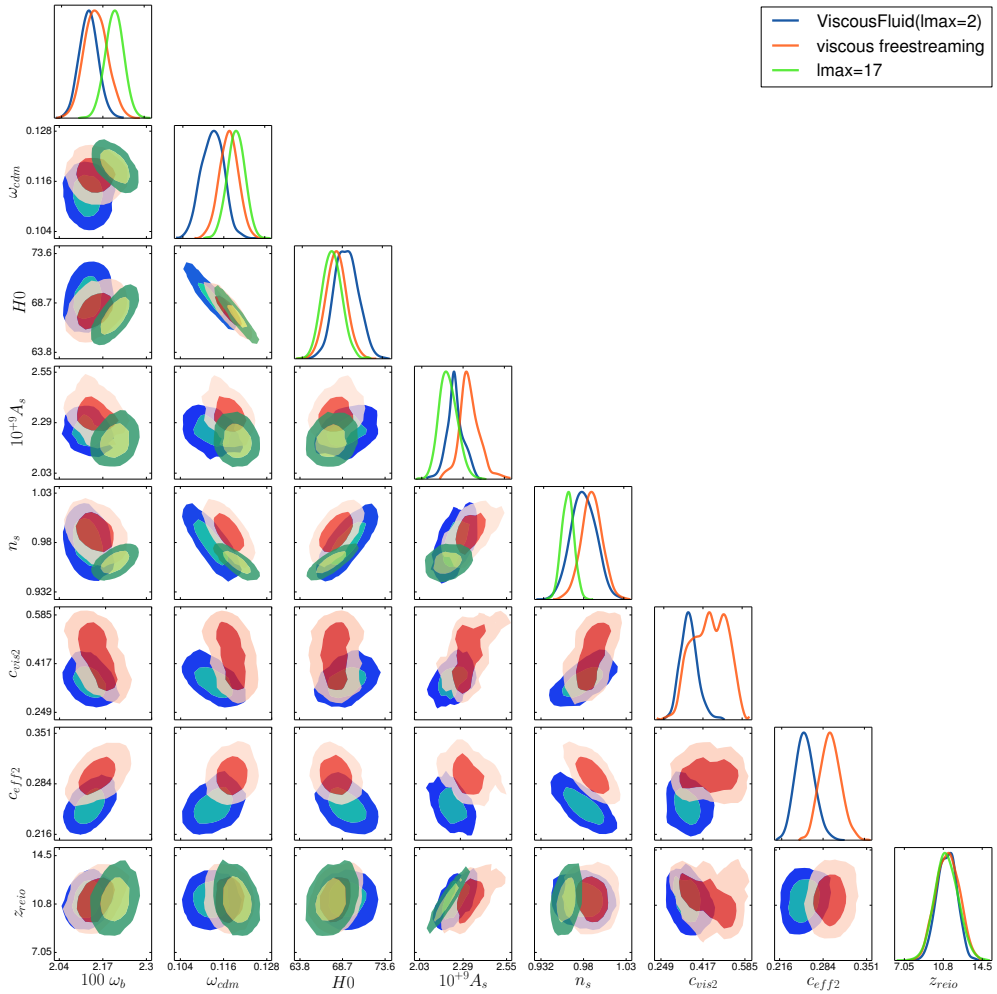

Supernova Cosmology Project
Suzuki, et al., Ap.J. (2011)
Union2.1 SN Ia Compilation
BAO
CMB
SNe
w
Ωm

Triangle plots



Planck Collab.

