Numerical Methods for Model Based Probabilistic Inference

J. Jewell, JPL

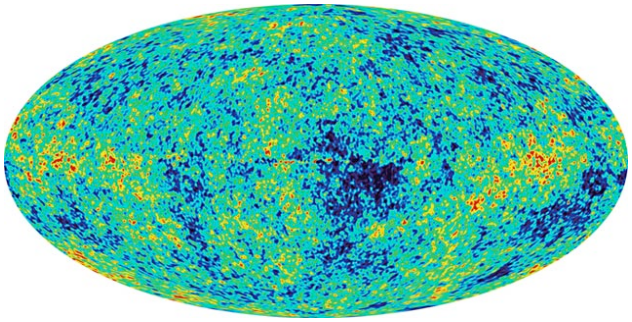
The Central Role of Probabilistic Inference in Modern Science and Engineering

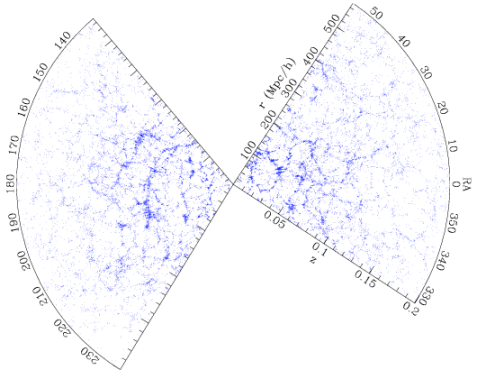
- There is an increasingly central role played by probability, and associated numerical methods for probabilistic reasoning, in many diverse problems of science in engineering.
- Several reasons for this - some aspects of the “world of interest” are conveniently described by probabilistic models (i.e. financial markets as random process, images, time series, etc.), deterministic systems might have random or unknown initial conditions, parameters of the dynamics, etc.
- General Problems we often face- 1) compare theory and experiment (in the presence of measurement and/or computational error), 2) uncertainty quantification
- Goal for this course - awareness of random or uncertain elements of models or problems, idea of how to reformulate as probabilistic inference, and numerically solve

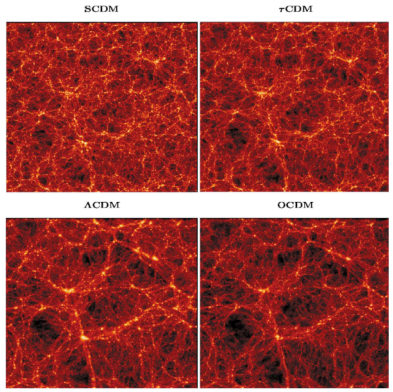
Computationally Enabled Probabilistic Inference

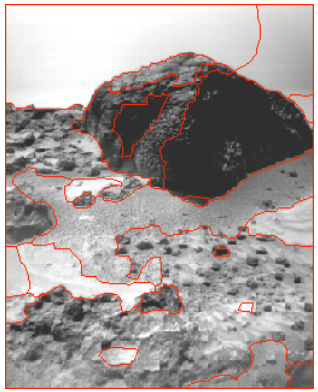
- Probability needed when we are not after single solutions, but after the ability to quantify states of knowledge - I.e. the representation, quantification, and control of uncertainty.
- For almost all problems with a probabilistic component (either due to the model, quantifying uncertainty, or both) we almost always wind up looking at probabilities which are non-Gaussian, and for which we do not have ``direct'' (i.e. in one step) sampling methods.
- Much more computationally intensive (as opposed to solving for single solutions or estimates), and only enabled through advances in computing (memory and speed)
- Why is this an interesting subject to discuss in the context of ``computationally enabled science''?
- It is directly due to advances in computing hardware (memory and speed) that numerical methods to solve these problems are within reach!

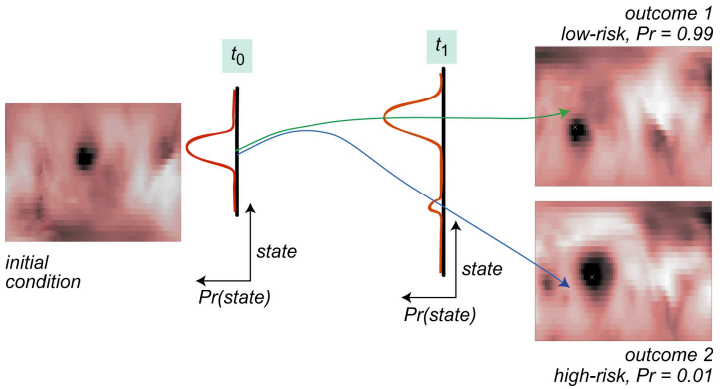
Example Inference Problems



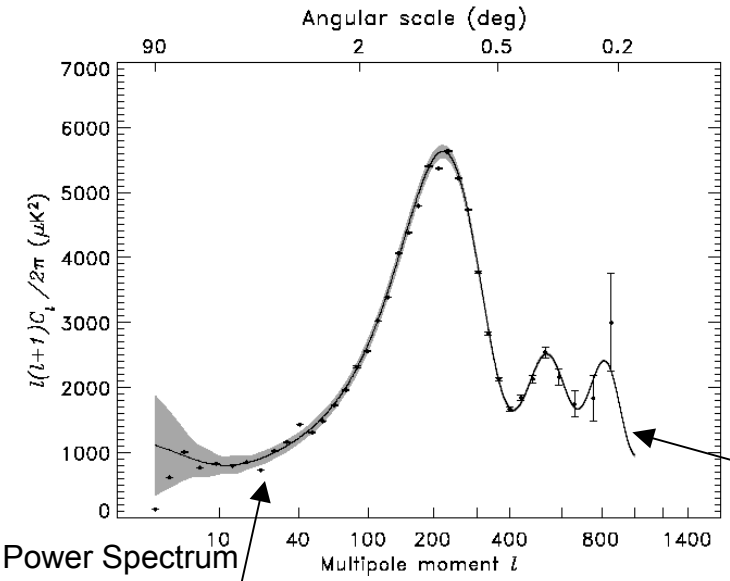






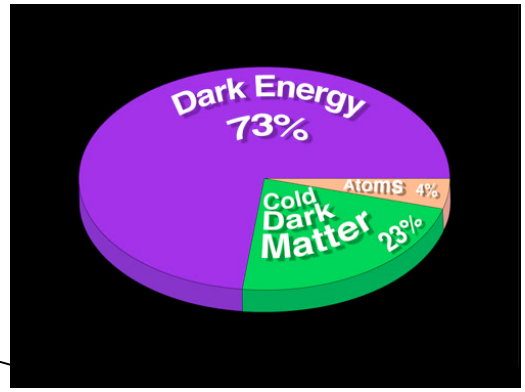


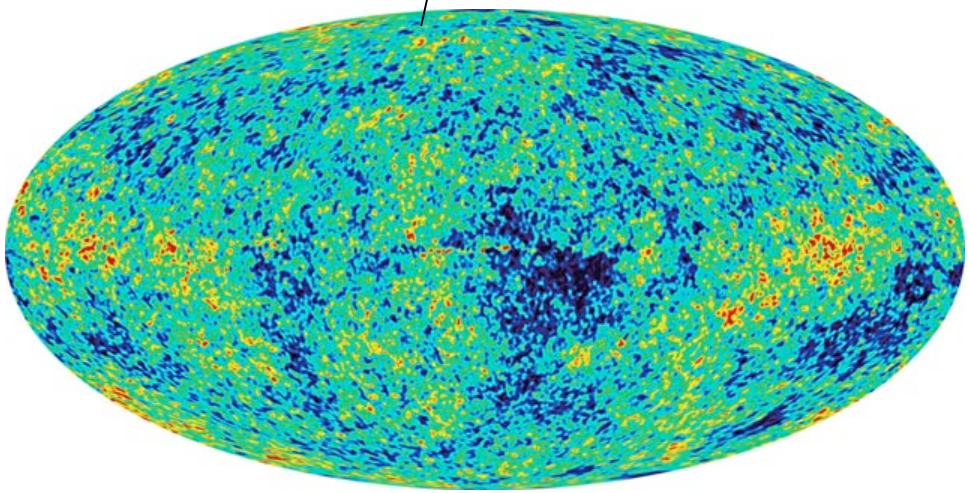
Comparison of Theory and Experiment in Cosmology I.

