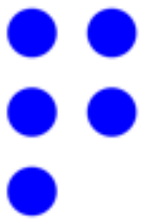
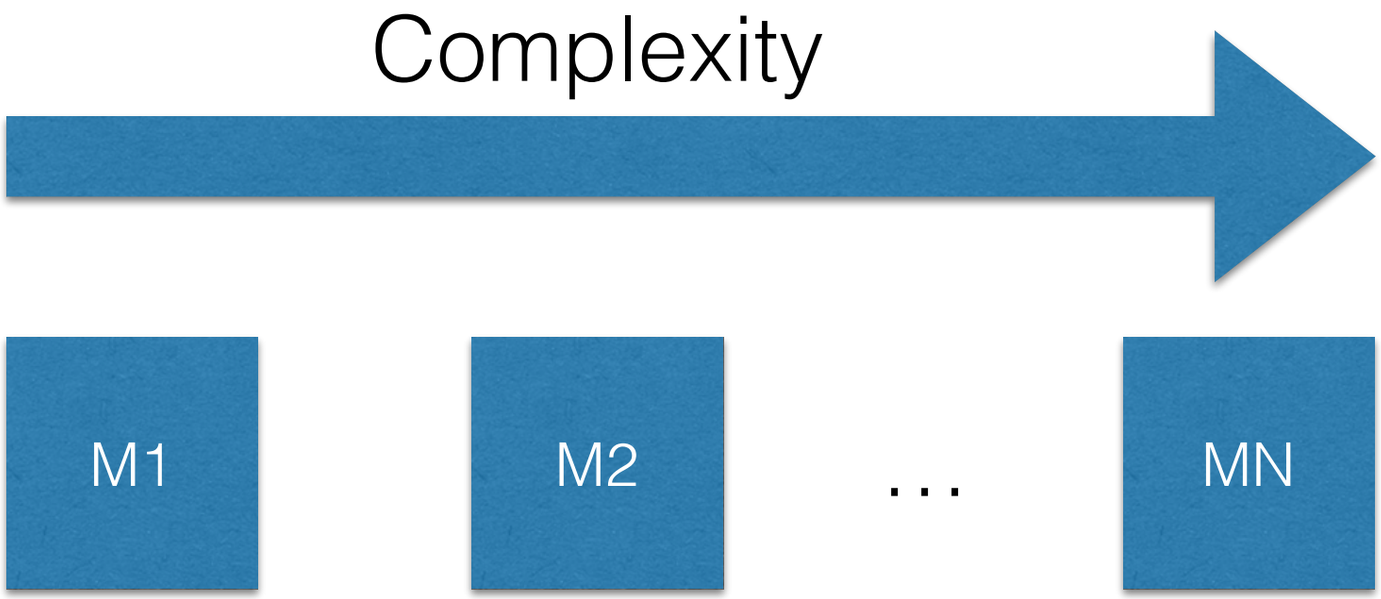
Lecture 26
Bayesian Model Comparison



PREDICTIVE
SCIENCE LABORATORY

The problem



What is the best model?

PREDICTIVE
SCIENCE LABORATORY

“We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.”

–Isaac Newton

PREDICTIVE
SCIENCE LABORATORY

“Everything should be made as simple as possible, but not simpler.”

–Albert Einstein

Occam’s Razor

Pick the simplest model that explains the data…

Example 1: Polynomial Regression

- M1: linear regression
- M2: quadratic regression
- M3: cubic regression
- …

Example 2: Model Calibration

- M1: simple physical model
- M2: more complex physical model
- M3: super complex physical model
- …

Example 3: Hypothesis Testing

- H1
- H2
- H3
- …

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

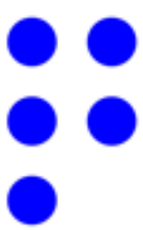
Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

Example 3: Hypothesis Testing

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*



Bayesian Model Selection

Prior over models:

$$M_i \sim p(M_i)$$

Prior

state of knowledge

Prior over model parameters:

$$\theta_i \mid M_i \sim p(\theta_i \mid M_i)$$

Likelihood of the data:

$$D \mid \theta_i, M_i \sim p(D \mid \theta_i, M_i)$$

Measurement process

Bayesian Model Selection

Prior state of knowledge

$$M_i \sim p(M_i)$$

$$\theta_i \mid M_i \sim p(\theta_i \mid M_i)$$

Measurement process

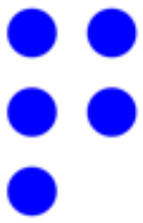
$$D \mid \theta_i, M_i \sim p(D \mid \theta_i, M_i)$$

Posterior over the parameters:

$$p(\theta_i \mid D, M_i) = \frac{p(D \mid \theta_i, M_i) p(\theta_i \mid M_i)}{p(D \mid M_i)}$$

where the **evidence** is the log of the normalizing constant:

$$p(D \mid M_i) = \int p(D \mid \theta_i, M_i) p(\theta_i \mid M_i) d\theta_i$$



Bayesian Model Selection

Prior state of knowledge

$$M_i \sim p(M_i)$$

$$\theta_i \mid M_i \sim p(\theta_i \mid M_i)$$

Measurement process

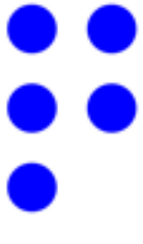
$$D \mid \theta_i, M_i \sim p(D \mid \theta_i, M_i)$$

$$p(\theta_i \mid D, M_i) = \frac{p(D \mid \theta_i, M_i) p(\theta_i \mid M_i)}{p(D \mid M_i)} \qquad p(D \mid M_i) = \int p(D \mid \theta_i, M_i) p(\theta_i \mid M_i) d\theta_i$$

Posterior over models:

$$p(M_i \mid D) \propto p(D \mid M_i) p(M_i)$$

How do we calculate the model evidence is the big problem!



PREDICTIVE
SCIENCE LABORATORY

Model Averaging

Assume you have this:

$$p(M_i \mid D) \propto p(D \mid M_i) p(M_i)$$

How do you make predictions?

Predictive distribution:

$$p(y \mid x) = \sum_i p(y \mid x, D, M_i) p(M_i, D)$$

$$p(y \mid x, D_i, M_i) = \int p(y \mid x, \theta_i, M_i) p(\theta_i \mid D, M_i) d\theta_i$$

Model predictive distribution.

Bayesian Model Selection

Assume you have this:

$$M_{i^*} = \text{argmax}_i \, p(M_i \mid D) = \text{argmax}_i \, p(D \mid M_i) p(M_i)$$

How do you make predictions?

$$p(y \mid x, D, M_{i^*}) = \int p(y \mid x, \theta_{i^*}, M_{i^*}) p(\theta_{i^*} \mid D, M_{i^*}) d\theta_{i^*}$$

Model Evidence Calculation

$$p(D \mid M_i) = \int p(D \mid \theta_i, M_i) p(\theta_i \mid M_i) d\theta_i$$

Analytically…
if you are lucky enough

The Laplace Approximation

Why don't we just approximate this:

$$p(\theta_i \mid D, M_i) = \frac{p(D \mid \theta_i, M_i) p(\theta_i \mid M_i)}{p(D \mid M_i)}$$

by a Gaussian!

The Laplace Approximation

Let's do it for an arbitrary distribution:

$$p(z) = \frac{f(z)}{Z}$$

Use some Taylor expansion stuff:

$$\log f(z) \approx \log f(z_0) - \frac{1}{2}(z - z_0)^T \nabla^2 \log f(z_0)(z - z_0).$$

where:

$$z_0 = \text{argmax}_z \log f(z).$$

Taking the exponential:

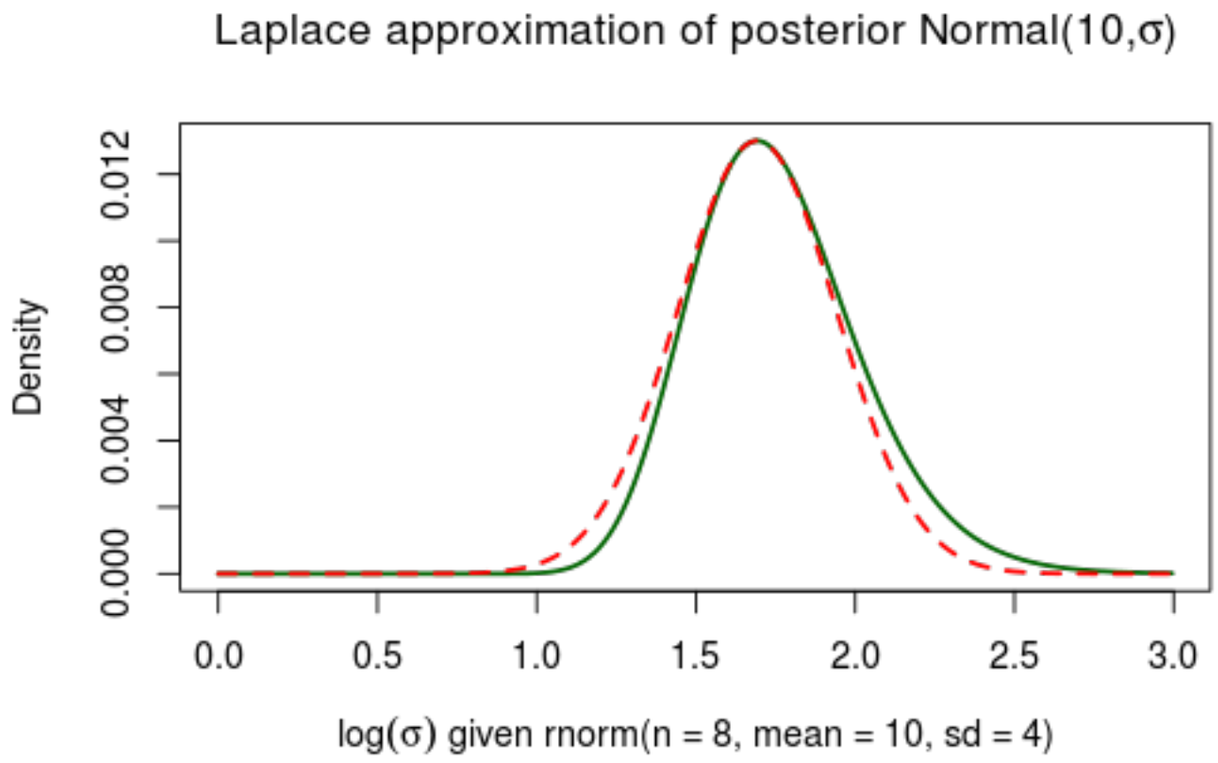
$$f(z) \approx f(z_0) \exp\left\{-\frac{1}{2}(z - z_0)^T \nabla^2 \log f(z_0)(z - z_0)\right\} = \mathcal{N}\left(z \mid f(z_0), \nabla^2 \log f(z_0)^{-1}\right)$$