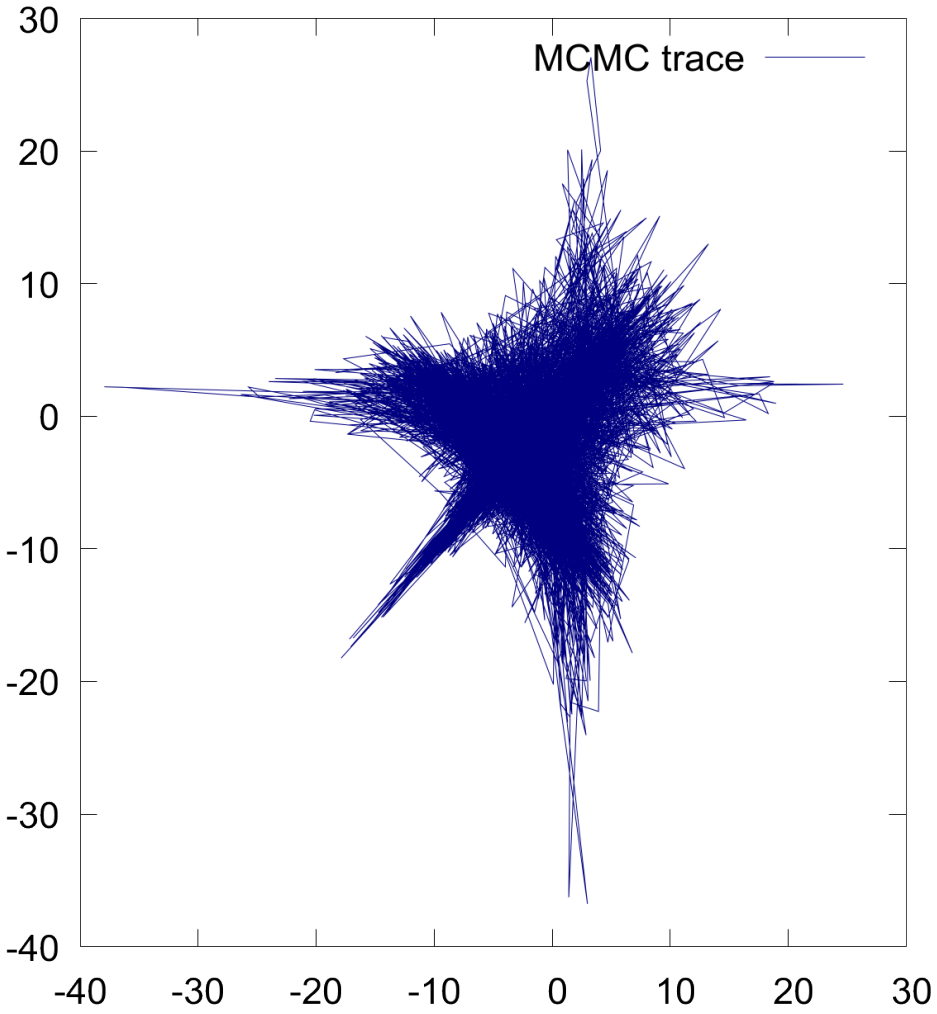
Triangle plots



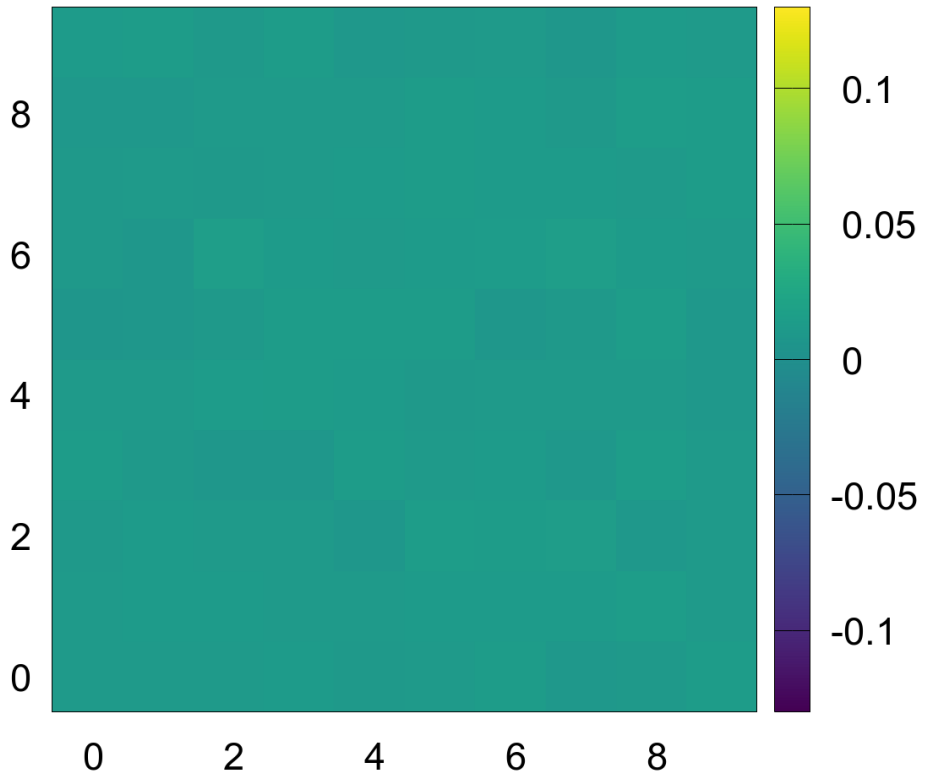
Sellentin & Durrer (2015)

Sampling for generating realizations

- Sampling to generate random fields
- Each ‘pixel’ is a random variable
- Sidelength N $\Rightarrow$ N^2 pixels
- Dimension = N^2



Generating random fields



The Metropolis-Hastings Algorithm

Detailed Balance

Equilibrium between the occupation of two states $\mathcal{P}_{i,j}$ is reached if

$$r_{i\to j}\mathcal{P}_i = r_{j\to i}\mathcal{P}_j, \tag{2}$$

and the transition probability from $\mathcal{P}_i$ to state $\mathcal{P}_j$ has rate $r_{i\to j}$.
$\Rightarrow$ Same principle as e.g. in photon emisson/absorption from electronic shells in atoms [search for 'Einstein coefficients']

$\rightarrow$ Let's turn the transition between two states into a running chain.

Metropolis Hastings Algorithm

1. Provide a guess for a Gaussian approximation $\mathcal{G}_P(\boldsymbol{\theta})$ to the posterior (Fisher matrix/previous sample covariance matrix)
2. For $i = 0$ to N_{MCMC}
   if $i = 0$ evaluate the posterior P at point $\boldsymbol{\theta}_0$ in parameter space. Else use the current $\boldsymbol{\theta}_i$ of the chain.
3. Draw a random step $\Delta\boldsymbol{\theta}_i \sim \mathcal{G}_P(\boldsymbol{\theta})$ and calculate $R = \frac{P(\boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i)}$.
4. If $R > 1$, then the posterior probability at the new point $\boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i$ is larger than the old probability; the new point is then accepted as $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i$.
5. If $R < 1$, then draw $\alpha \sim \mathrm{Uniform}[0, 1]$.
   If $\alpha > R$, then $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$, i.e. the point $\boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i$ is rejected because it has too low a probability.
   If, however, $\alpha < R$, then $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i$, i.e. the trial point is accepted because statistical equilibrium demands a population of the low-$\mathcal{L}$ states as well.
6. Store all points $\boldsymbol{\theta}_i$ in a chain.