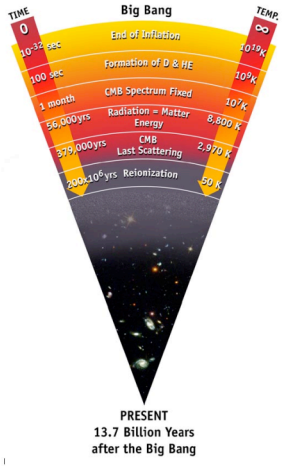


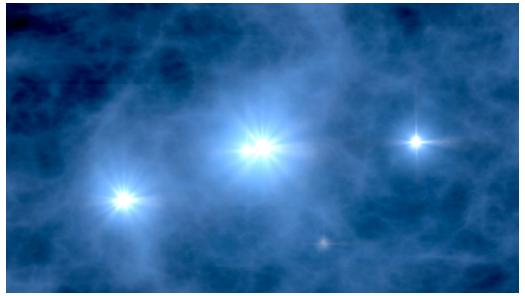
Measured Power Spectrum

Predicted Power Spectrum







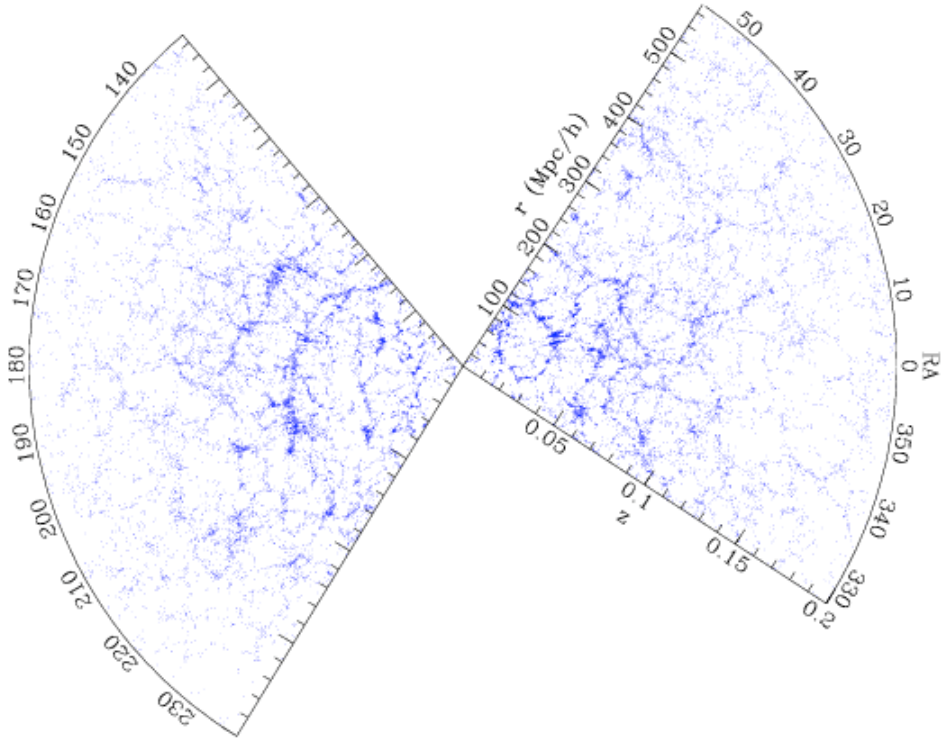


WMAP Observations of the Cosmic Microwave Background, 1/2 degree resolution

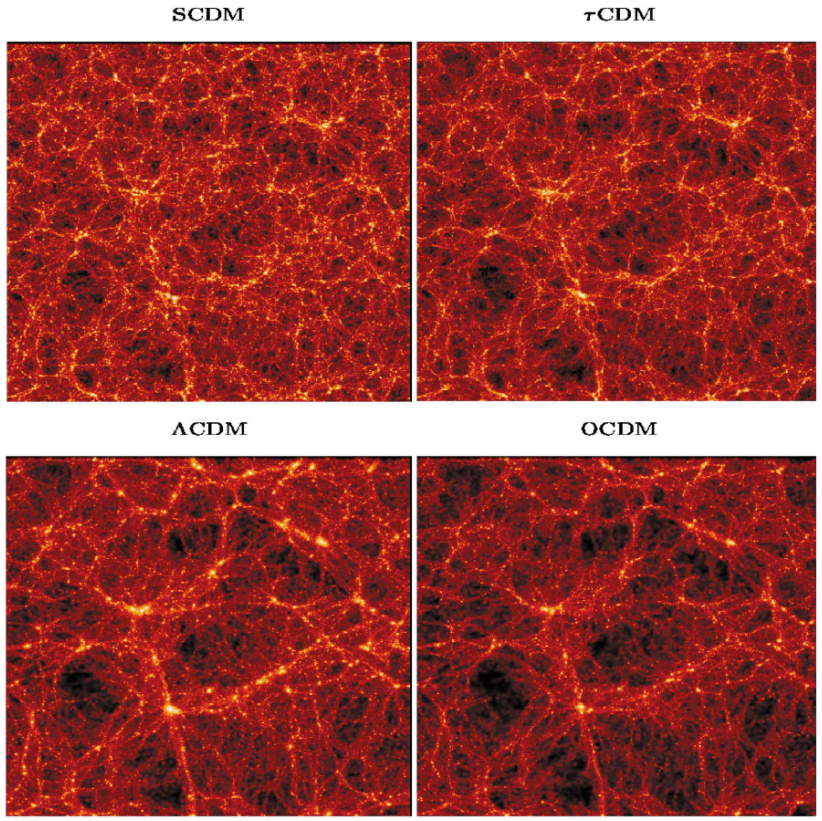
Cosmological models including details such as composition, age, star formation history, etc.

http://map.gsfc.nasa.gov

Comparison of Theory and Experiment in Cosmology II.

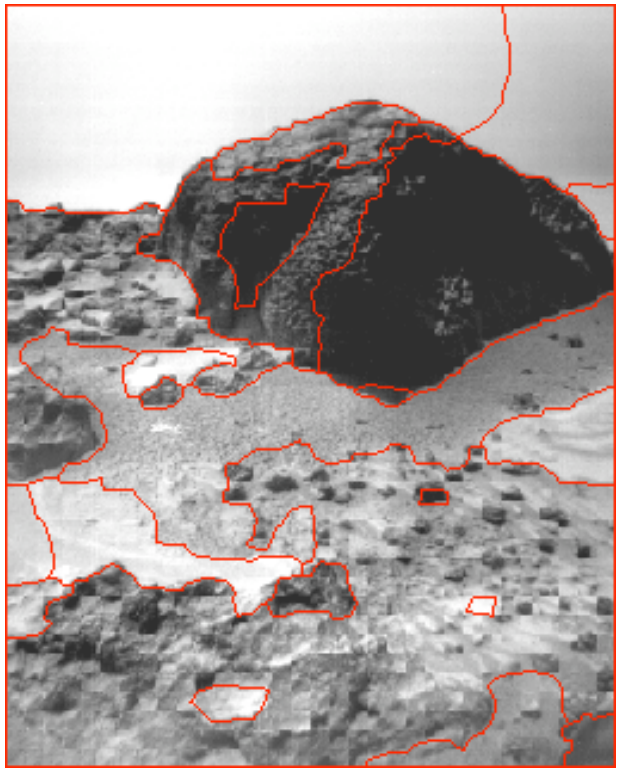


Redshift surveys map out the 3D distribution of matter, with distant "slices" seen as they were in the *PAST*…