Thus:

$$Z \approx \int f(z)dz = f(z_0)\frac{(2\pi)^{d/2}}{|\nabla^2 \log f(z_0)|^{1/2}}.$$

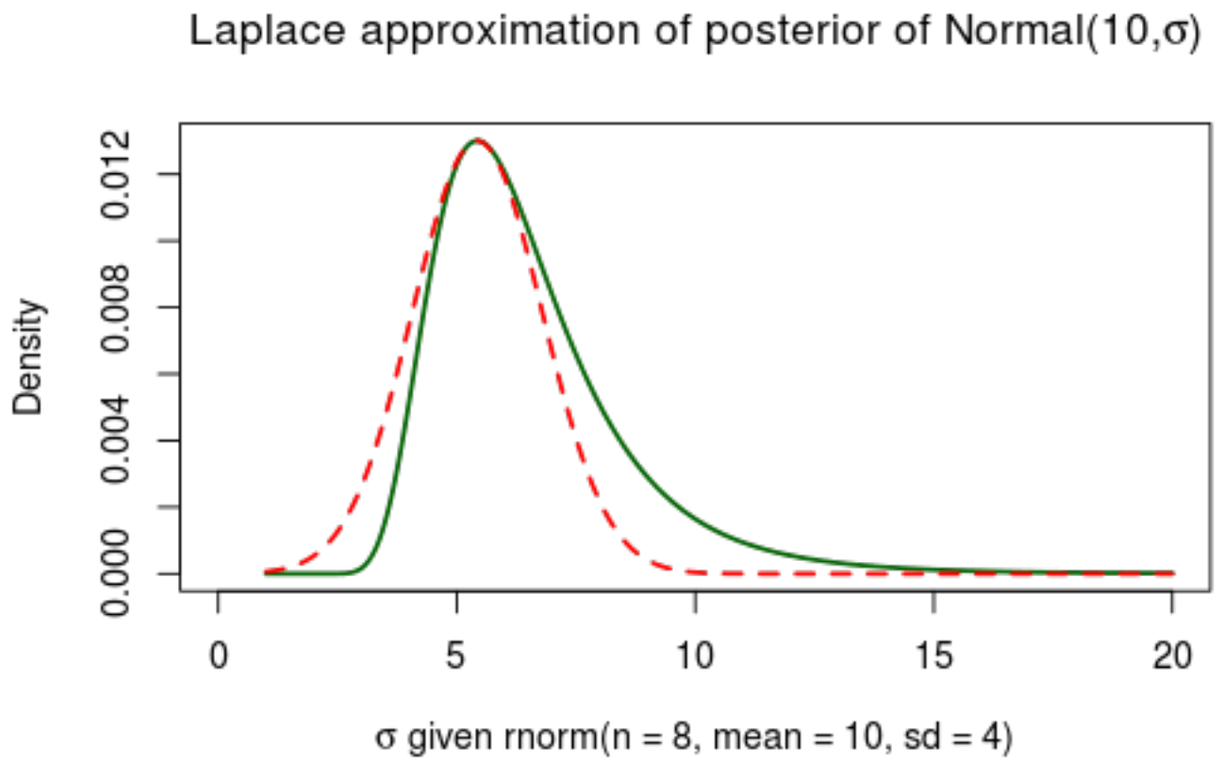
dimension of z

The Laplace Approximation



PREDICTIVE
SCIENCE LABORATORY

The Laplace Approximation



PREDICTIVE
SCIENCE LABORATORY

The Laplace Approximation

Why don't we just approximate this:

$$p(\theta_i \mid D, M_i) = \frac{p(D \mid \theta_i, M_i)\, p(\theta_i \mid M_i)}{p(D \mid M_i)}$$

by a Gaussian!

$$p(z) = \frac{f(z)}{Z}$$

Laplace

$$\log Z \approx \log f(z_0) + \frac{1}{2} \log \left| \nabla^2 \log f(z_0) \right| + \frac{d}{2} \log 2\pi.$$

$$\log p(D \mid M_i) \approx \log\left(p(D \mid \theta_i^*, M_i)\, p(\theta^*_i \mid M_i) \right) +$$

$$\frac{1}{2} \left| \nabla^2 \log\left(p(D \mid \theta_i^*, M_i)\, p(\theta^*_i \mid M_i) \right) \right| + \frac{d}{2} \log 2\pi$$

MAP estimate

PREDICTIVE
SCIENCE LABORATORY

Bayesian Information Criterion

$$\log p(D \mid M_i) \approx \log\Big(p(D \mid \theta_i^*, M_i) p(\theta^*_i \mid M_i)\Big) +$$

$$\frac{1}{2}\Big|\nabla^2 \log\Big(p(D \mid \theta_i^*, M_i) p(\theta^*_i \mid M_i)\Big)\Big| + \frac{d}{2}\log 2\pi$$

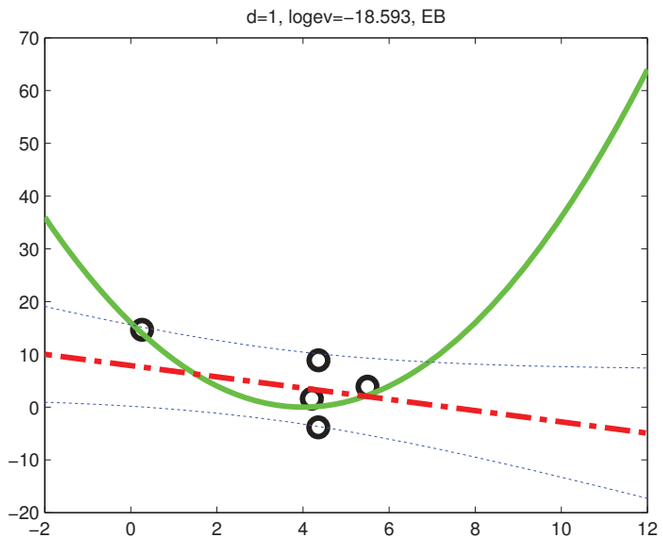
Crudest possible approximation
(flat prior, full rank Hessian, etc.)

$$\log p(D \mid M_i) \approx \log\Big(p(D \mid \theta_i^*, M_i)\Big) - \frac{d}{2}\log n$$

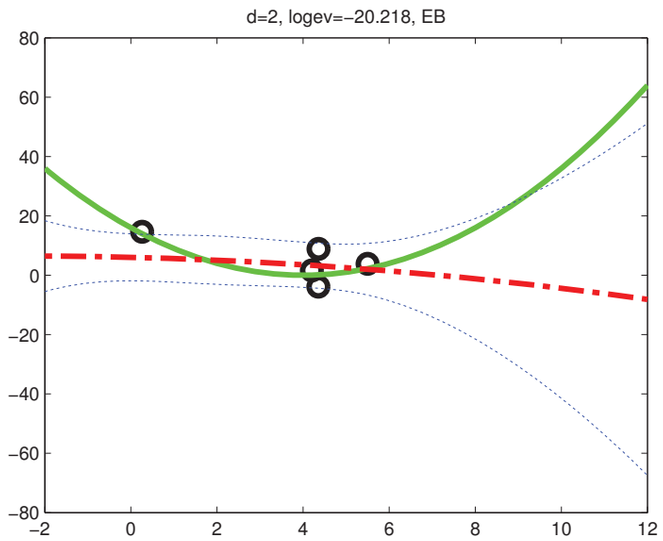
number of samples

PREDICTIVE
SCIENCE LABORATORY

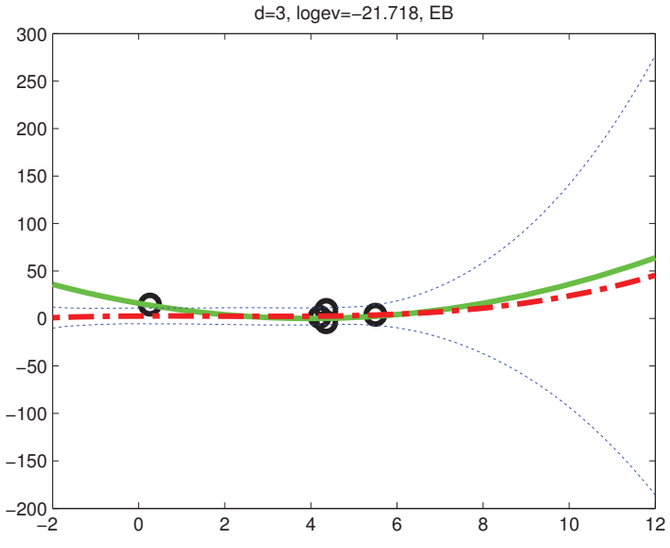
Example: Polynomial Regression



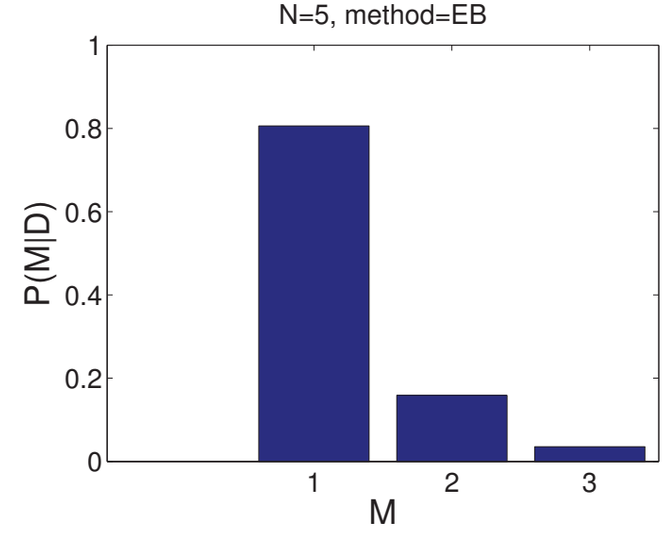
(a)



(b)



(c)

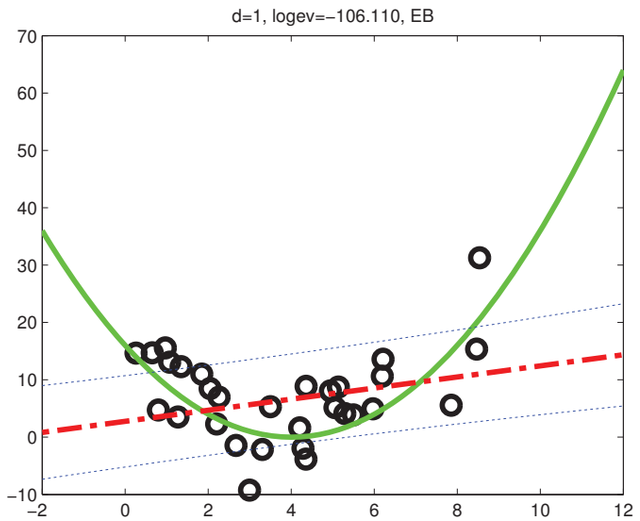


(d)