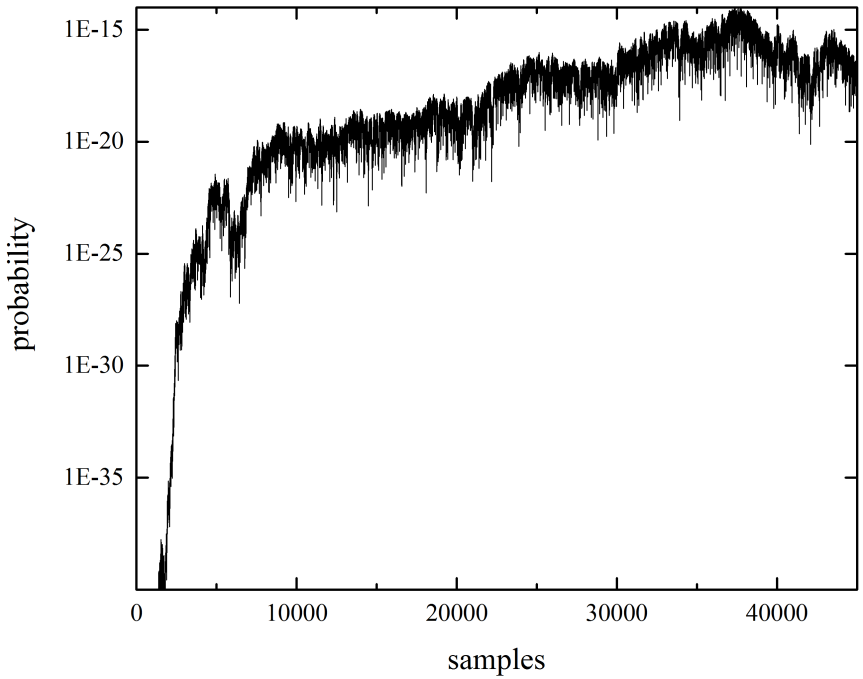
Convergence

- If the chain ran for long enough ('has converged'), then: $n(\boldsymbol{\theta}) \propto \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$ (density of samples proportional to posterior density).
- Detailed balance: reacts to $\mathcal{P}(\boldsymbol{\theta}_i)/\mathcal{P}(\boldsymbol{\theta}_j) \Rightarrow$ any normalization drops out.
- $\Rightarrow$ the normalization constant $\mathcal{N}$ of the distribution is unknown (but not needed for parameter inference).

Burn-in period

Reaching equilibrium needs time.

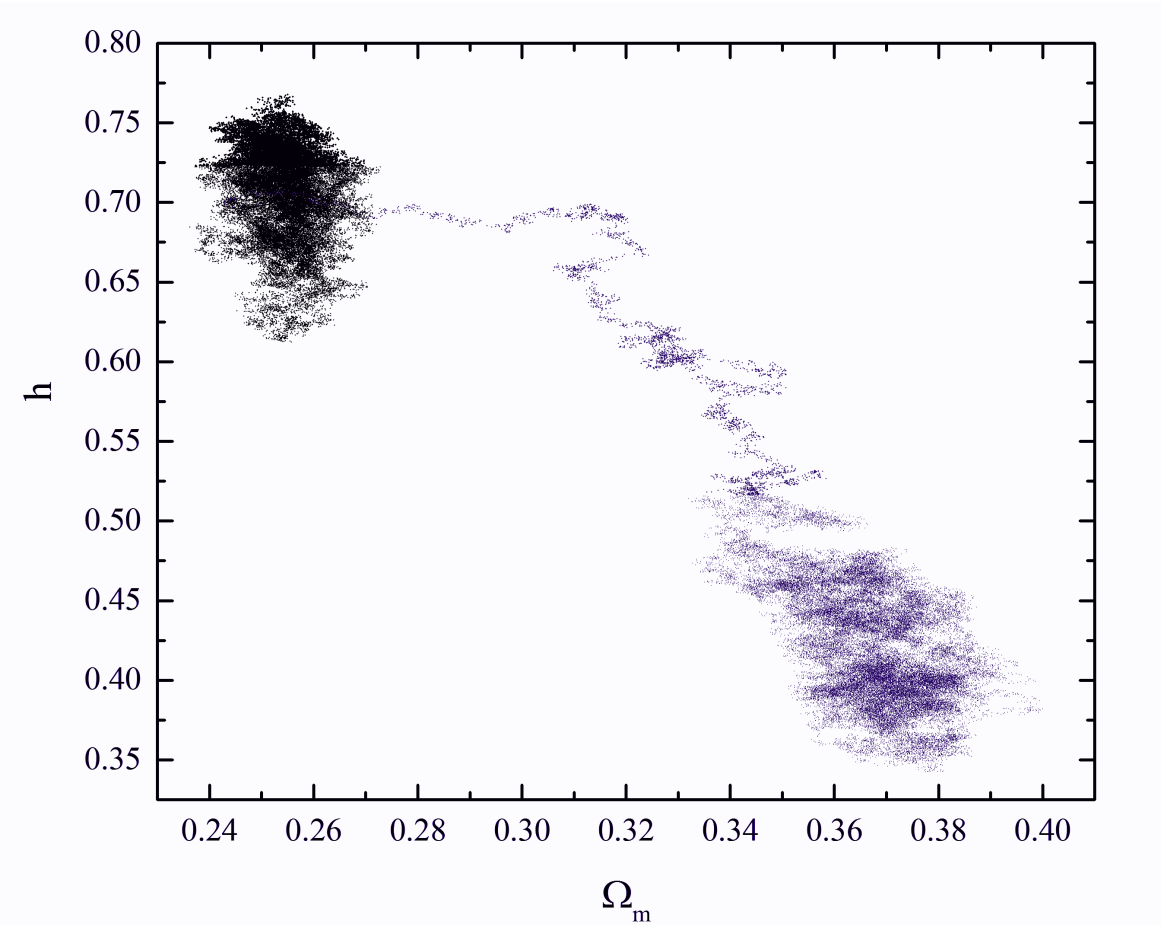
- Thermodynamics: put cold object into warmer environment
- Radiation physics: put a phosphorescent object into the dark
- MCMC: burn-in period (searching for the peak; log-likelihood increases)