1 "frame" from a movie of evolution of 4 different types of Universe (Virgo Consortium)

Image Segmentation - Inference in Non-stationary Spatial Random Fields

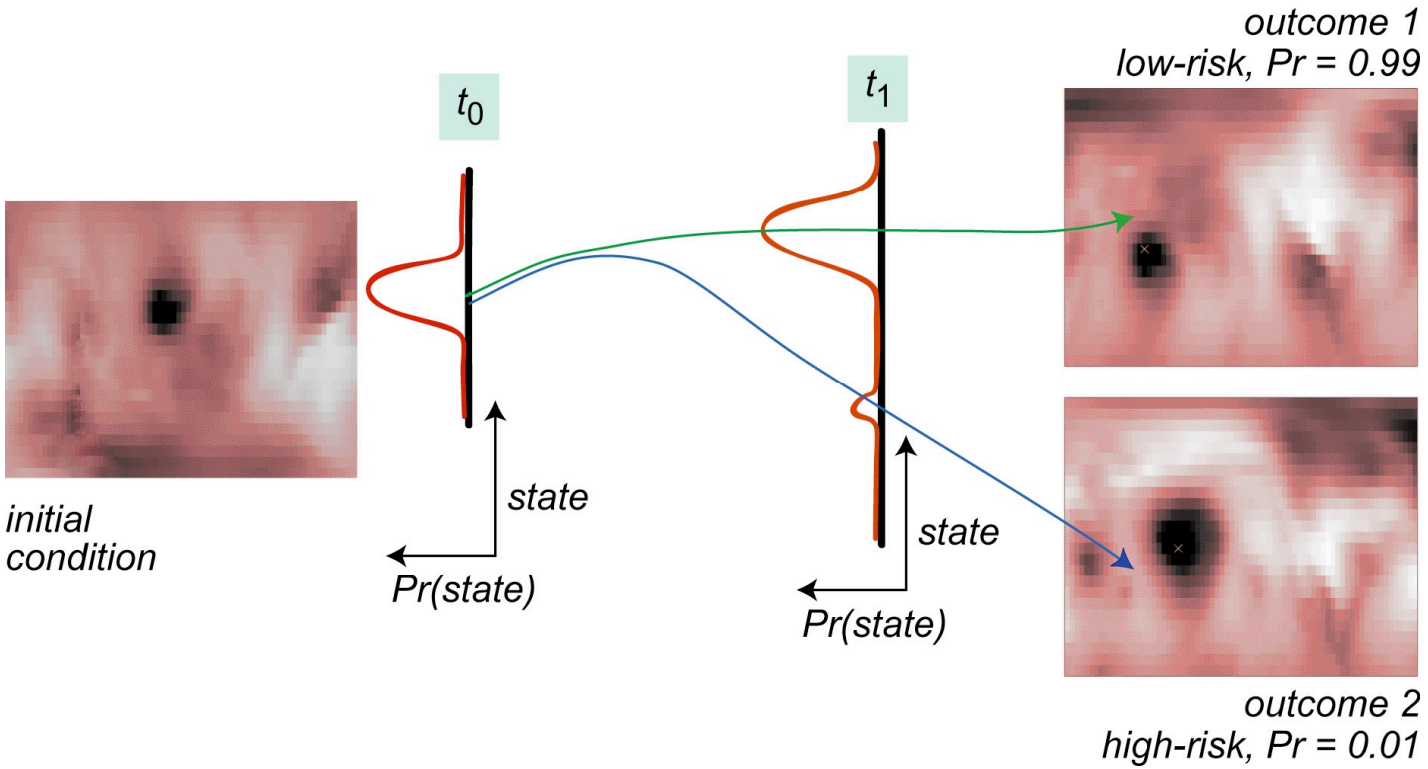


Mars Rover image -

- Modeled as a spatial random field with statistical properties that vary from region to region.
- Segmentation problem is an "inverse problem" - cluster the pixels together in regions with similar statistical properties (I.e. regions of uniform texture)

Courtesy, S.C. Zhu, UCLA

Smoothing, Filtering, and Prediction for Nonlinear Systems given Noisy Measurements



Comparing Theory and Observation

- For almost all problems of interest encountered, there is a complicated joint relationship among various degrees of freedom of a model and the measured quantities
- It is not purely the volume of data that makes comparison between theory difficult, but also the relation of the underlying theory to the observations (the data model can be complex, and/or have a complicated relation to the underlying parameters of interest to be inferred)
- Because of tremendous advances in computation, these problems can now be solved using the same basic strategy: “write down the probability of everything”, condition on data, and marginalize (numerically!) over everything else.
- In the past, this program was often computationally out of reach, but not anymore!

Brief Review of Probability…

We have some K dimensional state space, and a joint probability (measure if continuous variables):

$$p(x_1,\ldots,x_k)$$

We have the (1D) conditional densities (by definition):

$$p(x_i \mid x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k) = \frac{p(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_k)}{\int dy\ p(x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_k)}$$

Similarly for the 2D, etc. conditional densities.

Joint density factors into conditional and marginal…

Gibbs Random Fields

Hammersley-Clifford Theorem:

$$\log\left(\frac{p(x_1,\ldots,x_k)}{p(0_1,\ldots,0_k)}\right)=\sum_{2^k}V_{\{i,\ldots,j\}}(x_i,\ldots,x_j)$$

Where the "potentials" non-zero IFF a mutual conditional probability dependence on the variables in the set

Proven using the useful identity: $\frac{p(x_1,x_2)}{p(y_1,y_2)}=\frac{p(x_1\mid x_2)}{p(y_1\mid x_2)}\frac{p(x_2\mid y_1)}{p(y_2\mid y_1)}$

Conditional probability structure important algorithmically (I.e. sites which are conditionally independent can be adjusted In parallel!)

Simulation and Inference

- Different conditional densities inherited from a common joint density
- Simulation conditions on the model, Inference conditions on the data!

Joint density of "everything":

$$p(d,x,\theta) = p(d \mid x)\,p(x \mid \theta)\,p(\theta)$$

Data — Underlying "truth" — Model parameters

Simulation: Sample data given the model

$$p(d,x \mid \theta) = p(d \mid x)\,p(x \mid \theta)$$

Inference: Sample model parameters given data

$$p(\theta,x \mid d) = \frac{p(d \mid x)\,p(x \mid \theta)\,p(\theta)}{\int d(\theta',x')\; p(d \mid x')\,p(x' \mid \theta')\,p(\theta')}$$

Example: Bayesian Inference of the Hubble Constant

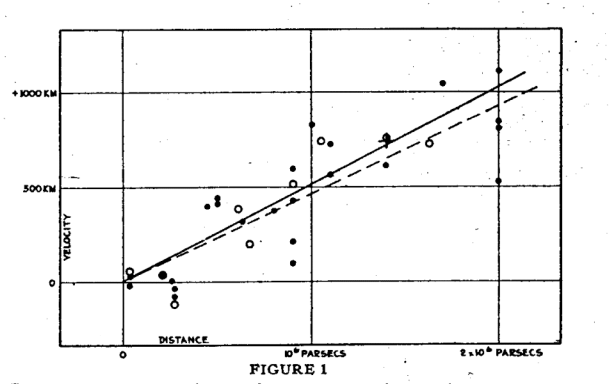


FIGURE 1

$$v_i = H_0 r_i$$

Joint density: Can be used to simulate observations

$$p(v, r, V, R, H_0) = p(v \mid V)\, p(r \mid R)\, p(V \mid R, H_0)\, p(R)\, p(H_0)$$