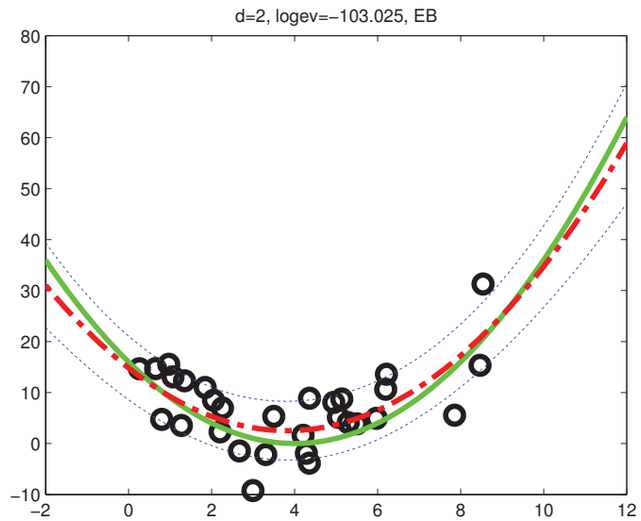
PREDICTIVE SCIENCE LABORATORY

Murphy 2012

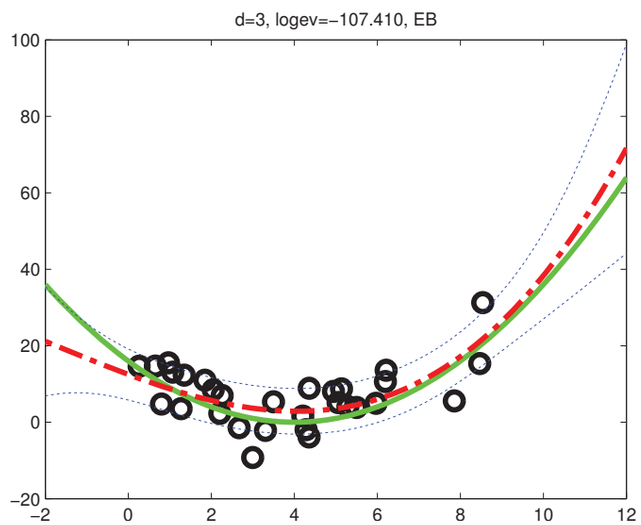
Example: Polynomial Regression



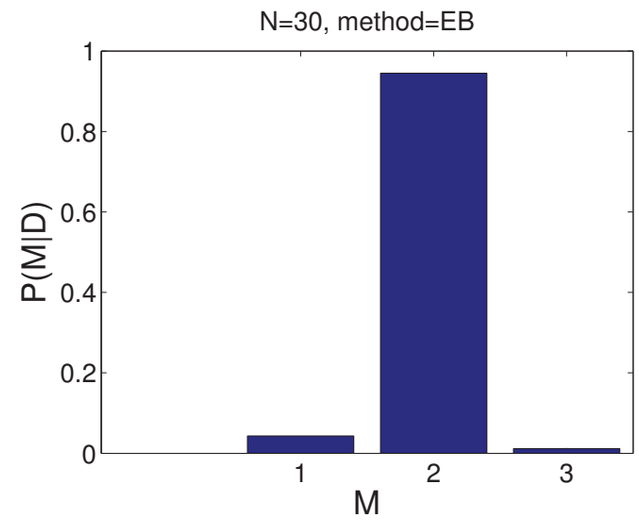
(a)



(b)



(c)



(d)

PREDICTIVE SCIENCE LABORATORY

Murphy 2012

Something more accurate? What about MCMC?

- MCMC (on its own) cannot be used to compute model evidence.
- There is an estimator called the **harmonic mean** estimator to the evidence, but it is garbage… (it should never be used).
- So, what do we do?

Sequential Monte Carlo Sampling Techniques

PREDICTIVE
SCIENCE LABORATORY

The problem

Sample from this:

$$p(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{Z}$$

Do it efficiently.

Compute the evidence (which can can be used for model selection):

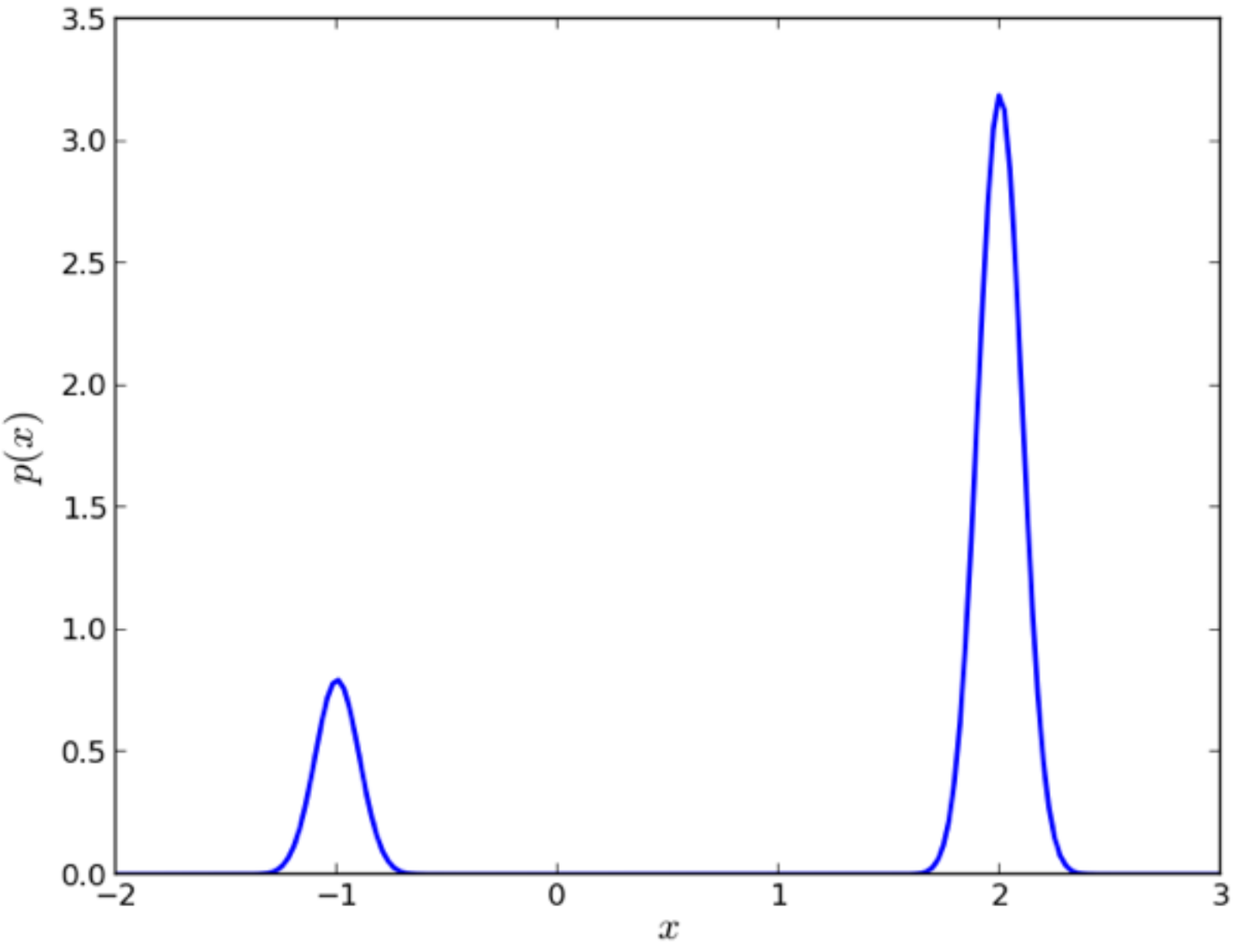
$$Z = \int p(D \mid \theta) p(\theta) d\theta$$

Why is it difficult?

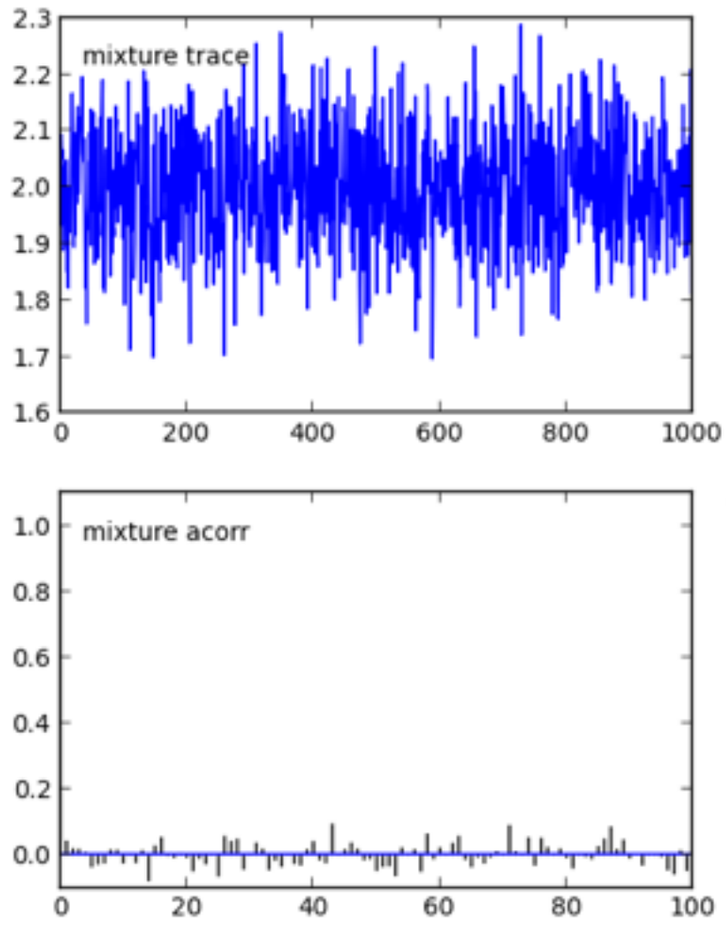
$$p(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{Z} \qquad Z = \int p(D \mid \theta) p(\theta) d\theta$$