Burn-in examples

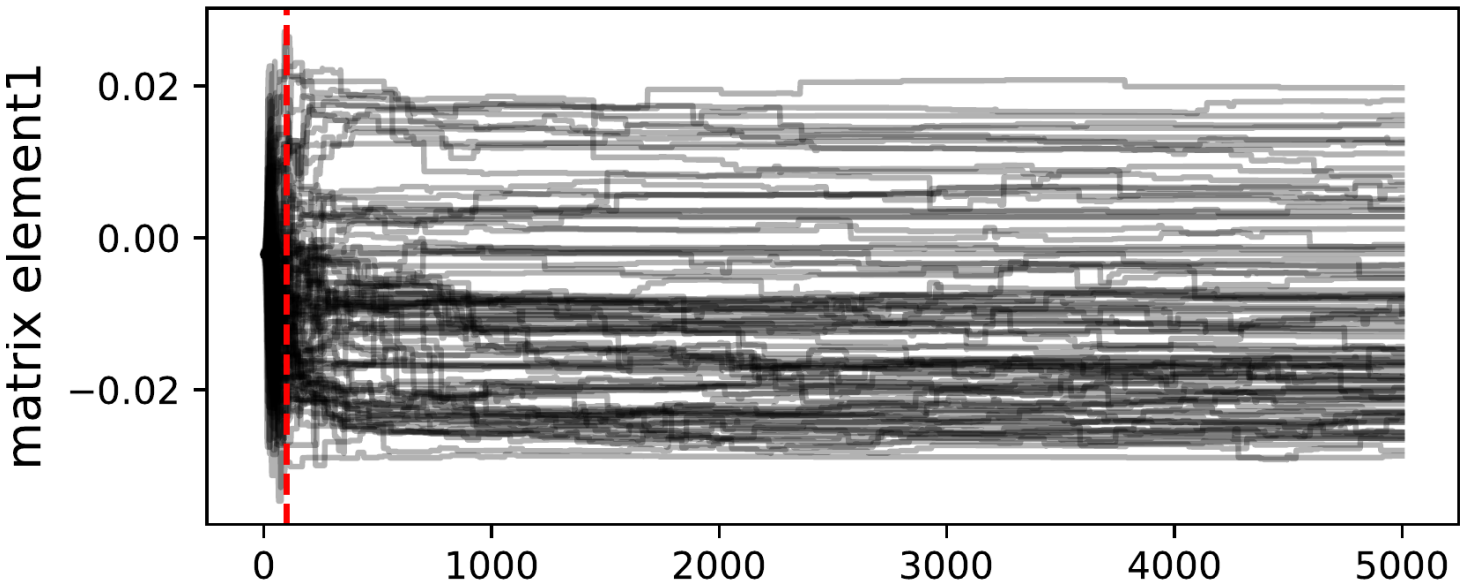
Remove the unequilibrated burn-in period!

- Rat tails in likelihood plots; log-likelihood increases.



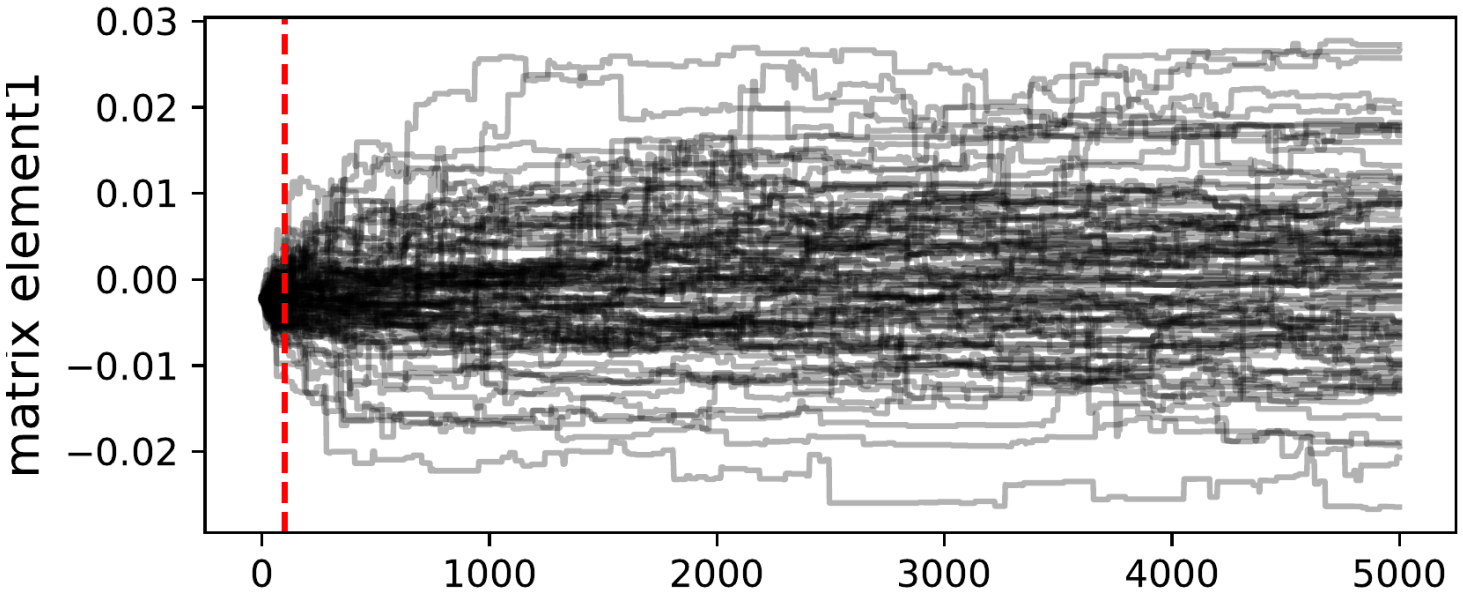
Getting stuck in high dimensions

18 dimensional parameter space (neutrino matrix)