Posterior: condition on data, and integrate over everything else

$$p(H_0 \mid v, r) \propto p(H_0) \int d(V, R)\; p(v \mid V)\, p(r \mid R)\, \delta(V - H_0 R)\, p(R)$$

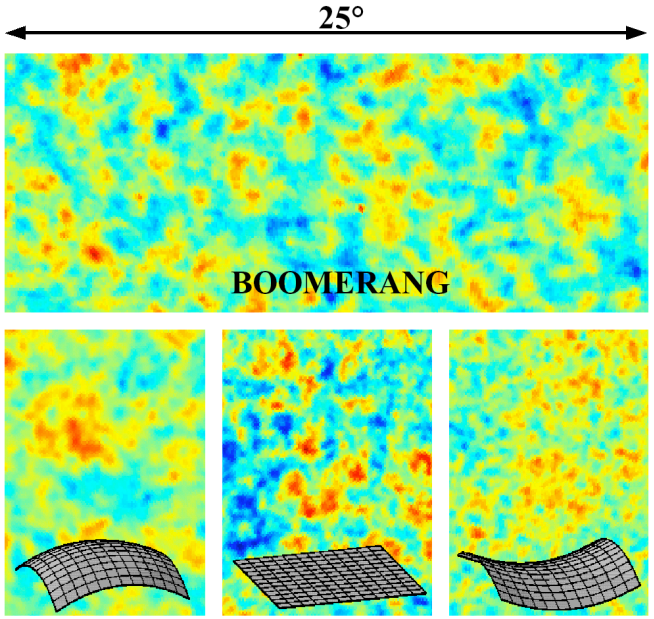
$$= p(H_0) \int d(R)\; e^{-(v - H_0 R)^2 / 2\beta^2}\, e^{-(r - R)^2 / 2\sigma^2}$$

Assume Uniform Prior…

$$p(H_0 \mid v_{1:n}, r_{1:n}) \propto p(H_0) \prod_i \frac{e^{-(v_i - H_0 r_i)^2 / 2(\beta_i^2 + H_0^2 \sigma_i^2)}}{\sqrt{2\pi}(\beta_i^2 + H_0^2 \sigma_i^2)^{1/2}}$$

For ‘n’ observed galaxies, each with different measurement error

Comparison of Theory and Experiment in Cosmology



1) Write down the "forward" probabilities needed for *simulation*

$$p(\Omega, s, d) = p(d \mid s)\, p(s \mid \Omega)\, p(\Omega)$$

2) Solve for the Bayesian Posterior as a conditional probability

$$p(\Omega \mid d) = \int ds\; p(\Omega, s \mid d)$$

Smoothing, Filtering, and Prediction for Discrete Time, Stochastic, Nonlinear Systems

Dynamics: ‘x’ is the underlying state, ‘y’ are observations

$$x_n = F(x_{n-1}, \theta) + u_n$$

$$y_n = G(x_n) + v_n$$

Joint Density: Simulate the observations…

$$p(y_{1:n}, x_{0:n}, \theta) = \left(\prod_{1:n} \frac{e^{-(y_i - G(x_i))^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \left(\prod_{1:n} \frac{e^{-(x_i - F(x_{i-1}, \theta))^2 / 2\kappa^2}}{\sqrt{2\pi}\kappa} \right) p(x_0)\, p(\theta)$$

- The “forward” conditional densities are all Gaussian here, but the various conditional densities are non-Gaussian for nonlinear dynamics!
- Leads to computationally challenging problems of state estimation and system identification

Smoothing, Filtering, and Prediction for Discrete Time, Stochastic, Nonlinear Systems

$$p(y_{1:n}, x_{0:n}, \theta) = \left(\prod_{1:n} \frac{e^{-(y_i - G(x_i))^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \left(\prod_{1:n} \frac{e^{-(x_i - F(x_{i-1}, \theta))^2 / 2\kappa^2}}{\sqrt{2\pi}\kappa} \right) p(x_0) p(\theta)$$

Problems: Given noisy state measurements, infer the past, present, and future state history, and possibly the dynamics

Smoothing

$$p(x_{0:n} \mid y_{1:n}, \theta) \propto \left(\prod_{1:n} \frac{e^{-(y_i - G(x_i))^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \left(\prod_{1:n} \frac{e^{-(x_i - F(x_{i-1}, \theta))^2 / 2\kappa^2}}{\sqrt{2\pi}\kappa} \right) p(x_0) p(\theta)$$

Filtering

$$p(x_i \mid y_{1:i}, \theta) = \int dx_{i-1} \left(\frac{e^{-(y_i - G(x_i))^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \left(\frac{e^{-(x_i - F(x_{i-1}, \theta))^2 / 2\kappa^2}}{\sqrt{2\pi}\kappa} \right) p(x_{i-1} \mid y_{1:i-1}, \theta)$$

Prediction

$$p(x_{i:n} \mid y_{1:i-1}, \theta) = \int dx_{i-1} \left(\prod_{i:n} \frac{e^{-(x_j - F(x_{j-1}, \theta))^2 / 2\kappa^2}}{\sqrt{2\pi}\kappa} \right) p(x_{i-1} \mid y_{1:i-1}, \theta)$$

System Id

$$p(\theta \mid y_{1:n}) \propto p(\theta) \int dx_{0:n} \left(\prod_{1:n} \frac{e^{-(y_i - G(x_i))^2 / 2\sigma^2}}{\sqrt{2\pi}\sigma} \right) \left(\prod_{1:n} \frac{e^{-(x_i - F(x_{i-1}, \theta))^2 / 2\kappa^2}}{\sqrt{2\pi}\kappa} \right) p(x_0)$$

If dynamics are linear, then everything is Gaussian, I.e. Kalman filtering

Markov Chain Monte Carlo

Goal - want to sample from some general probability density. In order to do so, run a Markov chain, such that:

$$\pi(x) \leftarrow_{n \to \infty} \int d(y_{0:n}) \left(\prod_{0<i\leq n} T(y_i \mid y_{i-1}) \right) p(y_0)$$

Sufficient conditions for convergence:

1). Stationarity

$$\pi(x) = \int dy \; T(x \mid y) \; \pi(y)$$

2). Irreducible: for any (x,y), an 'm' large enough such that

$$0 < T(x \mid y_m) \prod_{1:m} T(y_i \mid y_{i-1})$$

Metropolis-Hastings Algorithm

Construct transition matrix with any "proposal" matrix,
And find an "accept probability such that we satisfy
The condition of detailed balance:

$$\pi(x)w(y \mid x)A(y \mid x) = A(x \mid y)w(x \mid y)\pi(y)$$

Accept probability is any function such that: $\dfrac{A(y \mid x)}{A(x \mid y)} = \dfrac{\pi(y)w(x \mid y)}{\pi(x)w(y \mid x)}$

Maximal Accept Probability: $A(y \mid x) = \min\left[1, \dfrac{\pi(y)w(x \mid y)}{\pi(x)w(y \mid x)}\right]$

Algorithm:

1) Conditional on the past, propose new state
2) Accept with probability 0< A <= 1, otherwise keep past state
3) Continue

