Approximate Inference by Stochastic Simulation/Sampling Methods

Zhenke Wu
Department of Biostatistics
University of Michigan

October 20, 2016

Inference Techniques

- Central task of applying probabilistic models:
  - Evaluate the posterior: $p(Z \mid X^{obs})$
- Exact Inference Algorithms
  - Variable elimination
  - Message-passing (sum-product, max-product)
  - Junction-Tree algorithms
- Approximate Inference
  - To overcome the exponential (of graph treewidth) computational/space complexity for exact inference algorithms

Approximate Inference Techniques

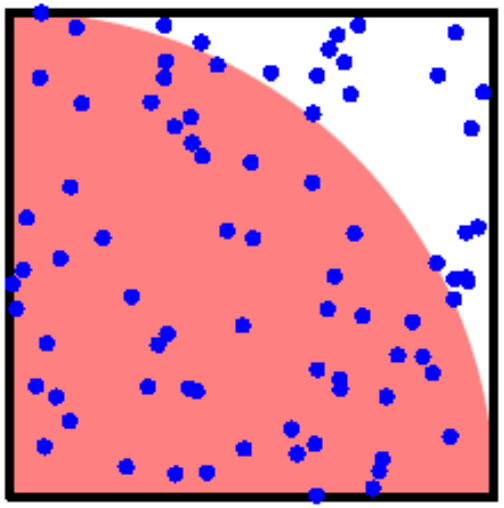
- Stochastic approximation
  - Given infinite computational resources, they can generate exact results; the approximation arises from the use of a finite amount of processor time
  - Monte Carlo
  - Buffon's needle;
  - Direct sampling (Box-Muller for bivariate Gaussian; Inverse Transformation)
  - Popular ones: Rejection sampling; Slice sampling; Likelihood weighting
  - Markov Chain Monte Carlo:
    - Metroplis-Hastings sampling (Metropolis N, Rosenbluth AW, Rosenbluth, Teller AH, Teller E (1953), Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics*); Extended by Hastings WK (1970) *Biometrika*.
    - Gibbs sampling (Geman and Geman, 1984), etc.
  - Hamiltonian Monte Carlo
  - Scalable Bayesian algorithms: Parallel and distributed MCMC (research frontier; e.g., Scott SL et al. 2013, consensus Monte Carlo)
- Need to address:
  - How to draw samples?
  - How to make efficient use of the obtained samples?
  - When to stop?

Approximate Inference Techniques

- Deterministic approximation (later lectures)
  - Scale well to large applications, natural language processing (Blei et al. (2003) JMLR, latent Dirichlet allocation); image processing
  - Based on analytic approximations to the posterior distribution, for example, assume specific factorization, or parametric form such as Gaussian (work with a smaller class of distributions that are close to the target)
  - Loopy belief propagation
  - Mean field approximation
  - Expectation propagation
  - …

Monte Carlo

1. Get expectation that is difficult to calculate



In general:

$$\int f(x)P(x)\,\mathrm{d}x \approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

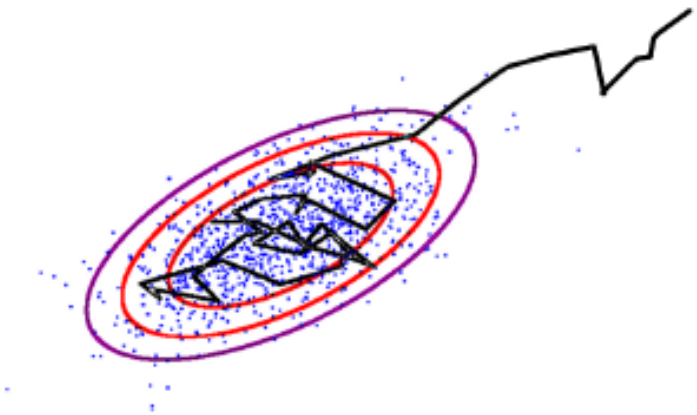
Example: making predictions

$$p(x|\mathcal{D}) = \int P(x|\theta,\mathcal{D})P(\theta|\mathcal{D})\,\mathrm{d}\theta$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} P(x|\theta^{(s)},\mathcal{D}), \quad \theta^{(s)} \sim P(\theta|\mathcal{D})$$