Getting stuck in high dimensions

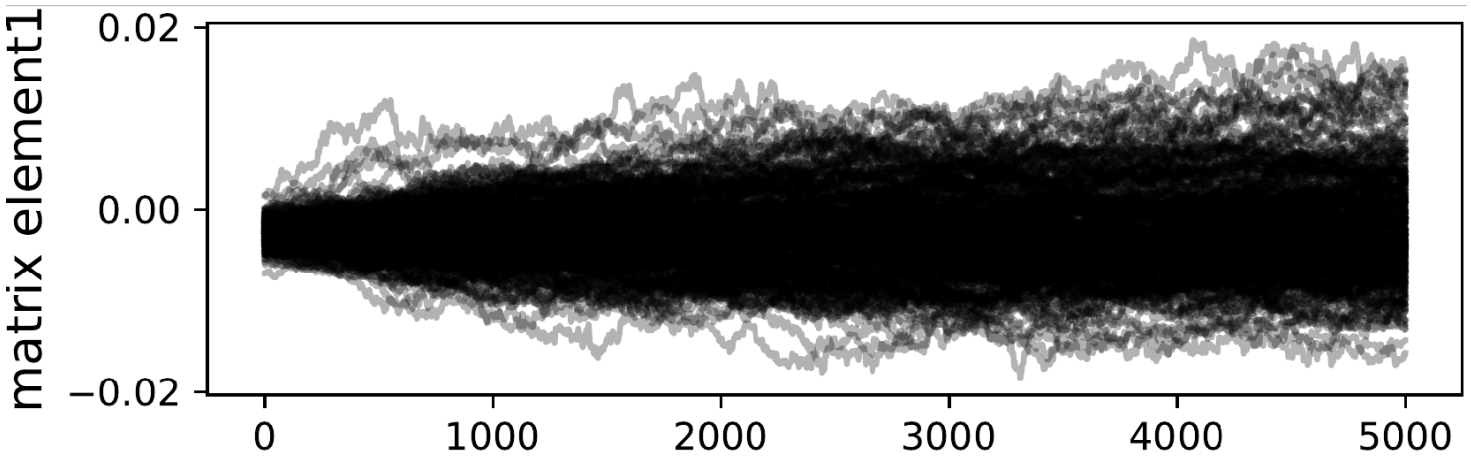
18 dimensional parameter space (neutrino matrix)



(Somewhat better than last slide.)

Getting stuck in high dimensions

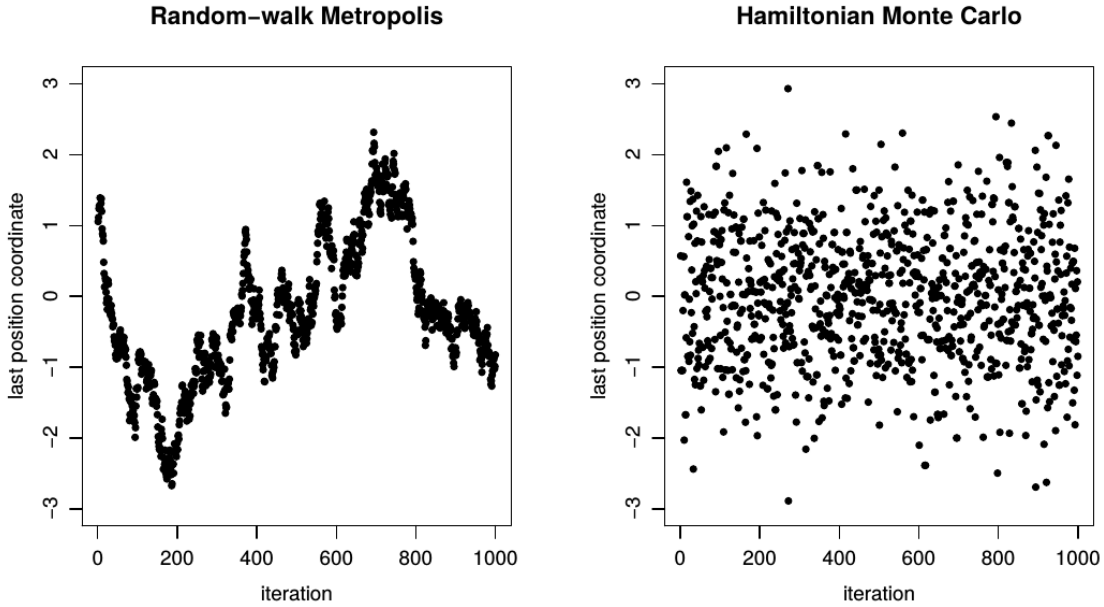
18 dimensional parameter space (neutrino matrix)



(Good mixing.)

Monitor MCMC convergence

- Trade-off: acceptance rate $\leftrightarrow$ correlation between samples.
- Measure the correlation length, and thin correlated chains.