Proof of Stationarity for the Metropolis-Hastings Algorithm

The condition of detailed balance:

$$\pi(x)w(y \mid x)A(y \mid x) = A(x \mid y)w(x \mid y)\pi(y)$$

$$p_1(x) = \underbrace{\left[1 - \int dy\ A(y \mid x)w(y \mid x)\right]}_{\text{Reject probability from State 'x'}}\pi(x) + \underbrace{\int dy\ A(x \mid y)w(x \mid y)\pi(y)}_{\text{Accept transition TO State 'x' from any other 'y'}}$$

$$p_1(x) = \left[1 - \int dy\ A(y \mid x)w(y \mid x)\right]\pi(x) + \underbrace{\left(\pi(x)\int dy\ A(y \mid x)w(y \mid x)\right)}_{\text{(by detailed balance)}}$$

Proof of Convergence

$$\int dx \left|\pi(x) - p_n(x)\right| = \int dx \left|\int dy\, T(x \mid y)\left(\pi(y) - p_{n-1}(y)\right)\right|$$

$$\leq \int dy \int dx\, T(x \mid y) \left|\pi(y) - p_{n-1}(y)\right|$$

$$= \int dy \left|\pi(y) - p_{n-1}(y)\right|$$

After repeated iterations, we always get closer in probability to the target equilibrium measure (follows from stationarity)…

Symmetric Proposals

Recall general accept probability: $A(y \mid x) = \min\left[1, \frac{\pi(y)w(x \mid y)}{\pi(x)w(y \mid x)}\right]$

Symmetric Proposal:

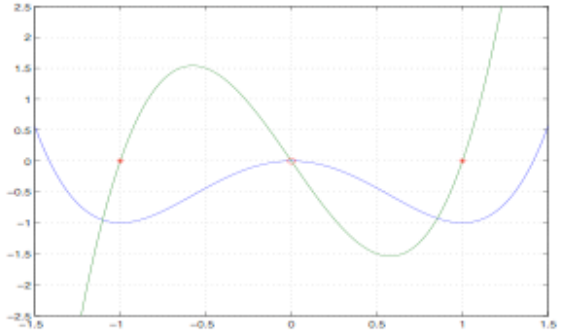
$w(y \mid x) \propto e^{-\beta\|y-x\|^2}$ gives $A(y \mid x) = \min\left[1, \frac{\pi(y)}{\pi(x)}\right]$

We always accept higher probability moves, and sometimes accept moves to lower probability

Double-Well Model

Particle in a potential well given by
$f(x) = -2x^2 + x^4$ and $g = f'$

Two minima at ± 1 and one stationary point at 0



Trivial climate
cf. Miller et al., Eyinck & Restrepo, etc.

Corresponds to the continuous-time diffusion

$$\begin{aligned} dx_t &= -g(x_t) + \kappa\, dB_t & \kappa \text{ constant} \\ y_t &= x_t + s_t & s_t \sim N(0, \sigma^2) \end{aligned}$$

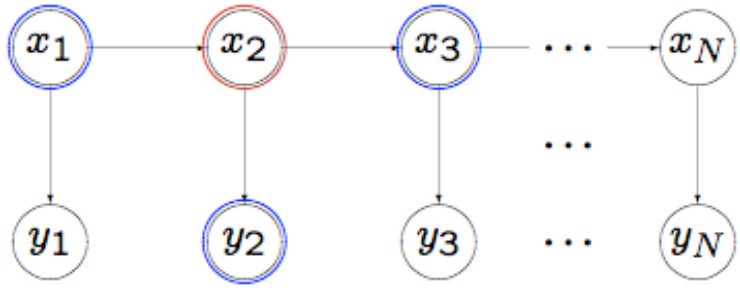
Approximate by δ-width finite differences

$$\begin{aligned} x_{n+1} &= x_n - \delta g(x_n) + r_n & r_n \sim N(0, \delta\kappa^2) \\ y_n &= x_n + s_n & s_n \sim N(0, \sigma^2) \end{aligned}$$

Need $\delta|g(\cdot)|, \sqrt{\delta}\kappa \ll 1$

MCMC: Time Series

Recall the Bayes network showing dependences



For now, we only want to estimate x_2 and other values are known

Apply Metropolis-Hastings recipe:

Propose a new value x_2', e.g. from $N(x_2, \delta\kappa^2)$ ($q(x_2' \mid x_2)$ is symmetric)

Compute the ratio

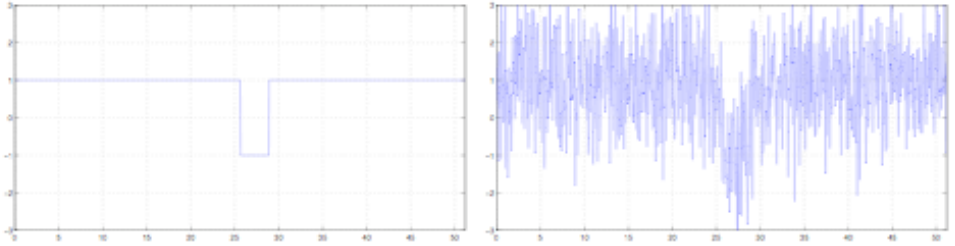
$$\frac{\pi(x')\, q(x \mid x')}{\pi(x)\, q(x' \mid x)} = \frac{p(x_2' \mid x_1)\, p(x_3 \mid x_2')}{p(x_2 \mid x_1)\, p(x_3 \mid x_2)} \times \frac{p(y_2 \mid x_2')}{p(y_2 \mid x_2)}$$

Combines a smoothness term and a data term

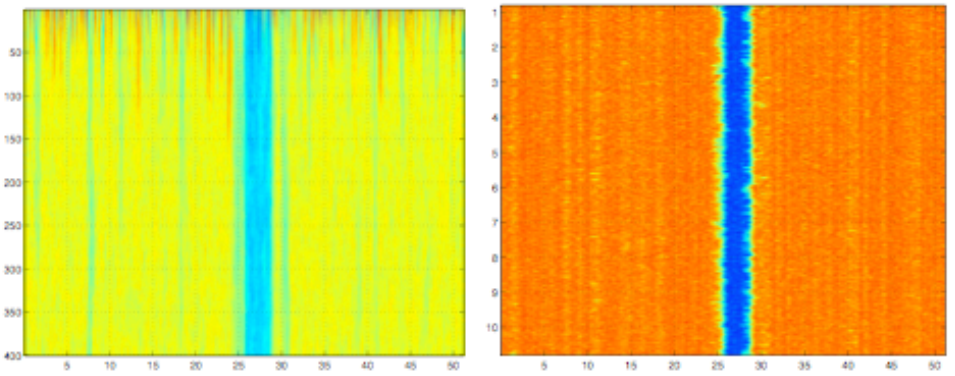
The full MCMC scheme sweeps over all unknown variables $x_1, \ldots, x_N$, proposing changes to each

MCMC Results

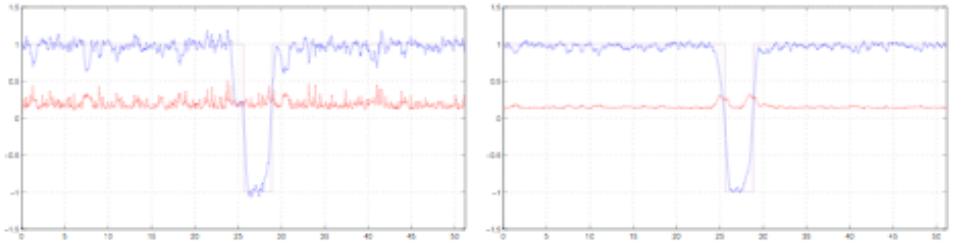
True x_n with data ($\kappa = 0.5$, $\sigma = 1$, $\delta = .05$, $T = 1024$)



Initial (1:400) and tail (10^4:10^5 downsampled)