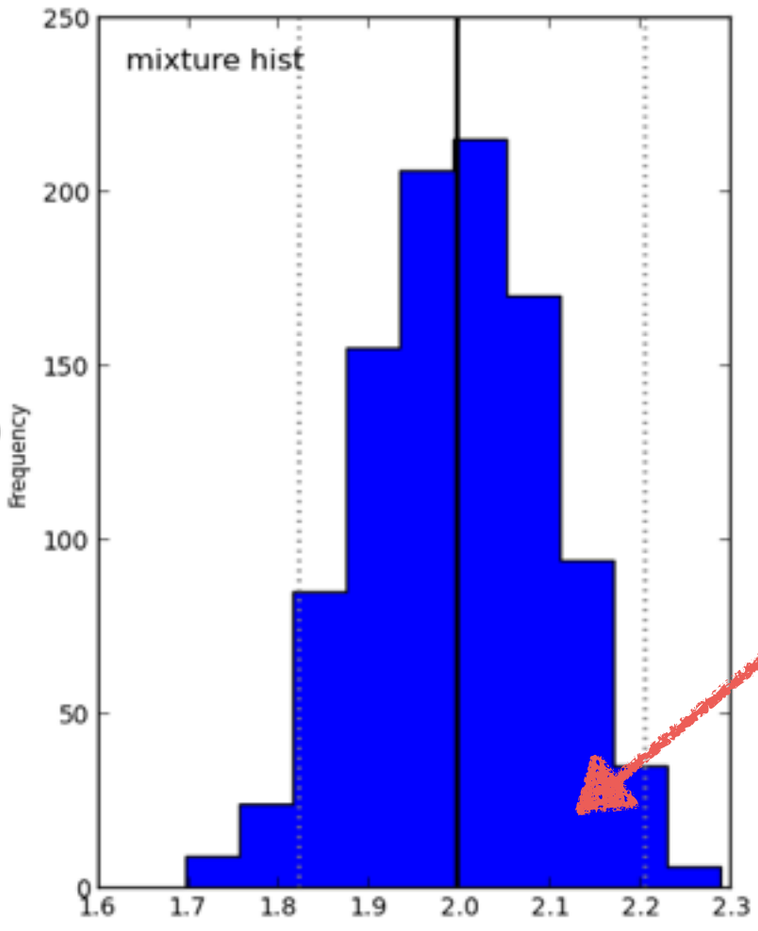
- Slow convergence (long thermalization).
- Trouble with multiple modes.
- Proposal must be hand-picked.
- The fact that the likelihood may be sharply picked makes the evaluation of the integral is extremely difficult.

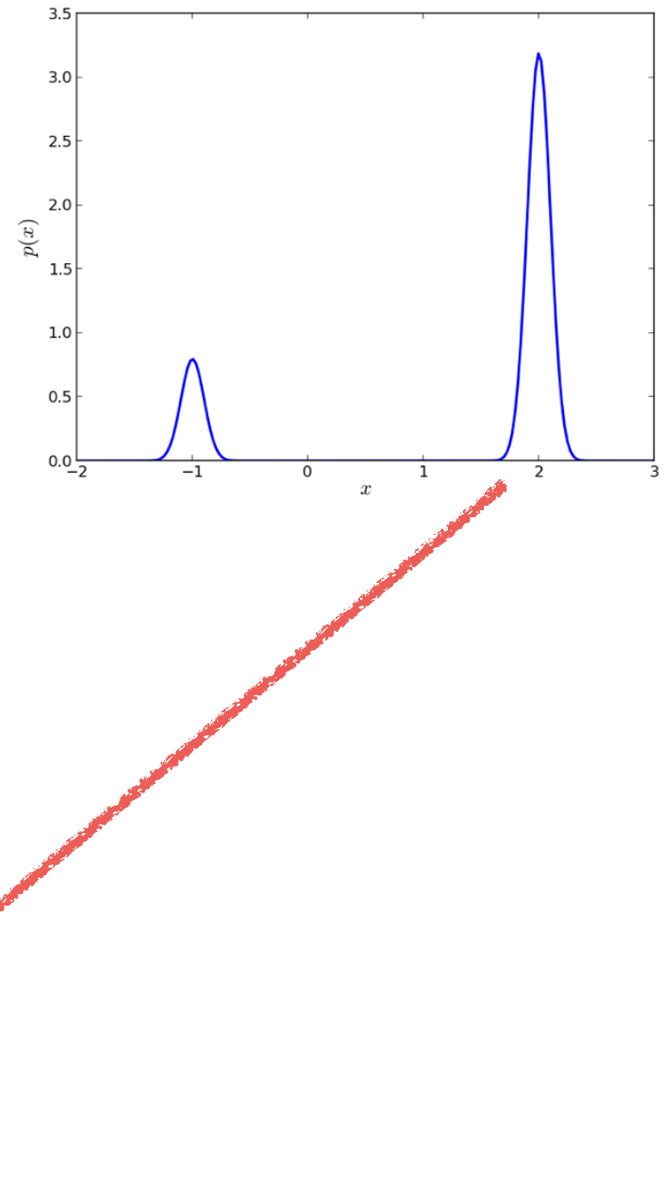
Example: Posterior with Two Modes



Example: Posterior with Two Modes

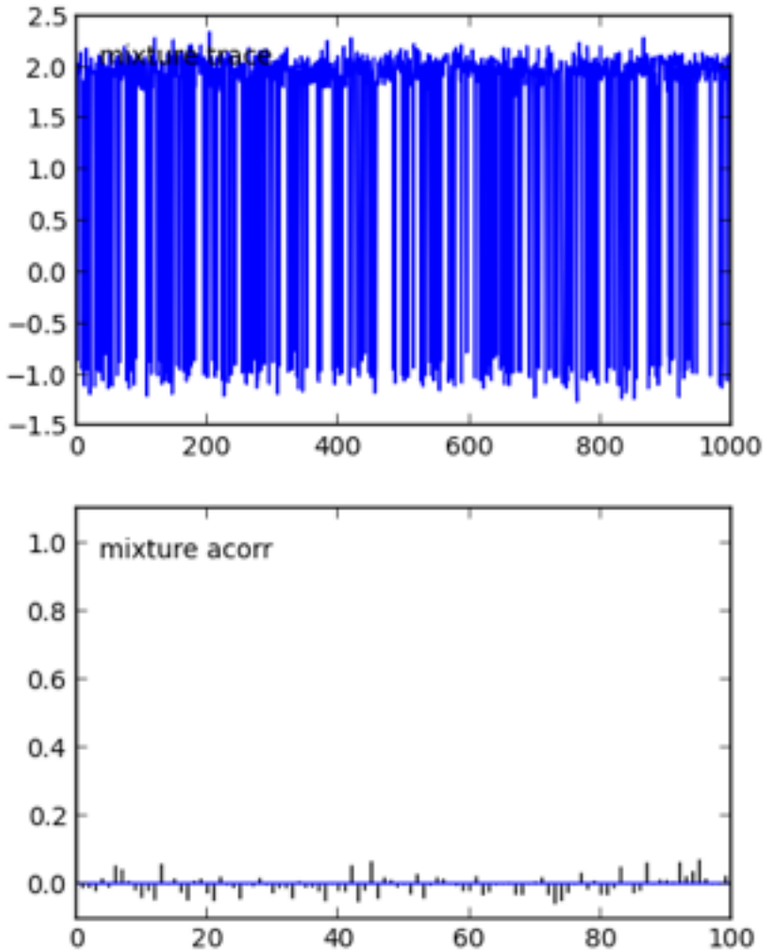


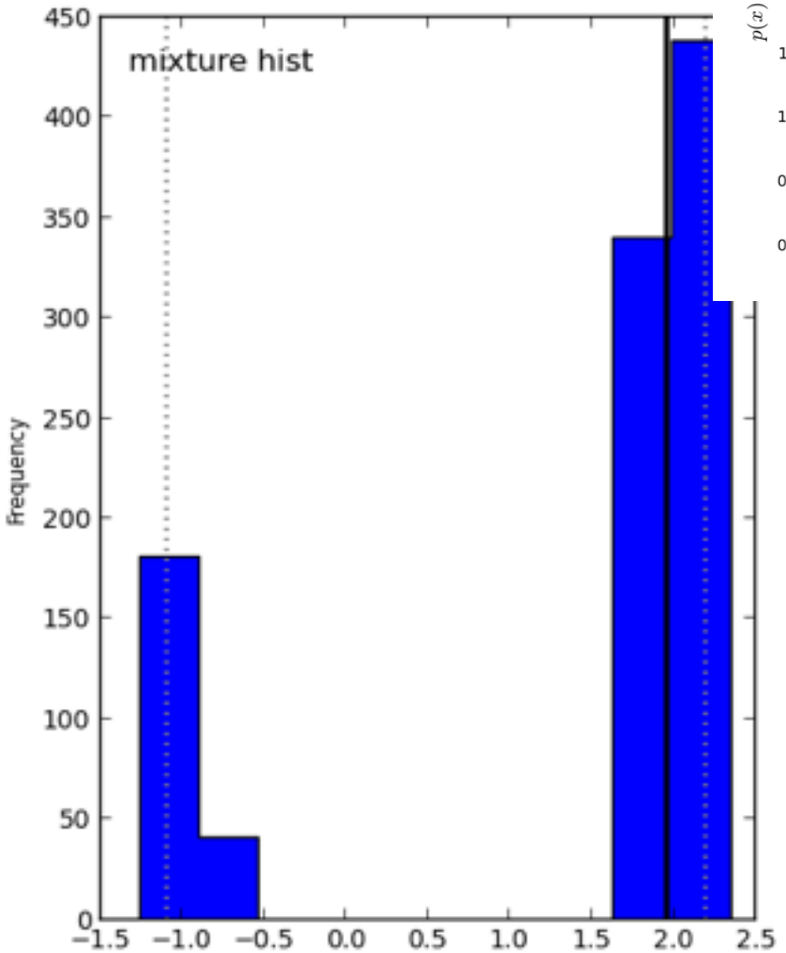


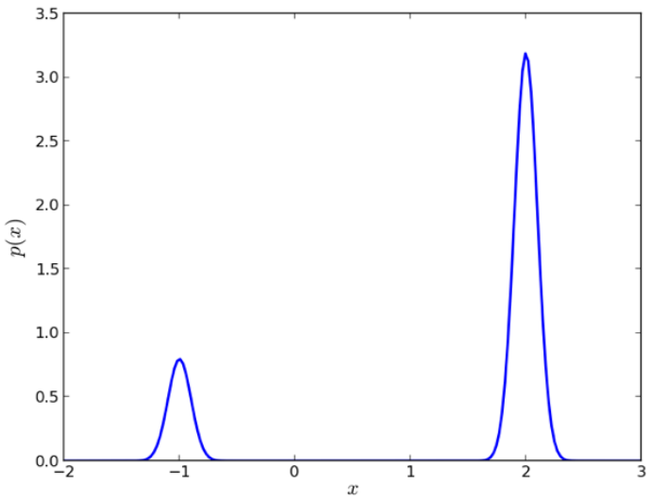


MCMC seems converged, but missed one mode…

Example: Posterior with Two Modes







MCMC works only if we hand-pick the right proposal…

Idea

Define a family of distributions that will take you from the prior to the posterior with increasing complexity:

$$p(\theta \mid D, \gamma) \propto \pi_\gamma(\theta) := p(D \mid \theta)^\gamma \, p(\theta)$$