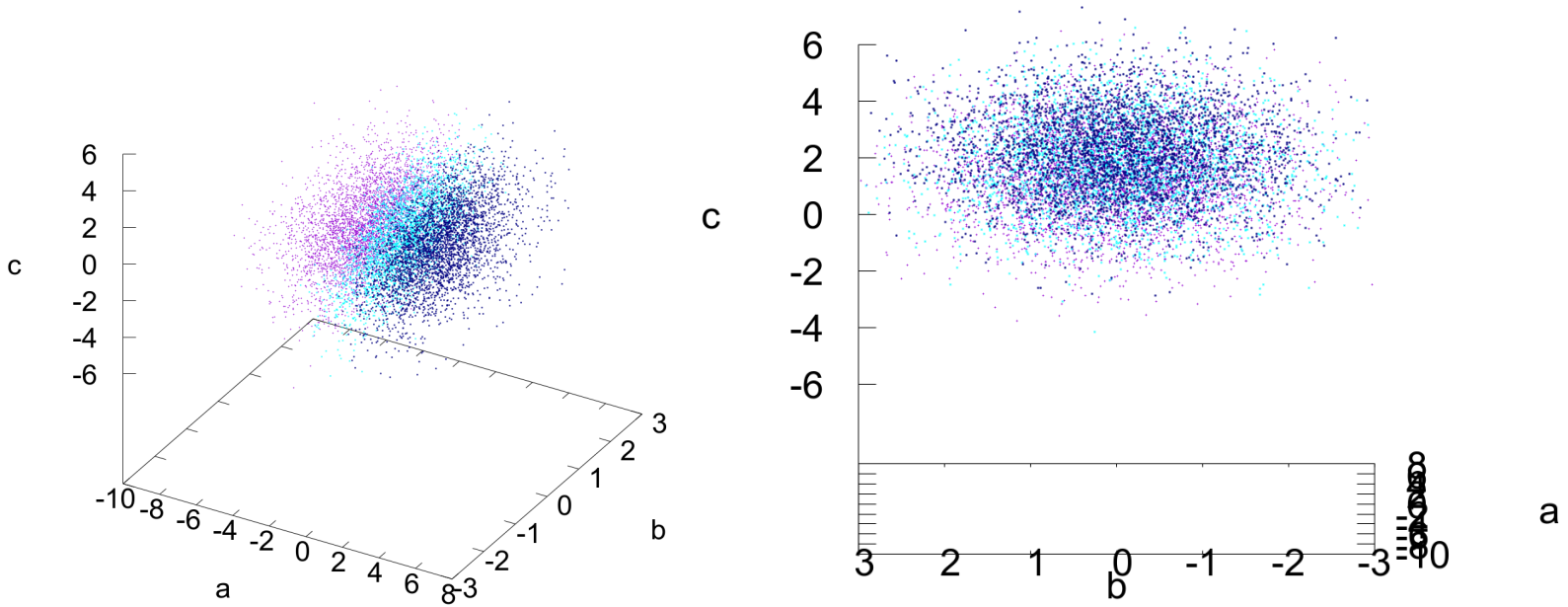


Neil 2012, arXiv: 1206.1901

→ Left: highly correlated (bad); right: uncorrelated (good).

Conditionals and Marginals

$$\mathcal{T} \propto [(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})]^{-\nu}$$



Left: Conditionals for different a, right: Marginal over a.

Convergence Diagnostics

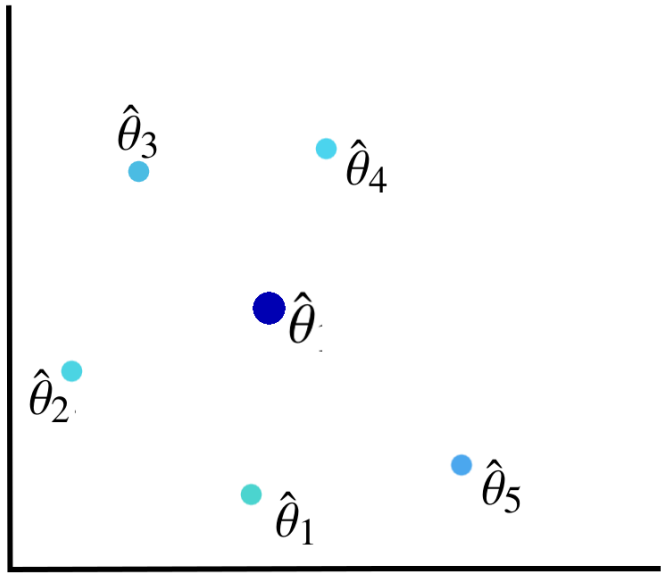
1. Plot the samples without gridding.
2. Find the best-fitting values of all your chains and compare.
3. Is half your chain still representative of the posterior?
4. Does binning still influence the credibility intervals?
5. Is the chain squeezed against any prior-boundaries?
6. Compute the auto-correlation length of your chain for all parameters & potentially thin until sequential points are uncorrelated ('potentially' depends on posterior-shape)

$$C(T) \approx \frac{1}{M-T} \sum_{m=1}^{M-T} [f(T+m) - \bar{f}][f(m) - \bar{f}]$$

Convergence Diagnostics

Gelman-Rubin Test: Intra-Chain variance vs. Inter-chain variance.

- Run M different chains with different starting points, let $m \in [1, M]$.
- mth chain: $\theta_1^m, \theta_2^m, \theta_3^m, ...\theta_{N_m}^m$.
- Discard the burnins.
- Calculate for each parameter θ, the posterior mean $\hat{\theta}_m = \frac{1}{N_m} \sum_i^{N_m} \theta_i^m$,
- ...and the intra-chain variance $\sigma_m^2 = \frac{1}{N_m - 1} \sum_i^{N_m} (\theta_i^m - \hat{\theta}_m)^2$.
- Calculate $\hat{\theta}$, the mean of all chains $\hat{\theta} = \frac{1}{M} \sum_m^M \hat{\theta}_m$.