Markov chain Monte Carlo

2. Construct correlated samples that explore target distribution.



Markov steps, $x^{(s)} \sim T\left(x^{(s)} \leftarrow x^{(s-1)}\right)$

MCMC gives approximate, correlated samples

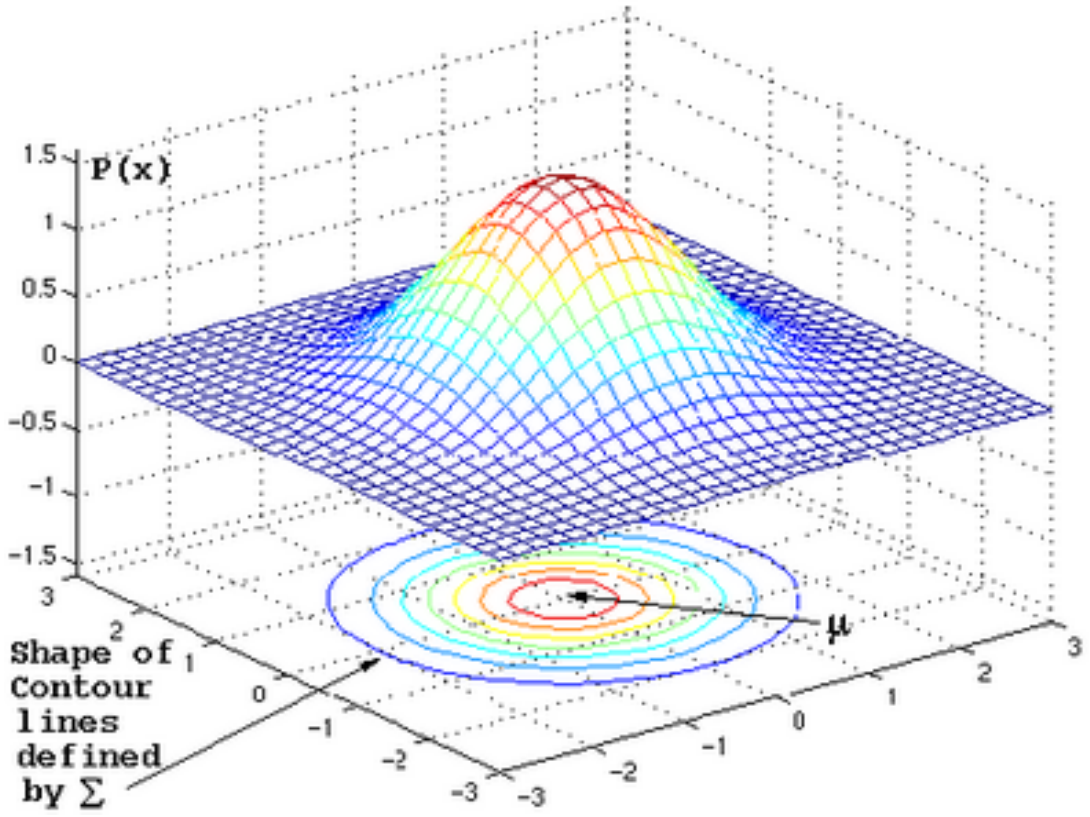
$$\mathbb{E}_P[f] \approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})$$

Example transitions:

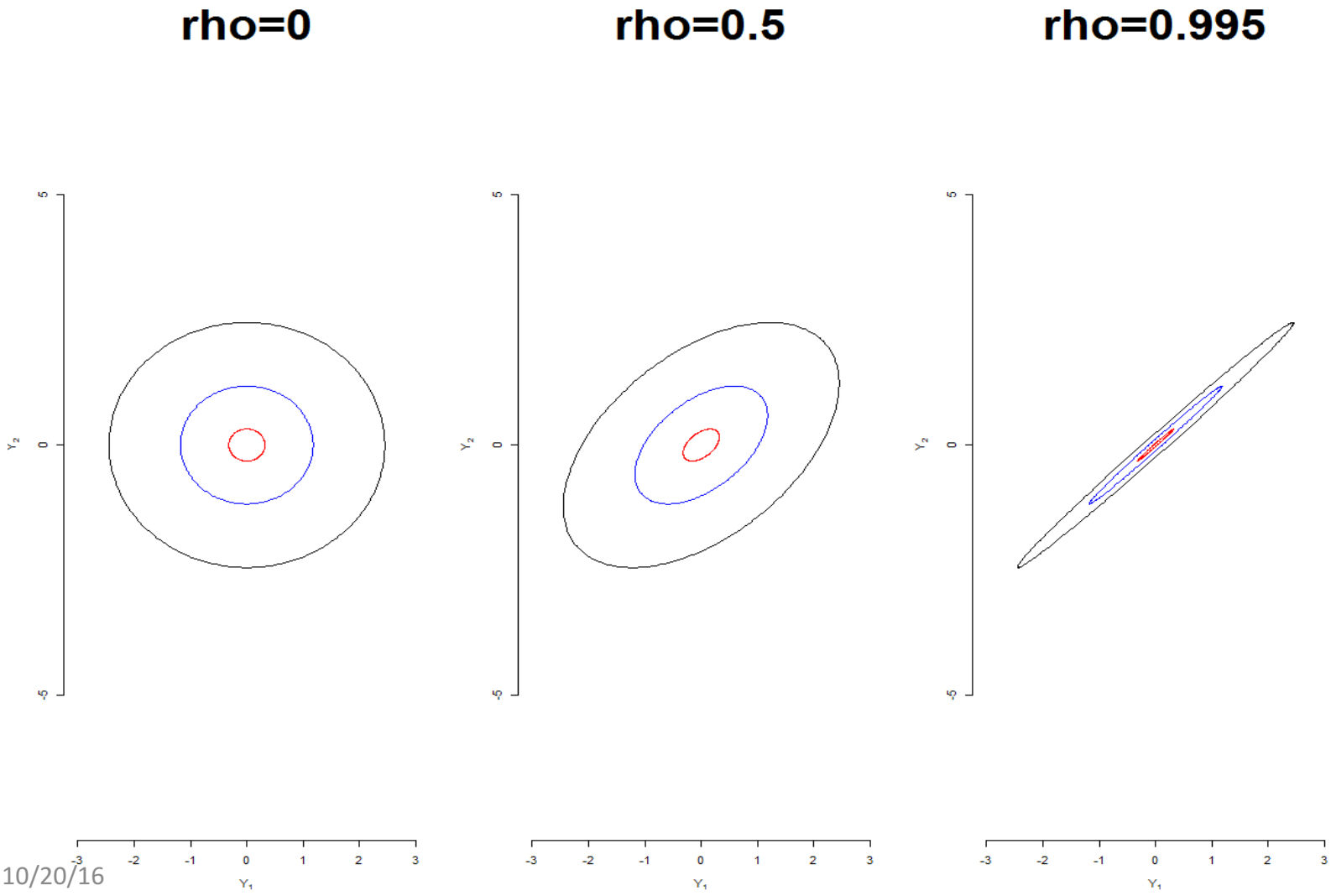
Metropolis–Hastings: $T(x' \leftarrow x) = Q(x'; x)\ \min\left(1,\ \frac{P(x')\,Q(x; x')}{P(x)\,Q(x'; x)}\right)$

Gibbs sampling: $T_i(\mathbf{x}' \leftarrow \mathbf{x}) = P(x_i' \,|\, \mathbf{x}_{j\neq i})\,\delta(\mathbf{x}'_{j\neq i} - \mathbf{x}_{j\neq i})$

Example: Bivariate Gaussian



Bivariate Gaussian



Gibbs Sampler

A method with no rejections:

- Initialize $\mathbf{x}$ to some value
- Pick each variable in turn or randomly and resample $P(x_i|\mathbf{x}_{j\neq i})$