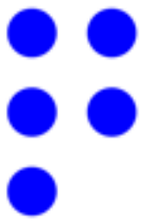
$$0 \leq \gamma \leq 1$$

From the prior:

$$p(\theta \mid D, \gamma = 0) \propto \pi_0(\theta) = p(\theta)$$

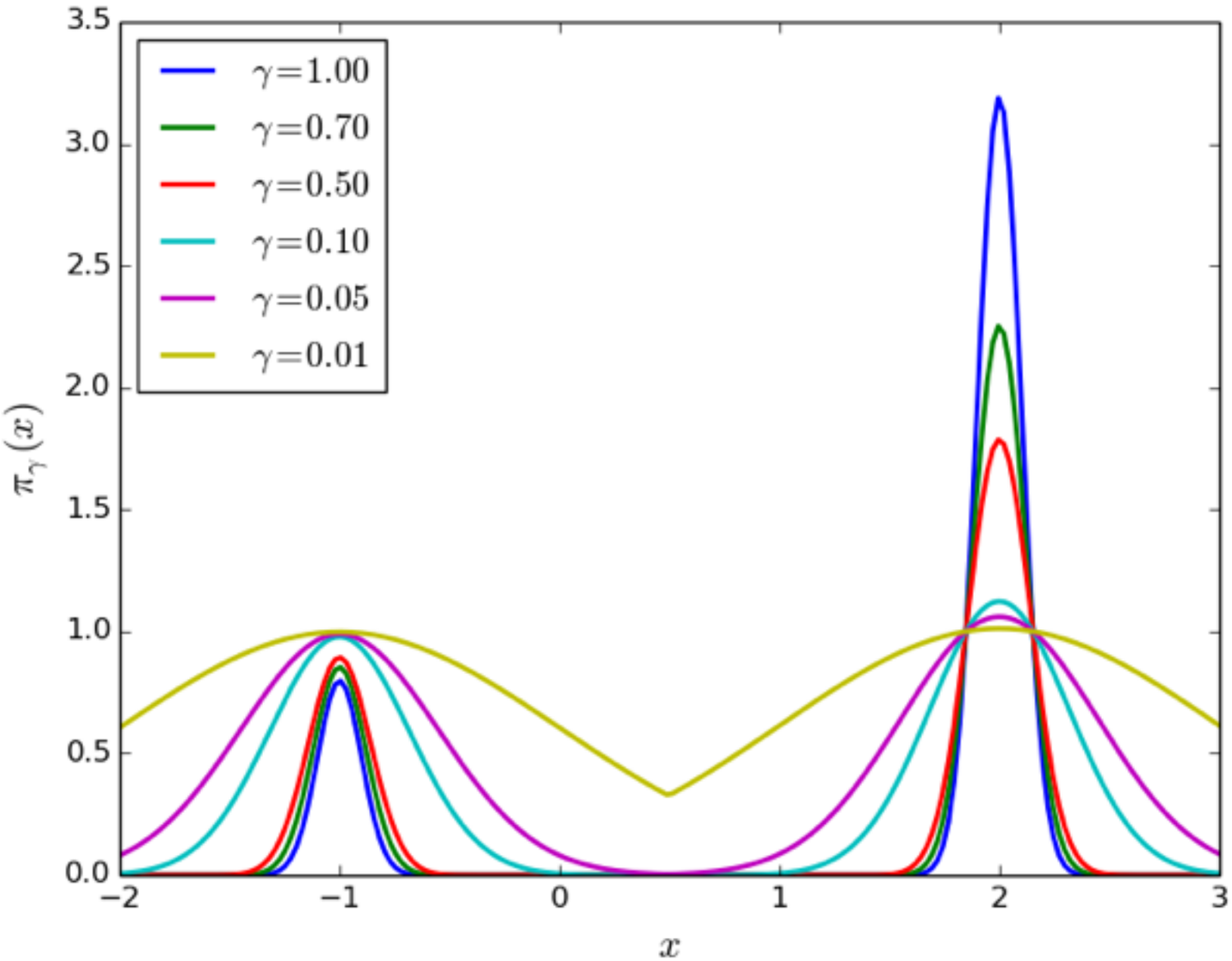
To the posterior:

$$p(\theta \mid D, \gamma = 1) = \pi_1(\theta) = p(D \mid \theta) p(\theta)$$



PREDICTIVE
SCIENCE LABORATORY

Example: Posterior with Two Modes



Idea

Start with a **particle approximation** to the prior:

$$p(\theta \mid D, \gamma = 0) = \sum_{i=1}^{N} w_{0,i} \delta(\theta_{0,i} - \theta)$$

$$\theta_{0,i} \sim p(\theta \mid D, \gamma = 0) = p(\theta) \qquad \sum_{i=1}^{N} w_{0,i} = 1$$

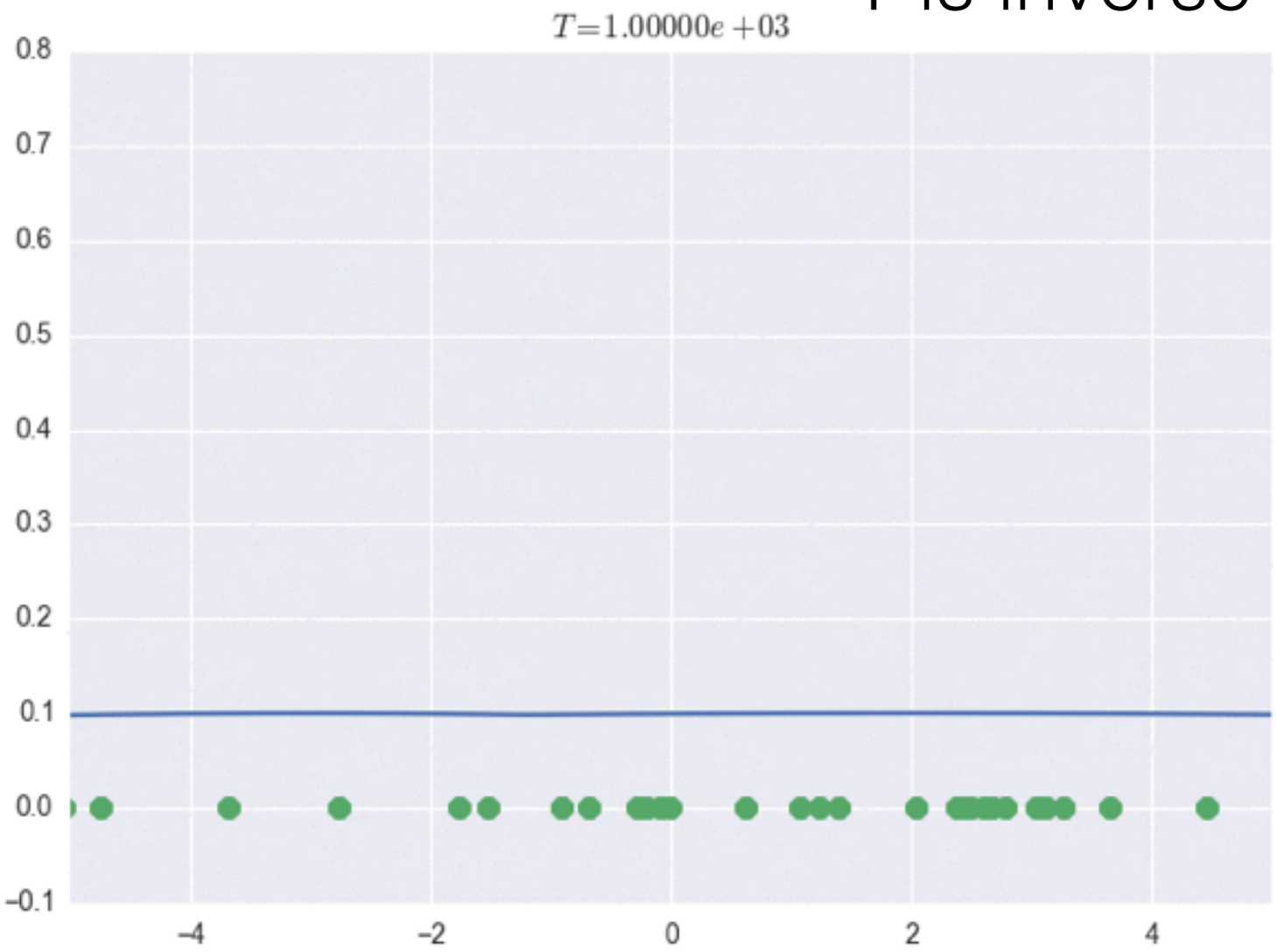
Gradually increase gamma and update particles and weights

$$p(\theta \mid D, \gamma) = \sum_{i=1}^{N} w_{\gamma,i} \delta(\theta_{\gamma,i} - \theta)$$

$$p(\theta \mid D, \gamma = 1) = \sum_{i=1}^{N} w_{1,i} \delta(\theta_{1,i} - \theta)$$