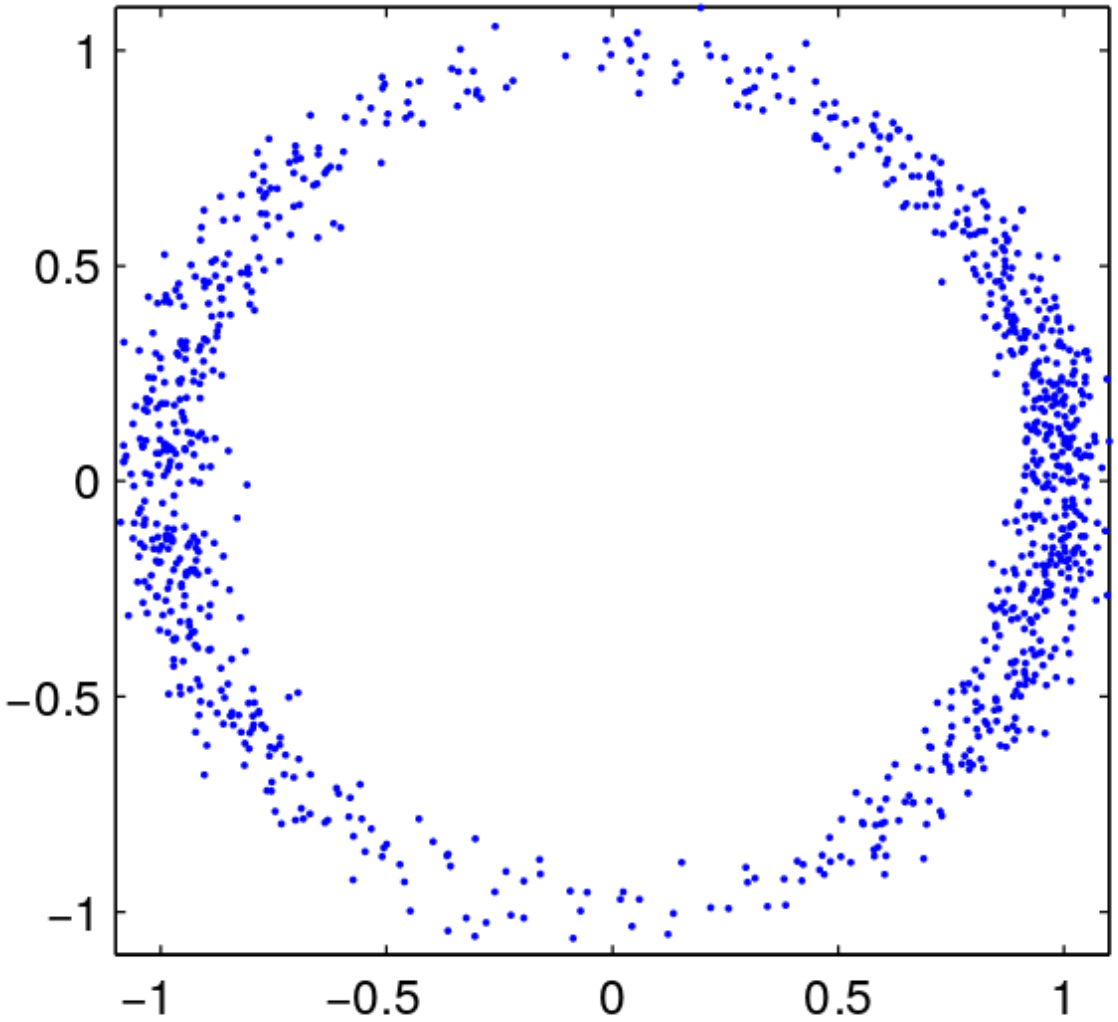


Gelman-Rubin cntd.

- Compute how the individual means vary around the joint mean
  $B = \frac{N}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta})^2$
- Compute the averaged variances of the chains
  $W = \frac{1}{M} \sum_{m=1}^{M} \sigma_m^2$
- Define $\hat{V} = \frac{N-1}{N} W + \frac{M+1}{MN} B$; under convergence, this is an unbiased estimator of the true variance. But if the chains have converged, then W is *also* an unbiased estimate of the true variance. Hence...
- ...test whether $R = \sqrt{\hat{V}/W} \approx 1$.
  If it is not, convergence has not been reached.
- Various refinements exist, see Gelman & Rubin (1992), Brooks & Gelman (1997).

Hamilton Monte Carlo
A clever way of distributing MH-samples

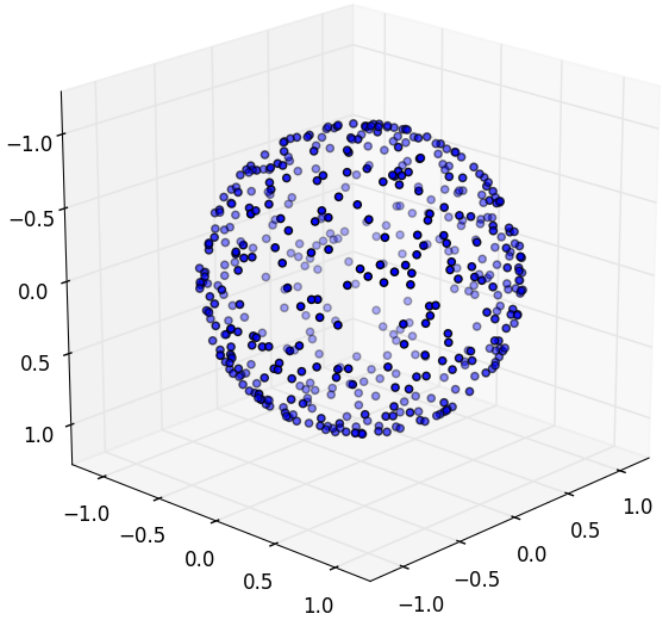
Difficult posterior shapes



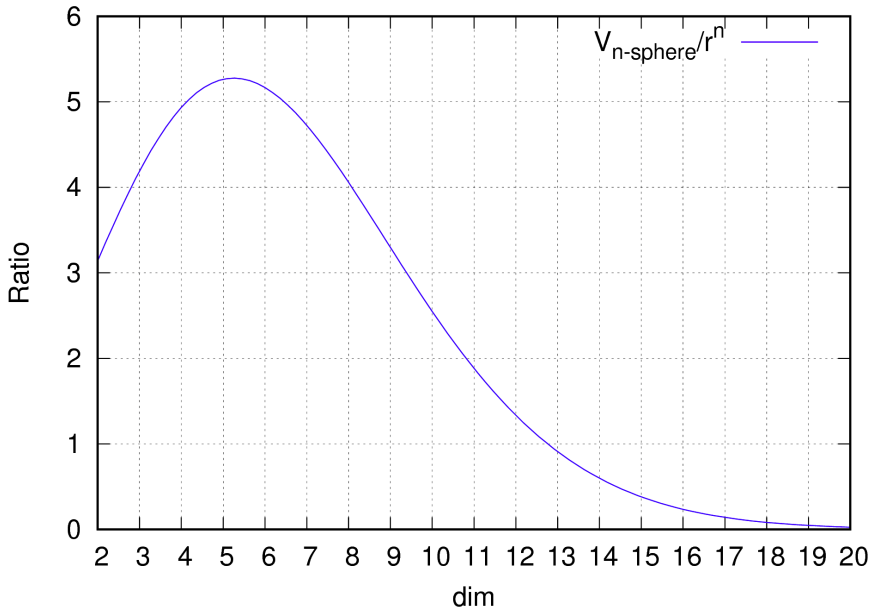
Hajian 2006

The Curse of Dimensionality

With increasing d, have ever more possibilities to go 'wrong' with a d-dimensional random step.