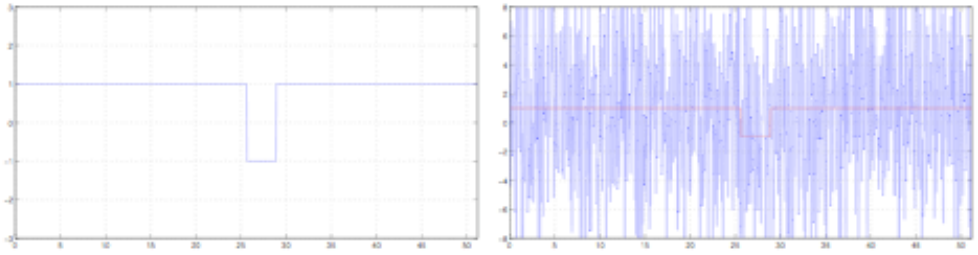


Statistics, initial and tail: Mean, RMS, and truth

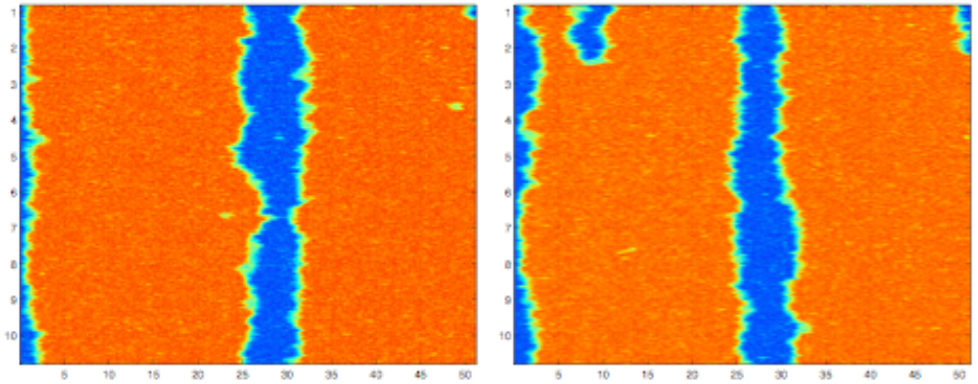


MCMC Results: High Noise

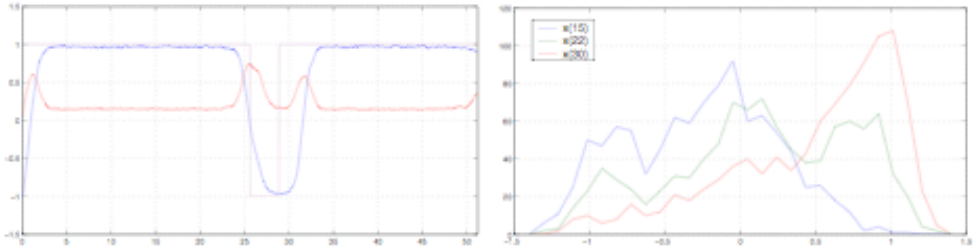
True x_n with data ($\kappa = 0.5$, $\sigma = 5$, $\delta = .05$, $T = 1024$)



Two MCMC series (10^4:10^5 downsampled 100:1)

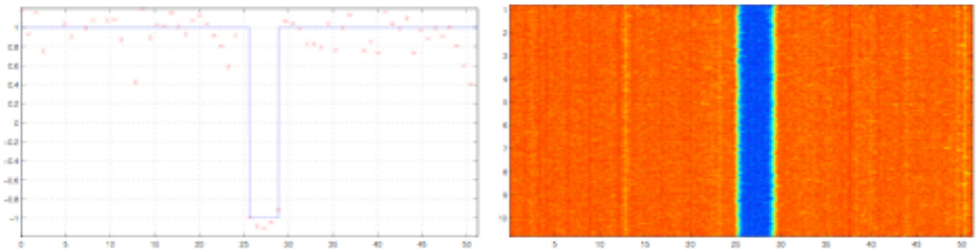


$E\,x_n$, $\sigma(x_n)$, truth; histograms of x_{15}, x_{22}, x_{30}

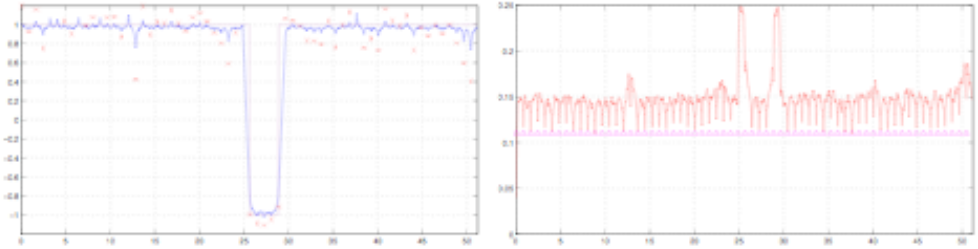


MCMC Results: Sparse Data

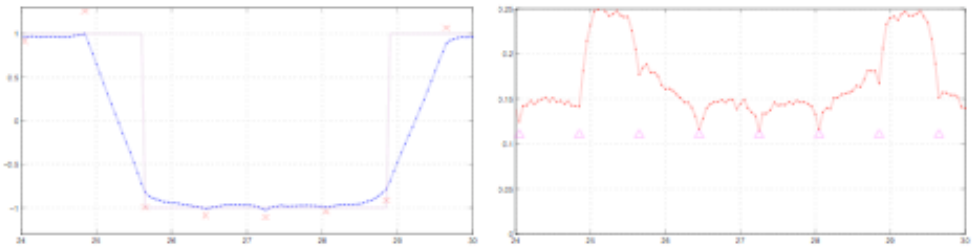
True x_n with data ($\kappa = 0.5$, $\sigma = 0.2$, $\delta = .05$, $T = 1024$)
But: downsample 16× (data rate $\ll$ simulation)