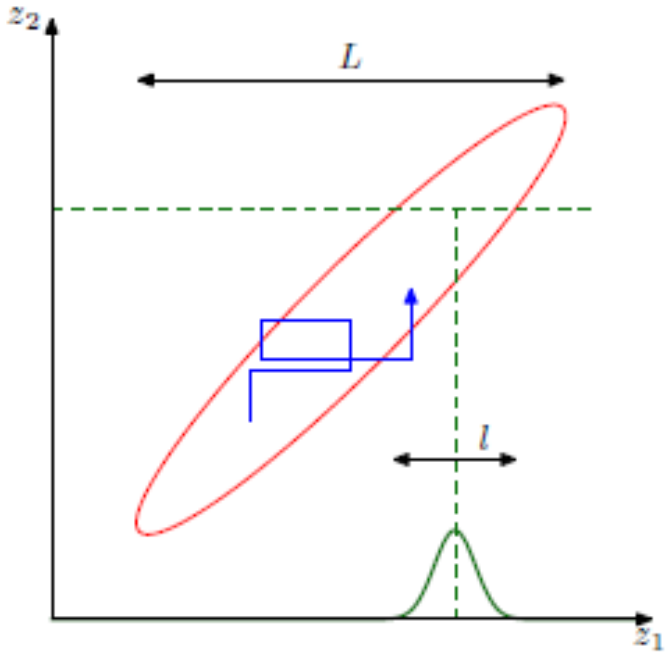
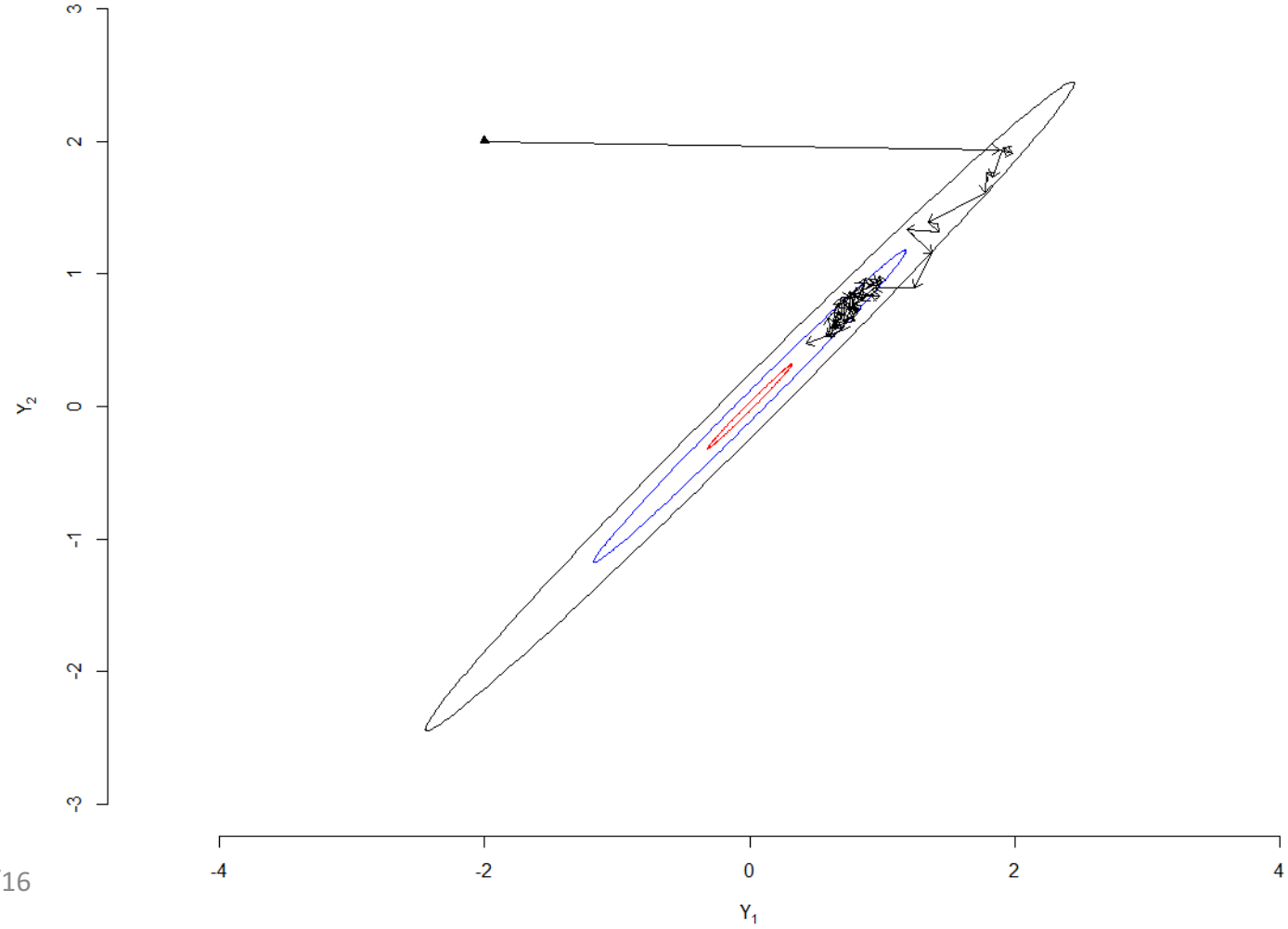