Credit (left): Wikipedia



We need a satnav for our high-dimensional space.

HMC

Hamiltonian: (governs the evolution of trajectories in phase-space)

$$H(\boldsymbol{\theta}, \boldsymbol{u}) = U(\boldsymbol{\theta}) + K(\boldsymbol{u}). \tag{3}$$

Introducing a potential $U(\boldsymbol{\theta})$

$$U(\boldsymbol{\theta}) = -\log P(\boldsymbol{\theta}). \tag{4}$$

Introduce kinetic energy

$$K(\boldsymbol{u}) = \boldsymbol{u}^T \boldsymbol{u}/2, \quad \boldsymbol{u} \sim \mathcal{G}(0, \mathcal{I}), \tag{5}$$

Know Hamiltonian equations of motion:

$$\dot{\boldsymbol{\theta}} = \boldsymbol{u}, \quad u_i = -\frac{\partial H}{\partial \theta_i}. \tag{6}$$

HMC

Solving Hamiltonian equations: deterministic.
Need a source of randomness $\Rightarrow$ randomize initial velocities.
Initial $\boldsymbol{u} \sim \mathcal{G}$.

$$\exp\left(-H(\boldsymbol{\theta}, \boldsymbol{u})\right) = P(\boldsymbol{\theta})\mathcal{G}(0, \mathcal{I}) \tag{7}$$

So if the auxiliary velocities $\boldsymbol{u}$ are marginalized over, which is equivalent to not protocolling them in the chain, then sampling $\exp(-H)$ samples $P(\boldsymbol{\theta})$.

Hamilton Monte Carlo Algorithm

1. For $i = 0$ to N_{MCMC}
   if $i = 0$, choose a starting point $\boldsymbol{\theta}_0$,
   else use the current $\boldsymbol{\theta}_i$ of the chain.
2. Draw a random velocity $\boldsymbol{u}_i \sim \mathcal{G}(0, \mathcal{I})$.
   Leapfrog loop
   1. Use $\boldsymbol{\theta}_i$ and $\boldsymbol{u}_i$ as initial conditions for the Hamiltonian equations of motions.
   2. For $j = 0$ to N_L
      make leapfrog steps that update $(\boldsymbol{\theta}_j, \boldsymbol{u}_j) \rightarrow (\boldsymbol{\theta}_{j+1}, \boldsymbol{u}_{j+1})$
3. Having arrived at $(\boldsymbol{\theta}_{N_L}, \boldsymbol{u}_{N_L})$, calculate
   $R = \exp[-H(\boldsymbol{\theta}_i, \boldsymbol{u}_i) + H(\boldsymbol{\theta}_{N_L}, \boldsymbol{u}_{N_L})]$.
4. If $R > 1$, the new point is accepted, $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$.
5. If $R < 1$, draw $\alpha \sim \text{Uniform}[0, 1]$.
   If $\alpha > R$, then $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$, i.e. the trial point $\boldsymbol{\theta}_{N_L}$ is rejected.
   If $\alpha < R$, then $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_{N_L}$, i.e. the trial point is accepted.

HMC vs. MH

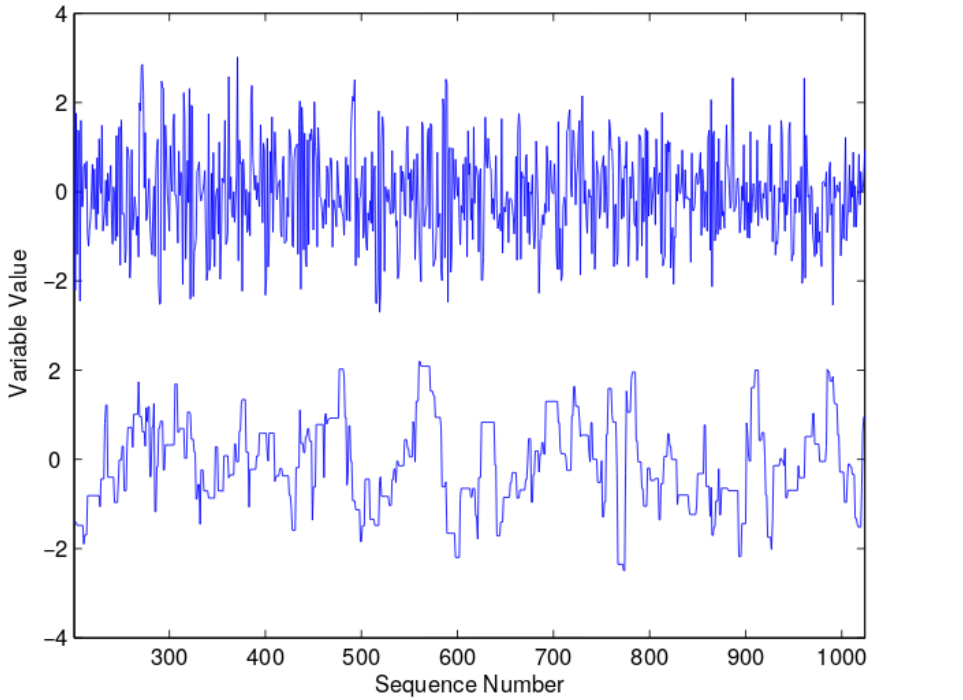


FIG. 1: Samples drawn from an isotropic six-dimensional Gaussian distribution using the HMC (top) and the Metropolis algorithm with optimal step-size (bottom).

Hajian 2006