$E\,x_n$ and $\sigma(x_n)$: sampling interval indicated

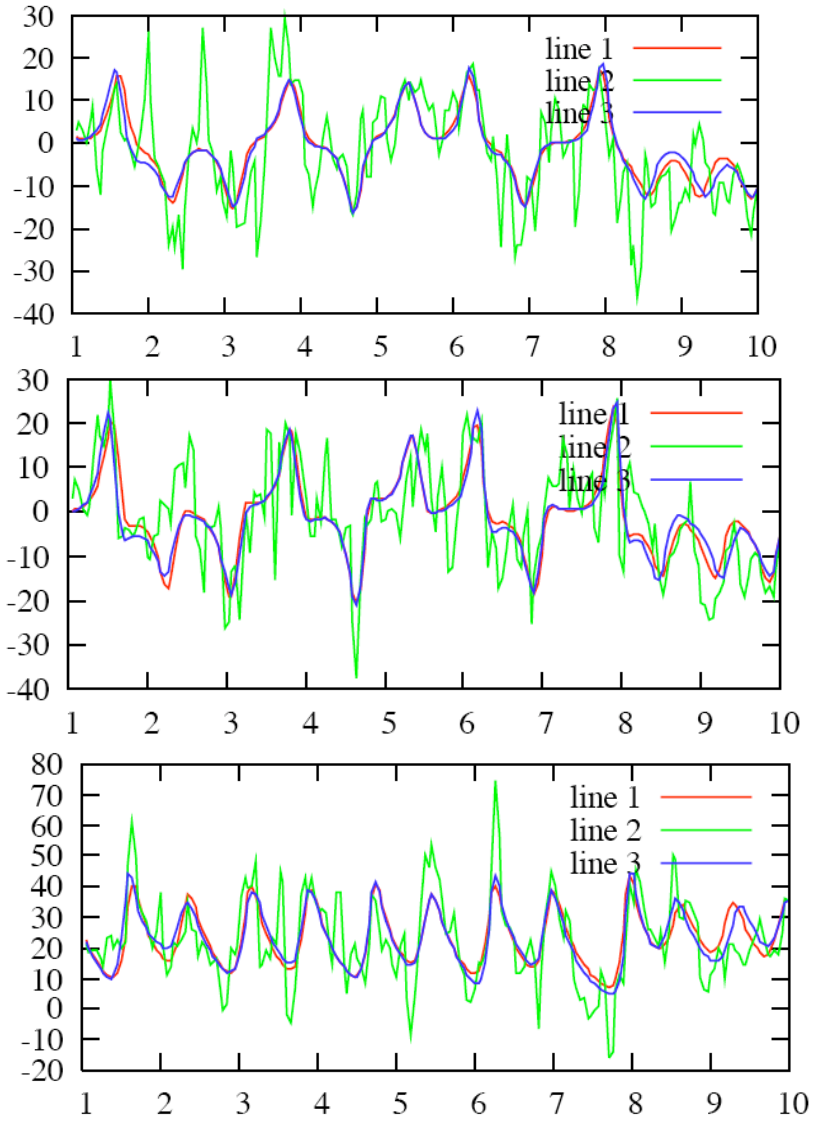


Data assimilation versus path smoothness



Near-gaussian statistics disallow fast jumps

Example - Inference of Solution Given Noisy Initial Guess (Lorenz Equations)



$$\dot{x} = a(y - x)$$

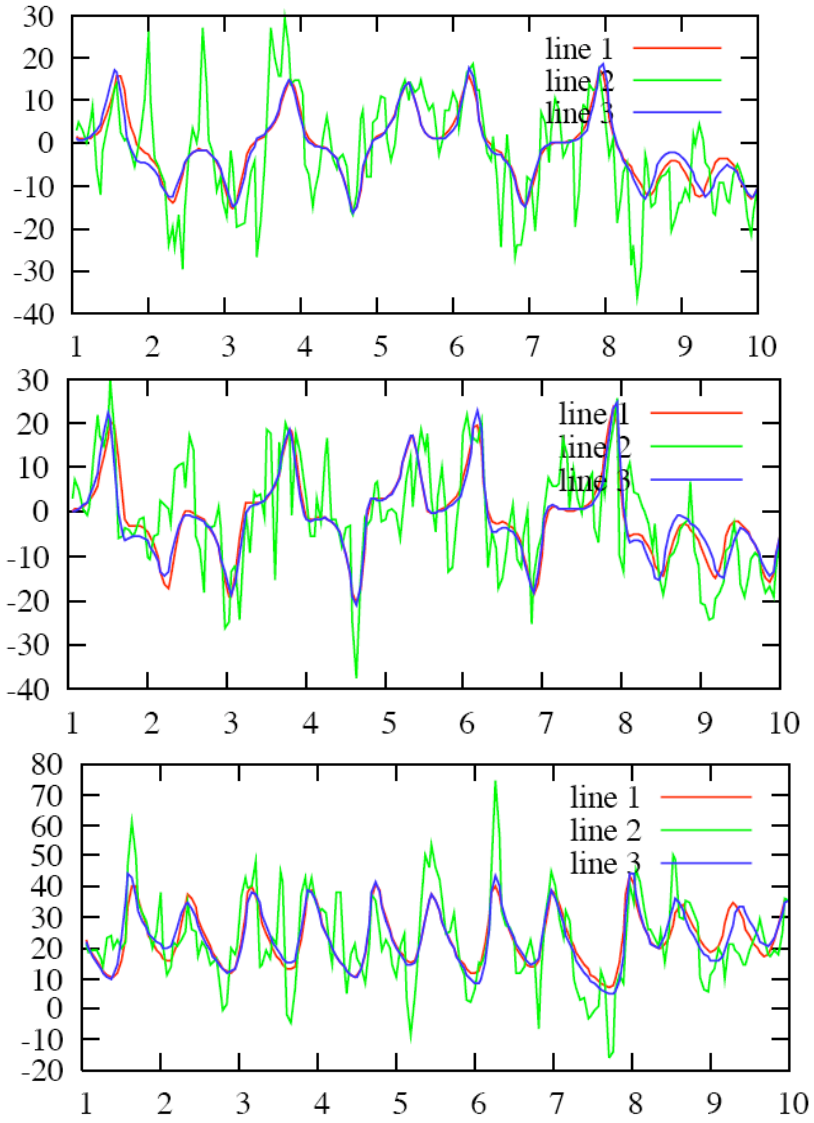
$$\dot{y} = bx - y - xz$$

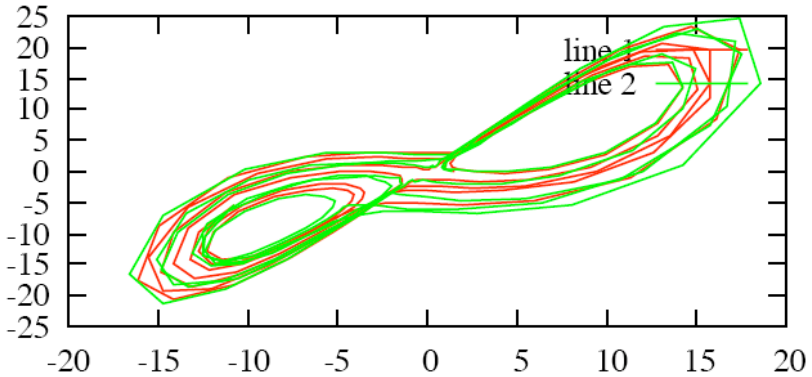
$$\dot{z} = xy - cz$$

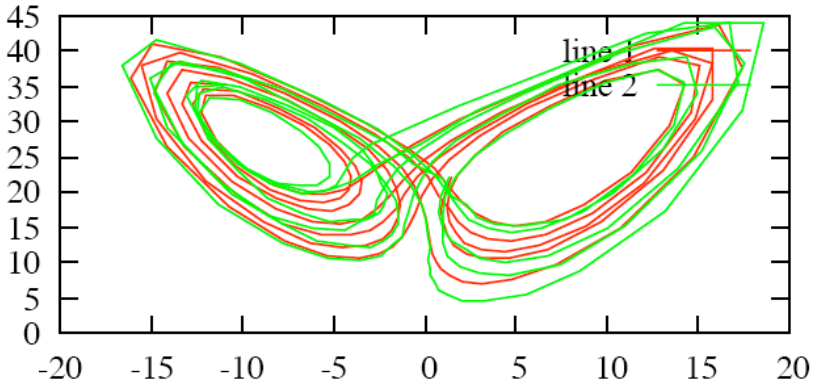
Observe, with noise…

$$d_{1:3} = x_{1:3} + n_{1:3}$$

Lorenz Equation Example…







Gibbs Sampling

Recall general accept probability: $A(y \mid x) = \min\left[1, \frac{\pi(y)w(x \mid y)}{\pi(x)w(y \mid x)}\right]$

K-dimensional state space state transition:
In sequential or random order, propose new state from 1D conditionals:

$$(x_1,\ldots,x_{j-1},x_j,x_{j+1},\ldots,x_n) \rightarrow (x_1,\ldots,x_{j-1},y_j,x_{j+1},\ldots,x_n)$$