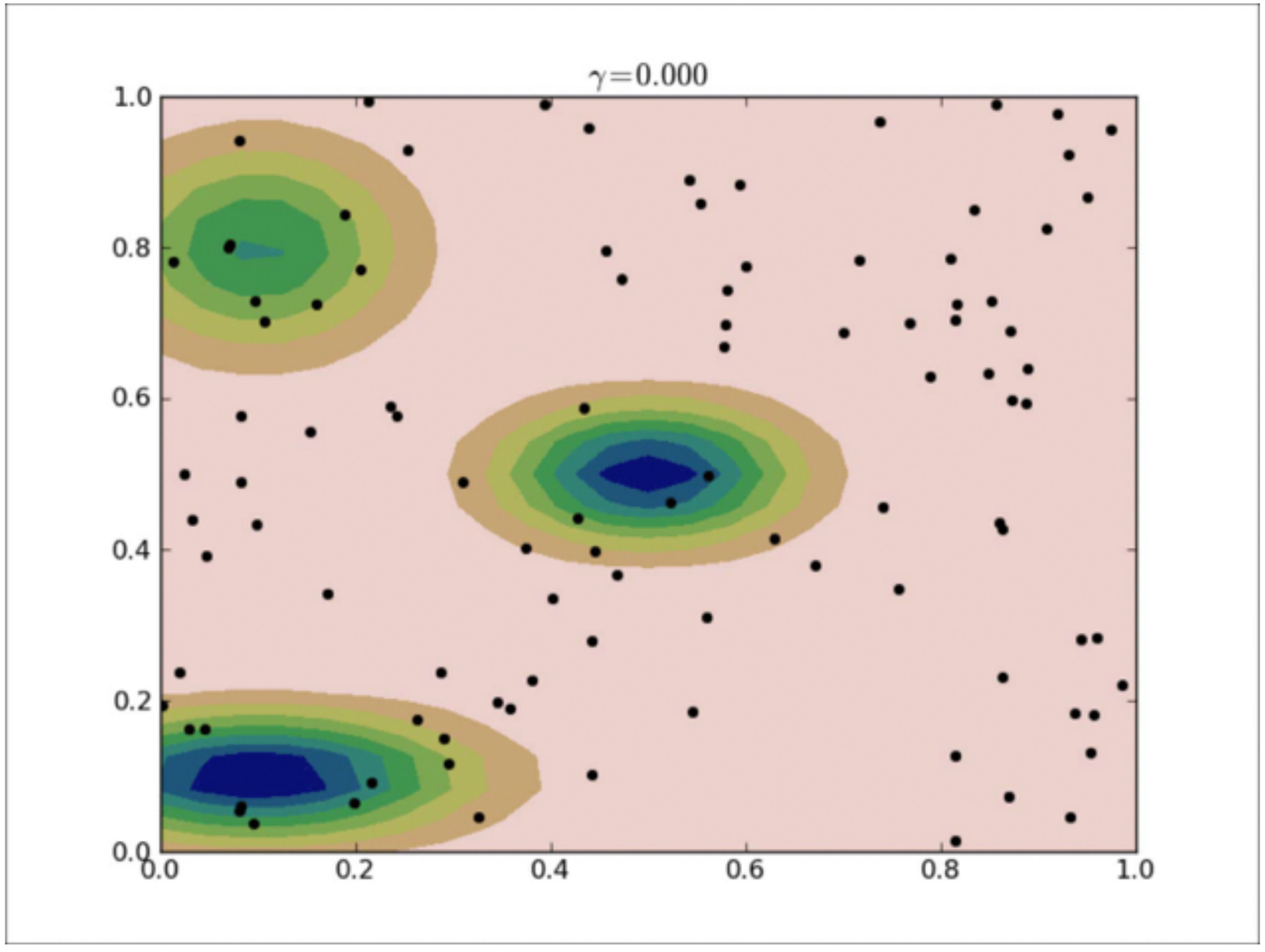
Example: Posterior with Two Modes

T is inverse gamma



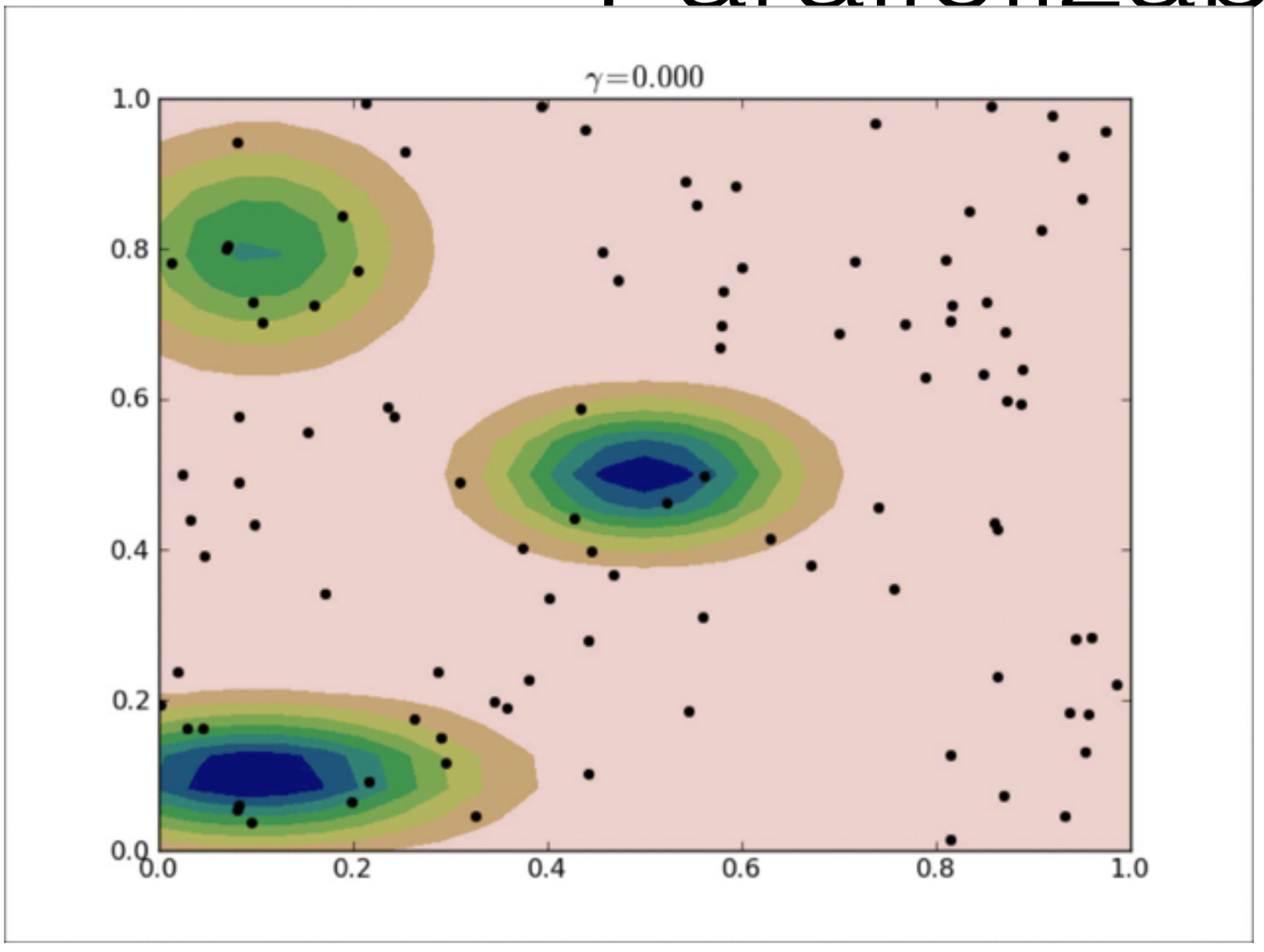
PREDICTIVE
SCIENCE LABORATORY

2D Three Modes



PREDICTIVE
SCIENCE LABORATORY

Embarrassingly Parallelizable



- Each particle can reside in a different CPU core.
- Minimal communication is required.

OK - How do we do it?

Consider two consecutive gammas:

$$0 \leq \gamma_t < \gamma_{t+1} \leq 1$$

We have a particle approximation for the first one:

$$p(\theta \mid D, \gamma_t) = \sum_{i=1}^{N} w_{\gamma_t,i} \delta(\theta_{\gamma_t,i} - \theta)$$

Build a particle approximation for the second one:

$$p(\theta \mid D, \gamma_{t+1}) = \sum_{i=1}^{N} w_{\gamma_{t+1},i} \delta(\theta_{\gamma_{t+1},i} - \theta)$$

It can be done in many ways, I will just give you one…

OK - How do we do it?

$$0 \le \gamma_t < \gamma_{t+1} \le 1$$

$$p(\theta \mid D, \gamma_t) = \sum_{i=1}^{N} w_{\gamma_t,i}\,\delta(\theta_{\gamma_t,i} - \theta) \quad \Rightarrow \quad p(\theta \mid D, \gamma_{t+1}) = \sum_{i=1}^{N} w_{\gamma_{t+1},i}\,\delta(\theta_{\gamma_{t+1},i} - \theta)$$

The new weights will be:

$$w_{\gamma_{t+1},i} = \frac{W_{\gamma_{t+1},i}}{\sum_{j=1}^{N} W_{\gamma_{t+1},j}} \quad \text{where} \quad W_{\gamma_{t+1}} = w_{\gamma_t,i}\,\frac{\pi_{\gamma_{t+1}}(\theta_{\gamma_t,i})}{\pi_{\gamma_t}(\theta_{\gamma_t,i})}$$

Sample the thetas from:

$$\theta_{\gamma_{t+1},i} \sim \pi_{\gamma_{t+1}}(\theta) = p(D \mid \theta)^{\gamma_{t+1}}\, p(\theta)$$

by doing a few steps of your favorite MCMC.

Where is the Normalization Constant?

$$0 \leq \gamma_t < \gamma_{t+1} \leq 1$$

$$p(\theta \mid D, \gamma_t) = \sum_{i=1}^{N} w_{\gamma_t,i} \delta(\theta_{\gamma_t,i} - \theta)$$

$$p(\theta \mid D, \gamma_{t+1}) = \sum_{i=1}^{N} w_{\gamma_{t+1},i} \delta(\theta_{\gamma_{t+1},i} - \theta)$$

$$w_{\gamma_{t+1},i} = \frac{W_{\gamma_{t+1},i}}{\sum_{j=1}^{N} W_{\gamma_{t+1},j}} \qquad W_{\gamma_{t+1}} = w_{\gamma_t,i} \frac{\pi_{\gamma_{t+1}}(\theta_{\gamma_t,i})}{\pi_{\gamma_t}(\theta_{\gamma_t,i})}$$

We can can approximate the ratio of two normalization constants by:

$$\frac{Z_{\gamma_{t+1}}}{Z_{\gamma_t}} \approx \sum_{i=1}^{N} W_{t,i} = \sum_{i=1}^{N} w_{\gamma_t,i} \frac{\pi_{\gamma_{t+1}}(\theta_{\gamma_t,i})}{\pi_{\gamma_t}(\theta_{\gamma_t,i})}$$

Where is the Normalization Constant?