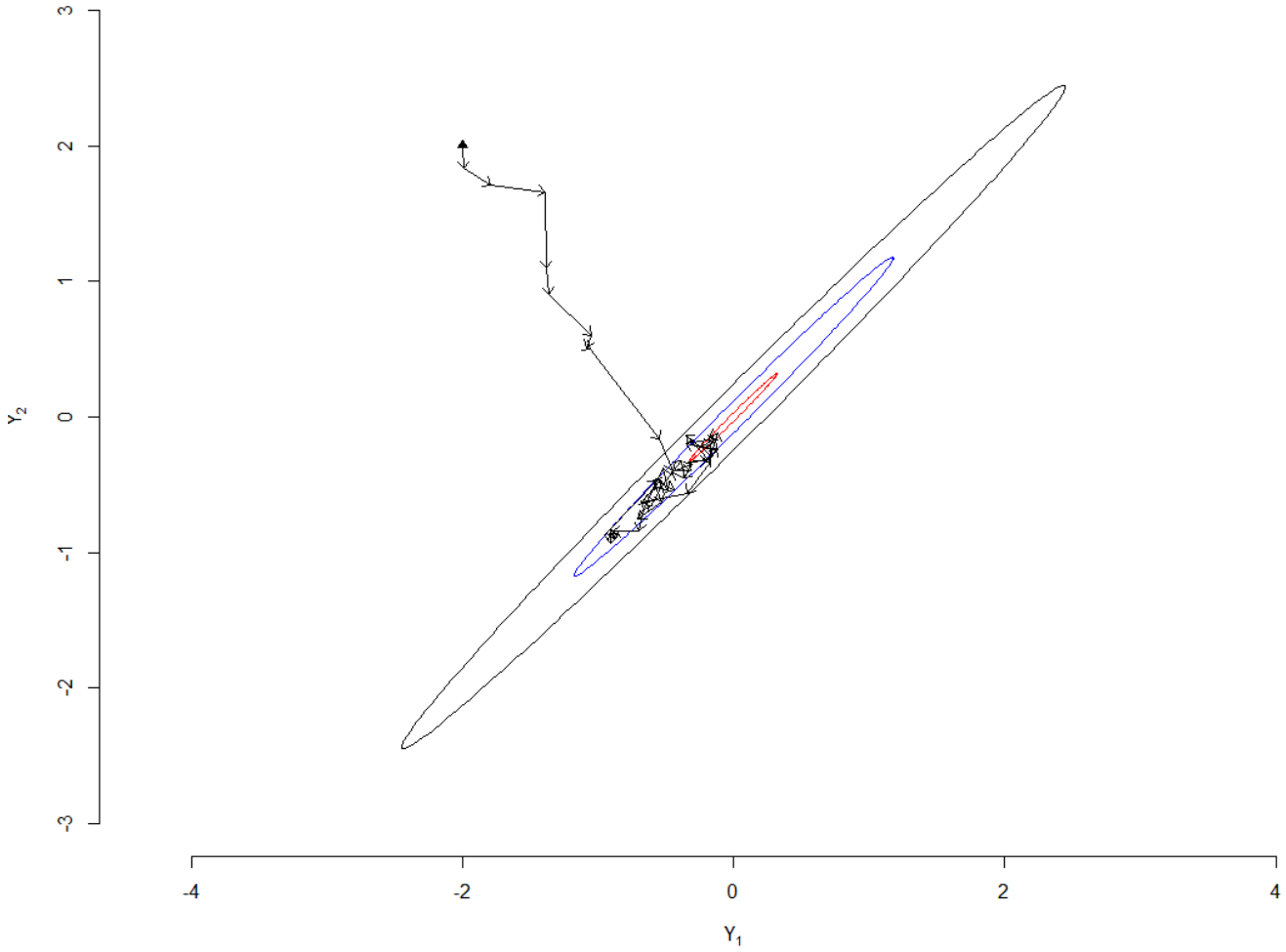