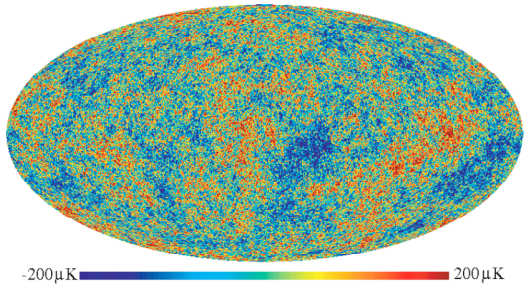
IF we propose exactly from $\pi(y_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)$

$$A = \min\left[\underbrace{\frac{\pi(y_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)}{\pi(x_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)}}_{\text{Ratio of new to old state}} \underbrace{\frac{\pi(x_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)}{\pi(y_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_n)}}_{\text{Proposal ratio (new to old/old to new)}}\right] = 1$$

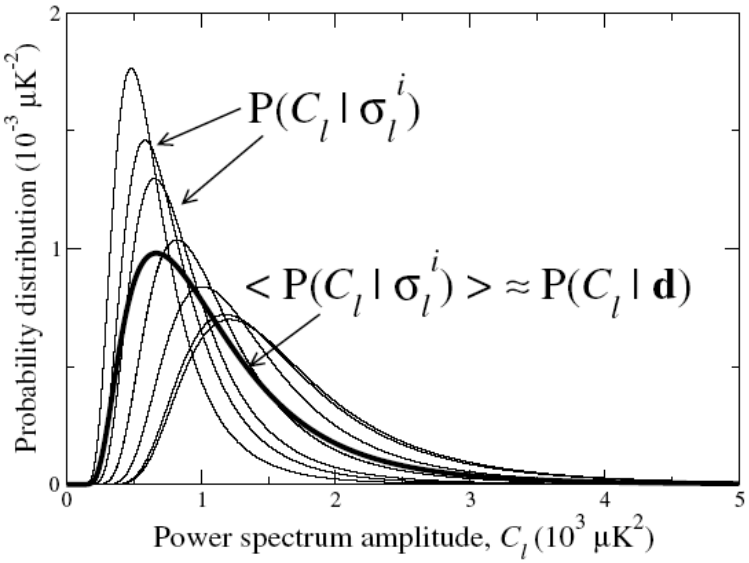
Gibbs Sampling Cosmological Parameters

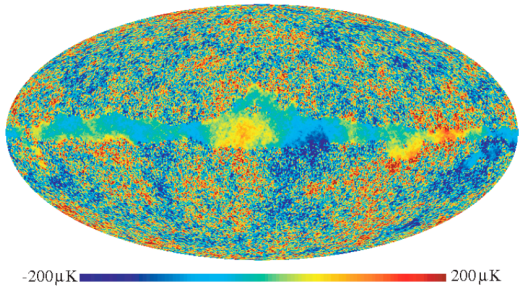
$$T(\Omega, s \mid \Omega', s') = p(\Omega \mid s)\, p(s \mid d, \Omega')$$

$$p(\Omega \mid s) \propto p(\Omega) \prod_{lm} \frac{e^{-\sigma_l / 2C_l(\Omega)}}{\sqrt{2\pi} C_l^{1/2}(\Omega)}$$

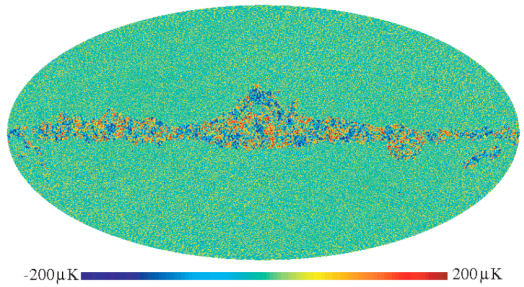


Sum of the two maps is a sample from the conditional





Random variation consistent with our uncertainty



Mean Field map given power spectrum guess

Auxiliary Variable Methods

Embed state space into higher dimensional space:

$$\pi(x) \rightarrow \pi(y \mid x)\pi(x)$$

The marginal density is still what we want to sample - so sample from the joint space, and "ignore" the new variable 'y'…

Propose according to: $w(x, y) = w(x \mid y)w(y)$

$$A(x', y' \mid x, y) = \min\left[1, \frac{\pi(y' \mid x')\pi(x')}{\pi(y \mid x)\pi(x)} \frac{w(x \mid y)w(y)}{w(x' \mid y')w(y')}\right]$$

- Can be useful for multi-modal measures (I.e. well approximated by a mixture of Gaussians)
- Good overview of these methods in: R. M. Neal, "Probabilistic Inference Using Markov Chain Monte Carlo", technical report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, Sept. 1993.

Summary

- Probabilistic methods for model based reasoning, because either model has stochastic elements, uncertainty (measurement error), or both
- MCMC and other sampling methods allow us to quantify what has been learned from either computation or measurement
- For high dimensional problems, and/or complicated (I.e. nonlinear) models, probabilistic inference for these cases was previously intractable, but now computationally within reach.
- The way in which simulations and measurement are used in reaching conclusions will dramatically change - much more detail of the measurement process and details of the theory can be directly addressed with algorithms such as MCMC and its variants…

Selected MCMC References for Additional Reading

- Julian Besag, “Spatial Interaction and the Statistical Analysis of Lattice Systems”, J. of the Royal Statistical Society B, 36,192-236, 1974.
- C. Andrieu, et al, “An Introduction to MCMC for Machine Learning”, Machine Learning, 50,5-32, 2003.
- R. M. Neal, “Probabilistic Inference Using Markov Chain Monte Carlo”, technical report CRG-TR-93-1, Department of Computer Science, University of Toronto, Sept. 1993.
- L. Tierney, A. Mira, “Some Adaptive Monte Carlo Methods for Bayesian Inference”, Statistics in Medicine, 18, 2507-2515, 1999.
- H. Haario et al, “An Adaptive Metropolis Algorithm”, Bernoulli, vol.7, no.2,2001, 223-242
- MCMC preprint site, with links to software, http://www.statslab.cam.ac.uk/~mcmc/

References for Presented Examples

- T. M. Chin, M. J. Turmon, J. B. Jewell, and M. Ghil (2007), "An ensemble-based smoother with retrospectively updated weights for highly nonlinear systems,", Mon.
- . Rev., 135(1), 186-202.
- "*Application of Monte Carlo Algorithms to the Bayesian Analysis of the Cosmic Microwave Background*", J. Jewell, S. Levin, C. A. Anderson, Astrophysical Journal, Volume 609, Issue 1, pp. 1-14
- "*Global, exact cosmic microwave background data analysis using Gibbs sampling*", B. D. Wandelt et al, Physical Review D, vol. 70, Issue 8, id. 083511
- "*Power Spectrum Estimation from High-Resolution Maps by Gibbs Sampling*", H.K. Eriksen et al, Astrophysical Journal Supplement Series, Volume 155, Issue 2, pp. 227-241
- "*Bayesian Power Spectrum Analysis of the First-Year Wilkinson Microwave Anisotropy Probe Data*", I.J. O'Dwyer et al, Astrophysical Journal, Volume 617, Issue 2, pp. L99-L102.
- "*Cosmological parameter constraints as derived from the Wilkinson Microwave Anisotropy Probe data via Gibbs sampling and the Blackwell-Rao estimator*", M. Chu, et al, Physical Review D, vol. 71, Issue 10, id. 103002
- "*A Reanalysis of the 3 Year Wilkinson Microwave Anisotropy Probe Temperature Power Spectrum and Likelihood*", H. K. Eriksen et al, Astrophysical Journal, Volume 656, Issue 2, pp. 641-652
- "*Estimation of Polarized Power Spectra by Gibbs Sampling*", D. Larson et al, Astrophysical Journal, Volume 656, Issue 2, pp. 653-660
- "*Bayesian analysis of the low-resolution polarized 3-year WMAP sky maps*", H.K. Eriksen et al., eprint arXiv:0705.3643, submitted to ApJL