Figure from PRML, Bishop (2006)

Simple Gibbs Sampler
First 50 Samples; Rho=0.995

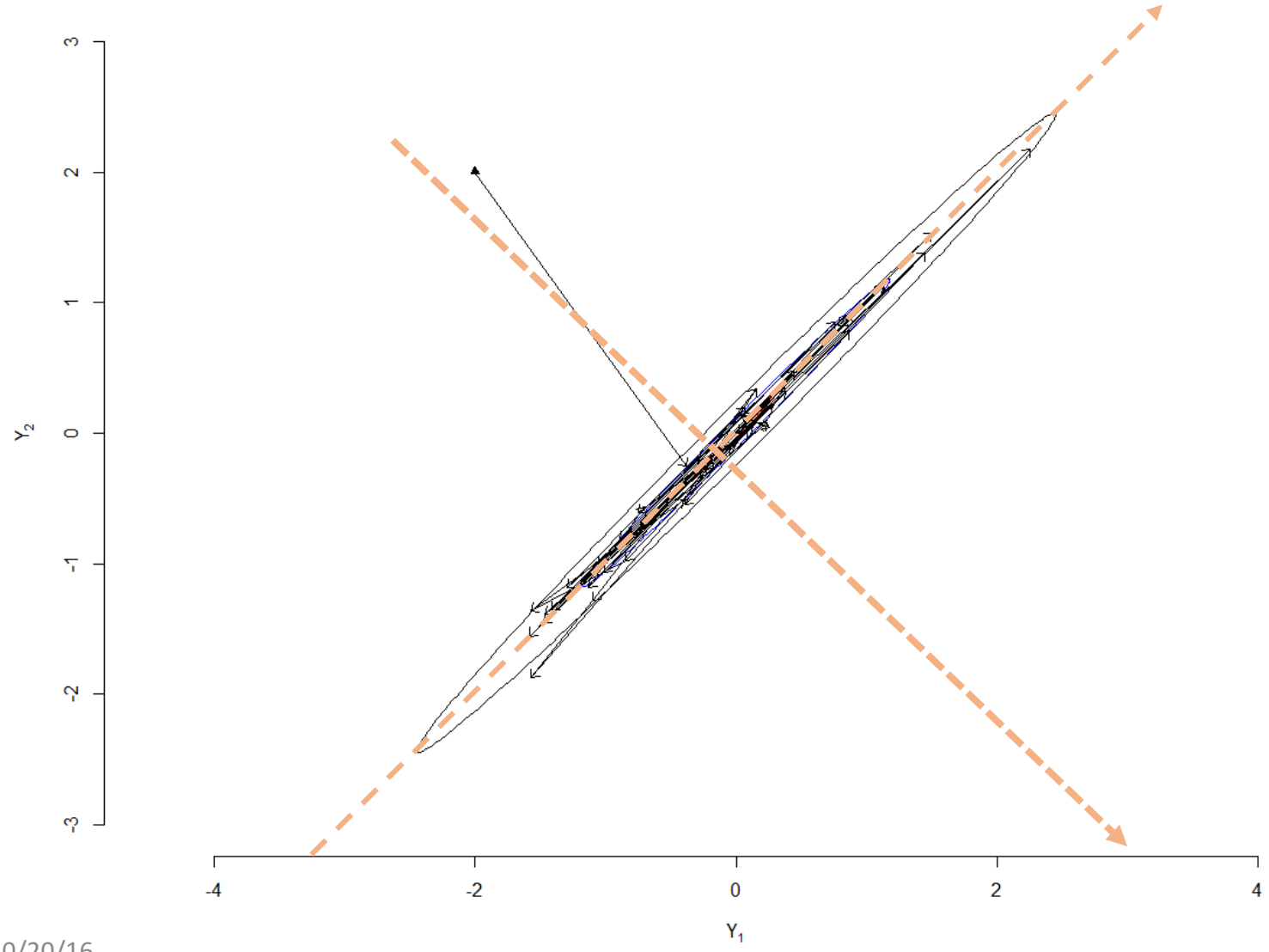


Slice Sampler
First 50 samples; Rho=0.995



Gibbs Sampler on Rotated Coordinate

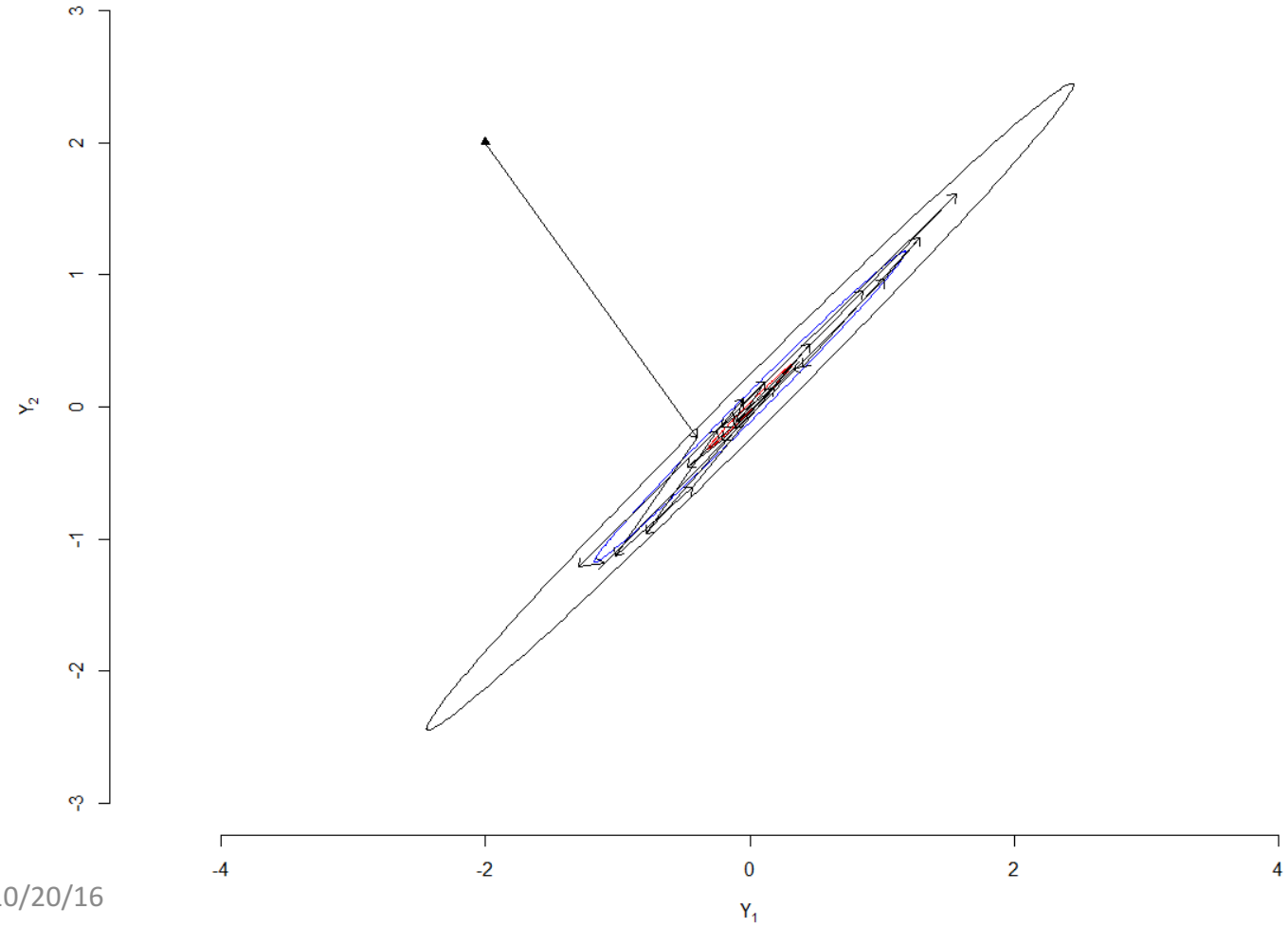
Rho=0.995



Lessons Learned

- Re-parametrize the model or de-correlate posterior shape when possible
- The covariance structure of the posterior density guides improvement of MCMC algorithm
- In WinBUGS, first 5,000 samples should not be used for inference: they are used to explore posterior shape and to tune proposal parameters

Hamiltonian Monte Carlo (HMC)
First 50 samples; Rho=0.995



Hamiltonian Monte Carlo (HMC)

- Computing core of Stan http://mc-stan.org/
- Advantage
  - Super fast
  - Cross-platform
  - Has algorithms to determine the number of leapfrog steps (No-U-Turn sampler)
- Limitation
  - Does not support sampling discrete parameters (no associated gradient required for sampling algorithm)
  - Can trick Stan to do the job in some parametric models

Comments

- A good posterior sampling algorithm is the one that
  - Use maximal information from the posterior terrain
  - Bold but wise explorations
- Play with the code: https://github.com/zhenkewu/demo_code
- Chapter 11, Bishop CM (2007) Pattern Recognition and Machine Learning.