We can can approximate the ratio of two normalization constants by:

$$\frac{Z_{\gamma_{t+1}}}{Z_{\gamma_t}} \approx \sum_{i=1}^{N} W_{t,i} = \sum_{i=1}^{N} w_{\gamma_t,i} \frac{\pi_{\gamma_{t+1}}(\theta_{\gamma_t,i})}{\pi_{\gamma_t}(\theta_{\gamma_t,i})}$$

Since, the normalization constant of the prior is just:

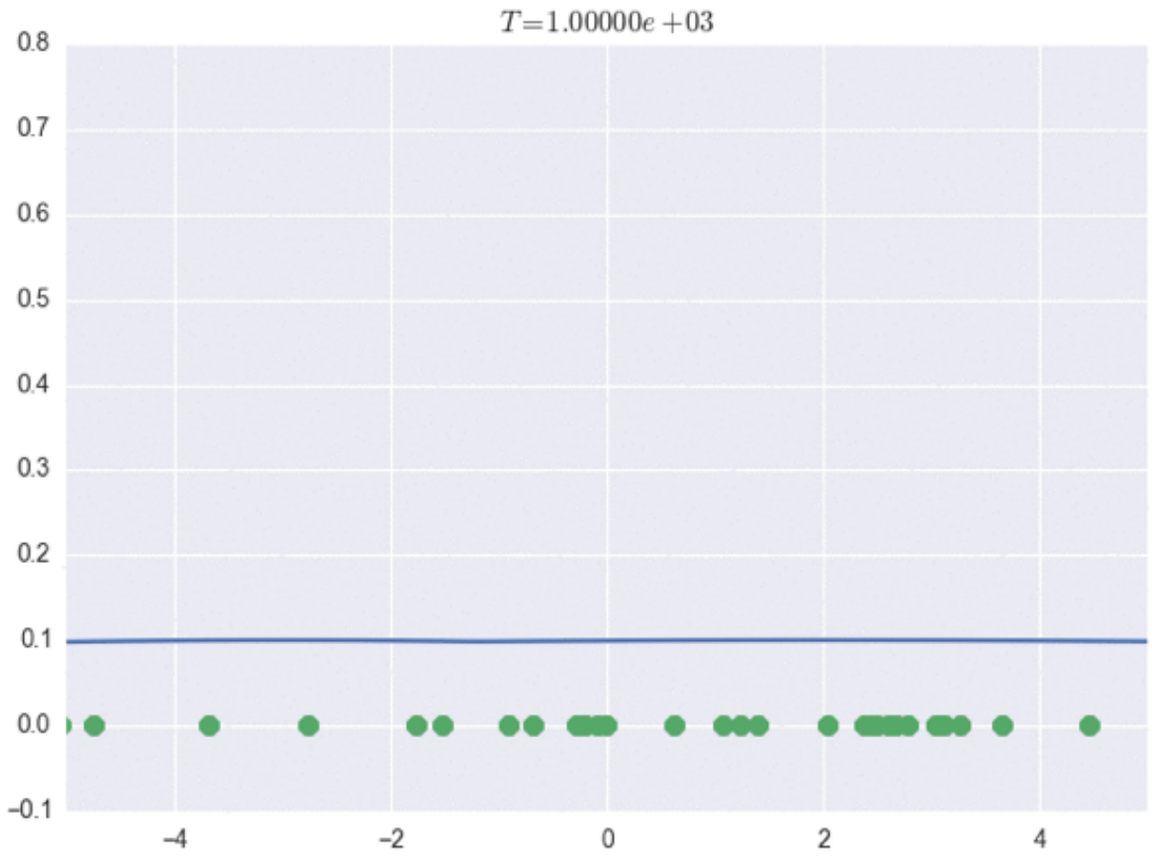
$$Z_{\gamma_0} = Z_0 = 1$$

we get by induction:

$$Z = Z_1 = \prod_{t=1}^{T} \frac{Z_{\gamma_{t+1}}}{Z_{\gamma_t}} \approx \prod_{t=1}^{N} \sum_{i=1}^{N} W_{t,i} = \prod_{t=1}^{N} \sum_{i=1}^{N} w_{\gamma_t,i} \frac{\pi_{\gamma_{t+1}}(\theta_{\gamma_t,i})}{\pi_{\gamma_t}(\theta_{\gamma_t,i})}$$

Resampling

- Some of the particles may be trapped in regions of low probability.
- Their weight will become extremely small.
- To avoid wasting computations, we should eliminate them.



The Effective Sample Size

The effective sample size (ESS):

$$\mathrm{ESS}(\gamma) = \frac{1}{\sum_{i=1}^{N} w_{\gamma,i}^{\,2}}$$

measures the degeneracy of the particle approximation.

Notice:

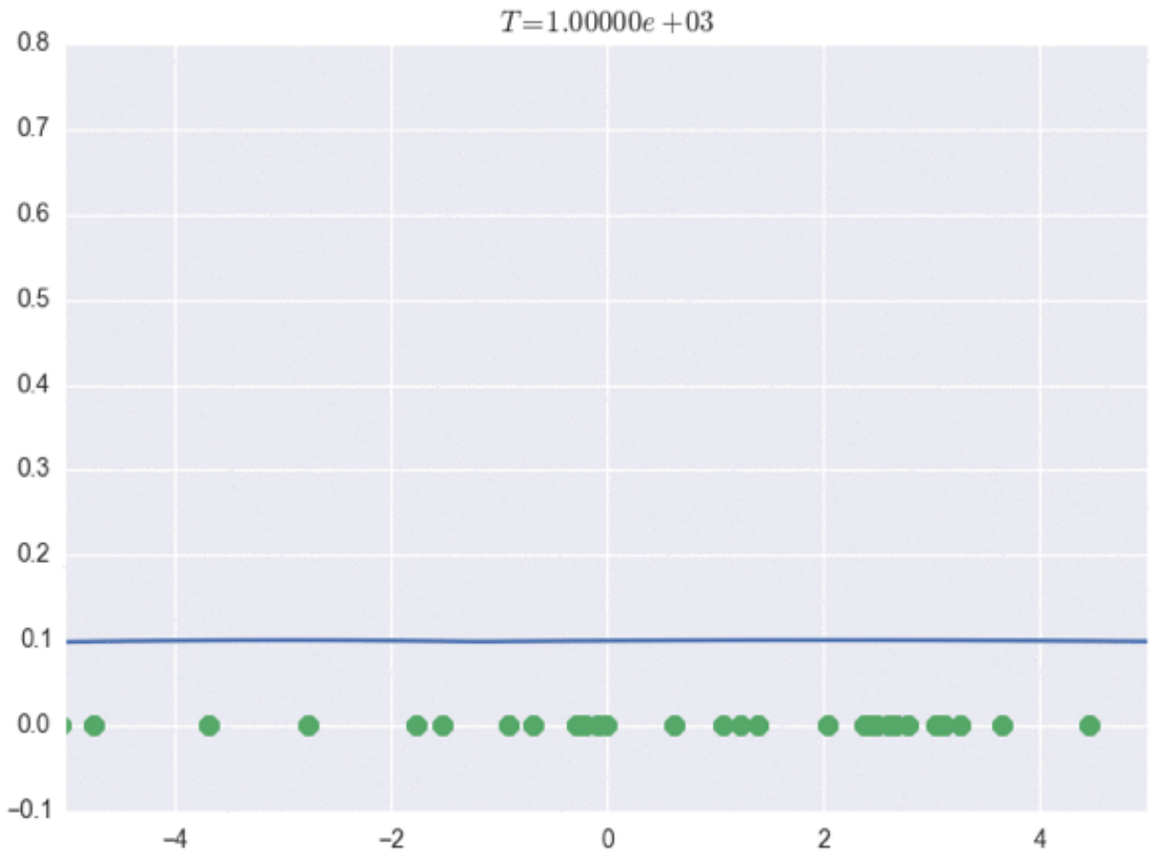
$$w_{\gamma,i} = \frac{1}{N} \quad \Rightarrow \quad \mathrm{ESS}(\gamma) = N$$

$$w_{\gamma,1} = 1, \text{ and } w_{\gamma,i} = 0, \text{ for } i = 2,\ldots,N \quad \Rightarrow \quad \mathrm{ESS}(\gamma) = 1$$

Resampling

We resample every time:

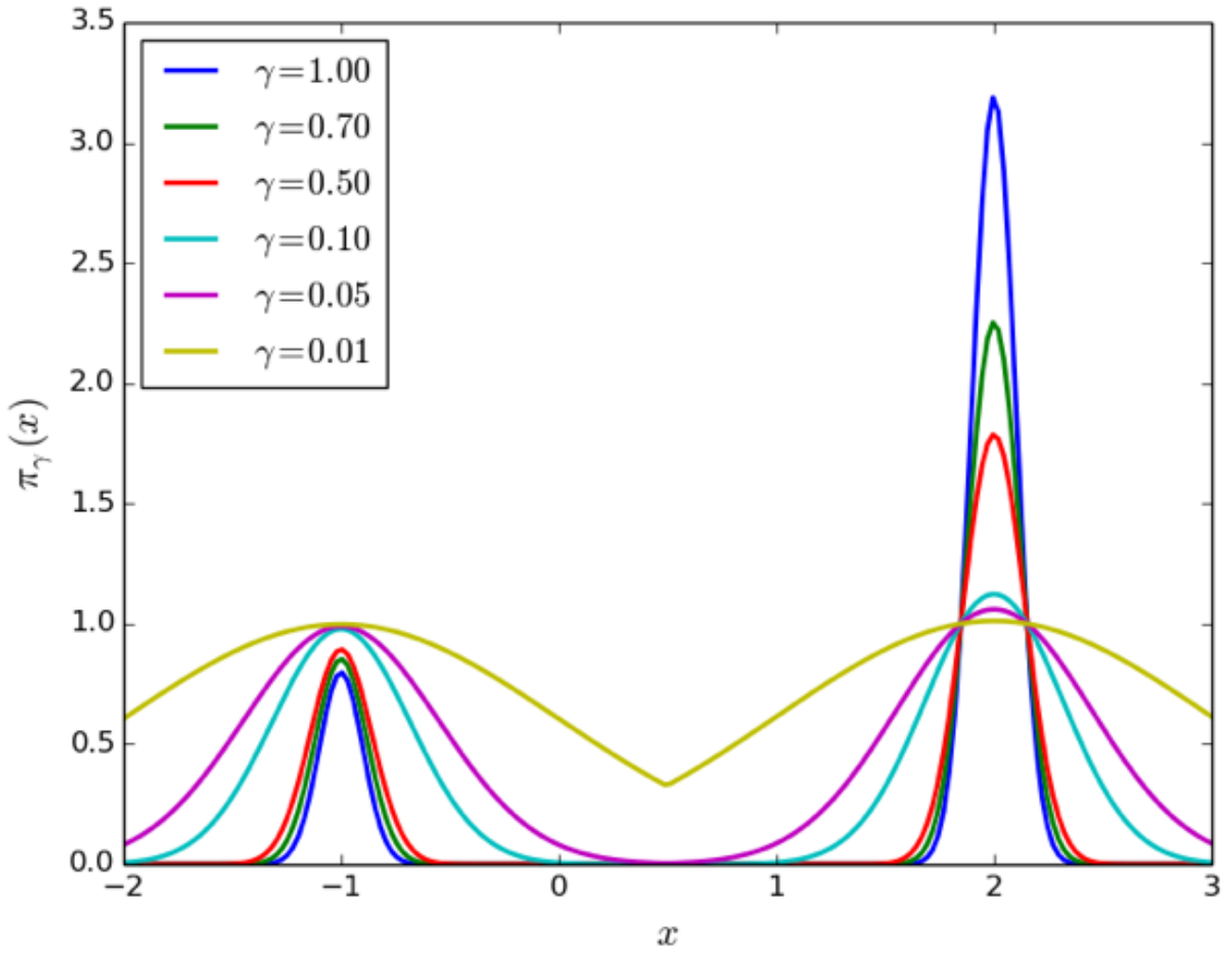
$$\mathrm{ESS}(\gamma_t) \leq 0.5N$$



PREDICTIVE
SCIENCE LABORATORY

Selecting the Intermediate Gammas

You can just pick them by hand:



PREDICTIVE
SCIENCE LABORATORY

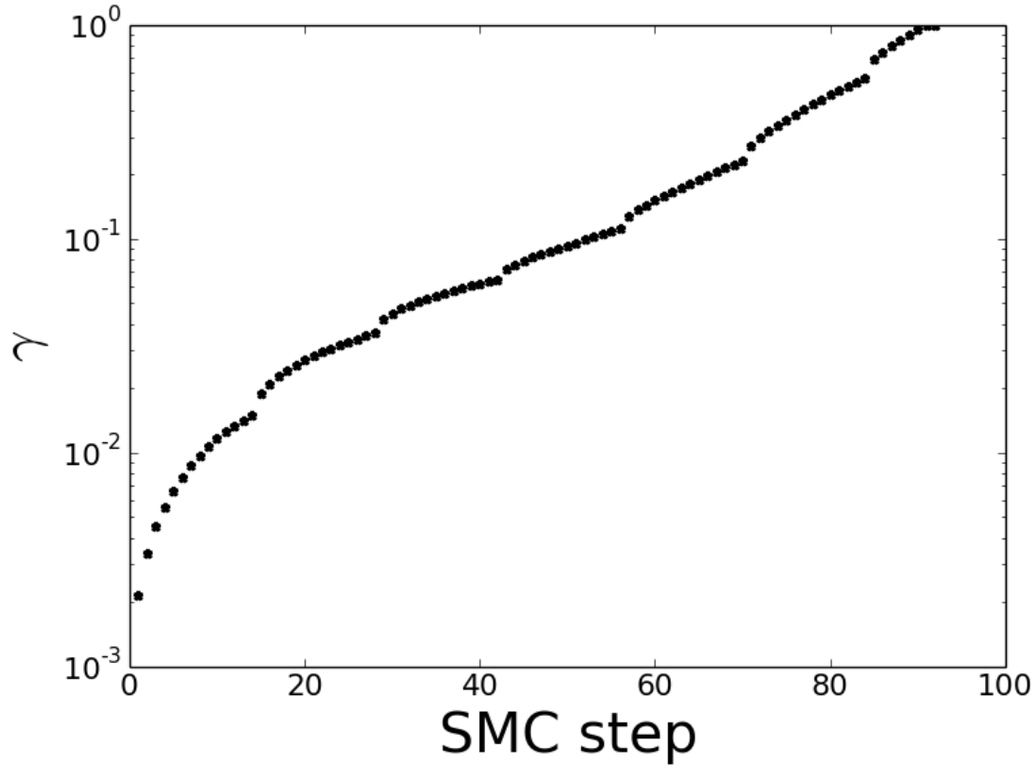
Selecting the Intermediate Gammas

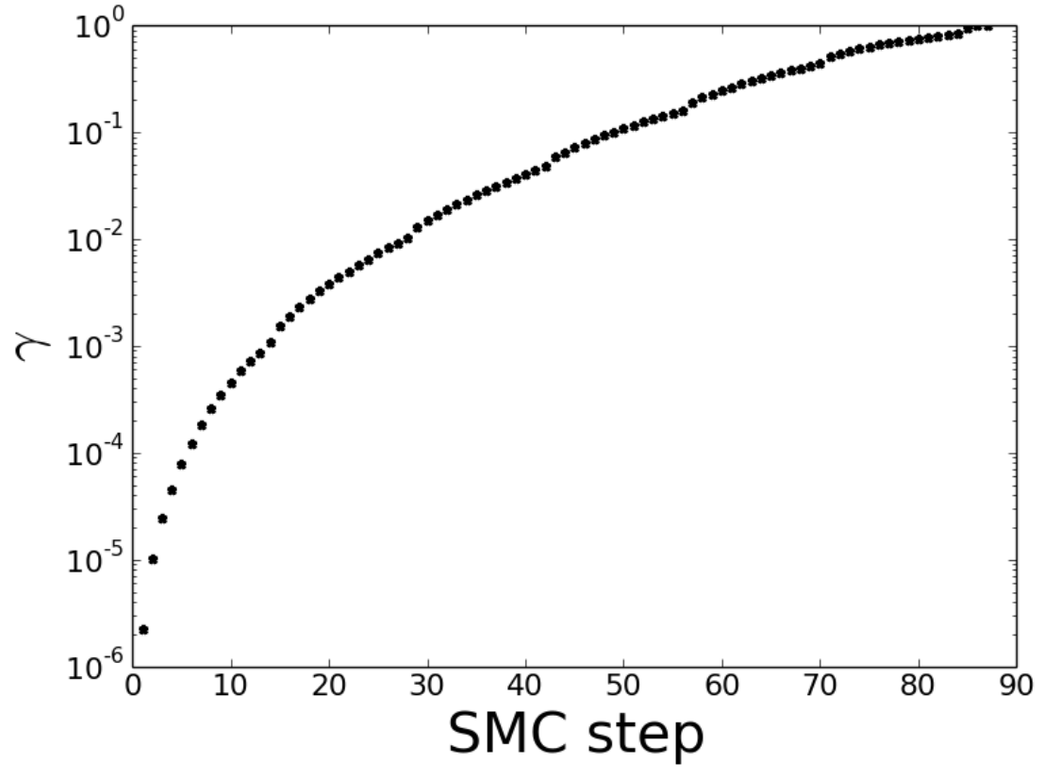
Or you can pick them so that the particles do not become too degenerate from step to step:

$$\mathrm{ESS}(\gamma_{t+1}) = 0.95\mathrm{ESS}(\gamma_t)$$

Example (Calibration 10D)

Automatic selection of inverse temperature schedule:





Bilionis, I., et al. (2015). "Crop physiology calibration in the CLM." Geoscientific Model Development **8**(4): 1071-1083.

PREDICTIVE
SCIENCE LABORATORY

The algorithm

1. Build initial particle approximation (sample the prior).
2. Find the next gamma by fixing the change in the ESS.
3. Resample if ESS falls below threshold.
4. Draw samples at new gamma starting at the old particles.

PREDICTIVE SCIENCE LABORATORY

Bilionis, I. and P. S. Koutsourelakis (2012). "Free energy computations by minimization of Kullback-Leibler divergence: An efficient adaptive biasing potential method for sparse representations." Journal of Computational Physics **231**(9): 3849-3870.

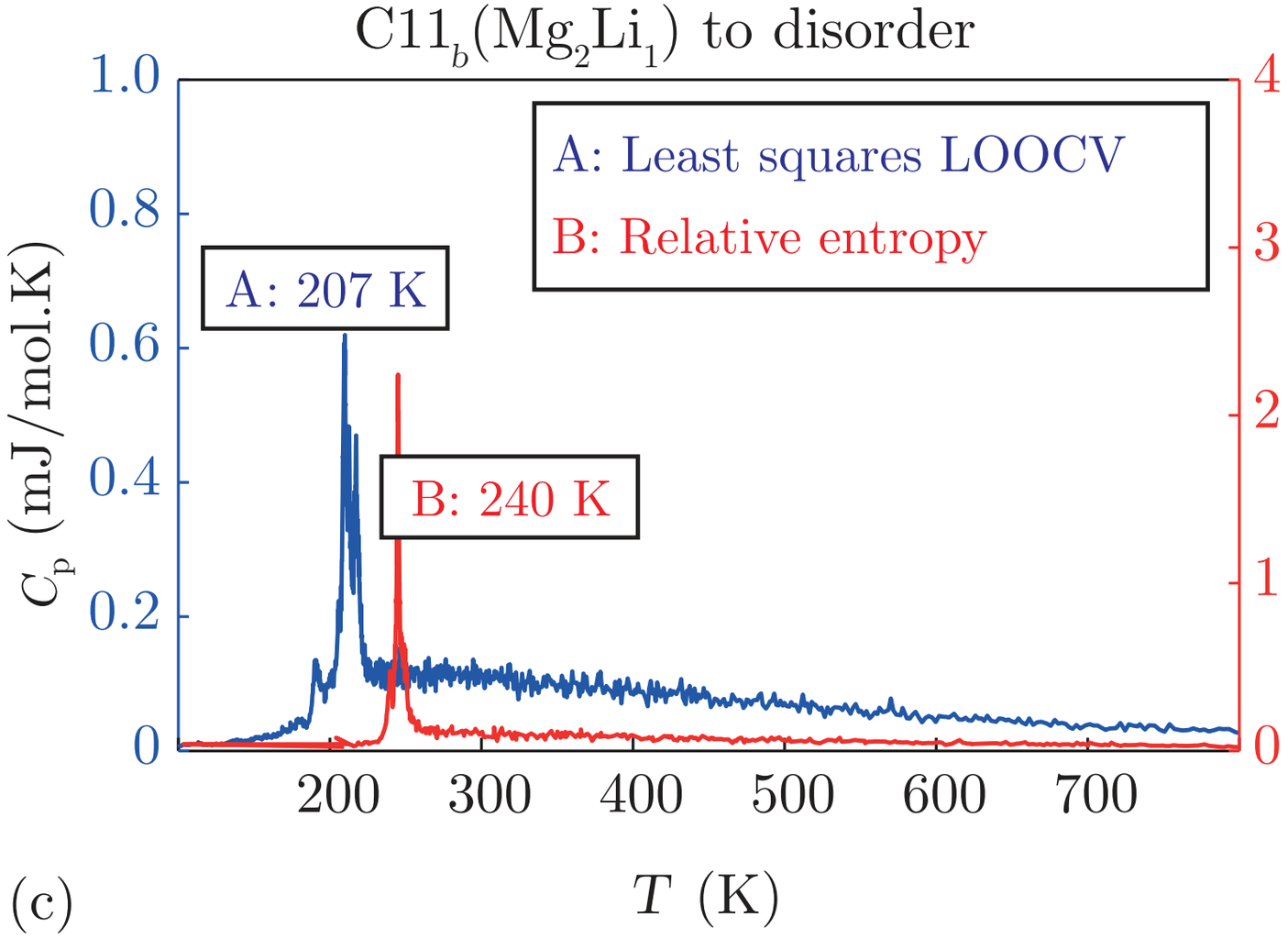
Why bother?

- Very easy to implement
- Embarrassingly parallelizable
- Makes use of existing sampling code
- Captures multiple modes
- Don’t have to fix gamma schedule.
- Can adjust parameters of stochastic samplers.

Example (Calibration 16KD)

Scanning the temperature to find where transition occurs in binary alloys (x 10K faster than thermodynamic integration)

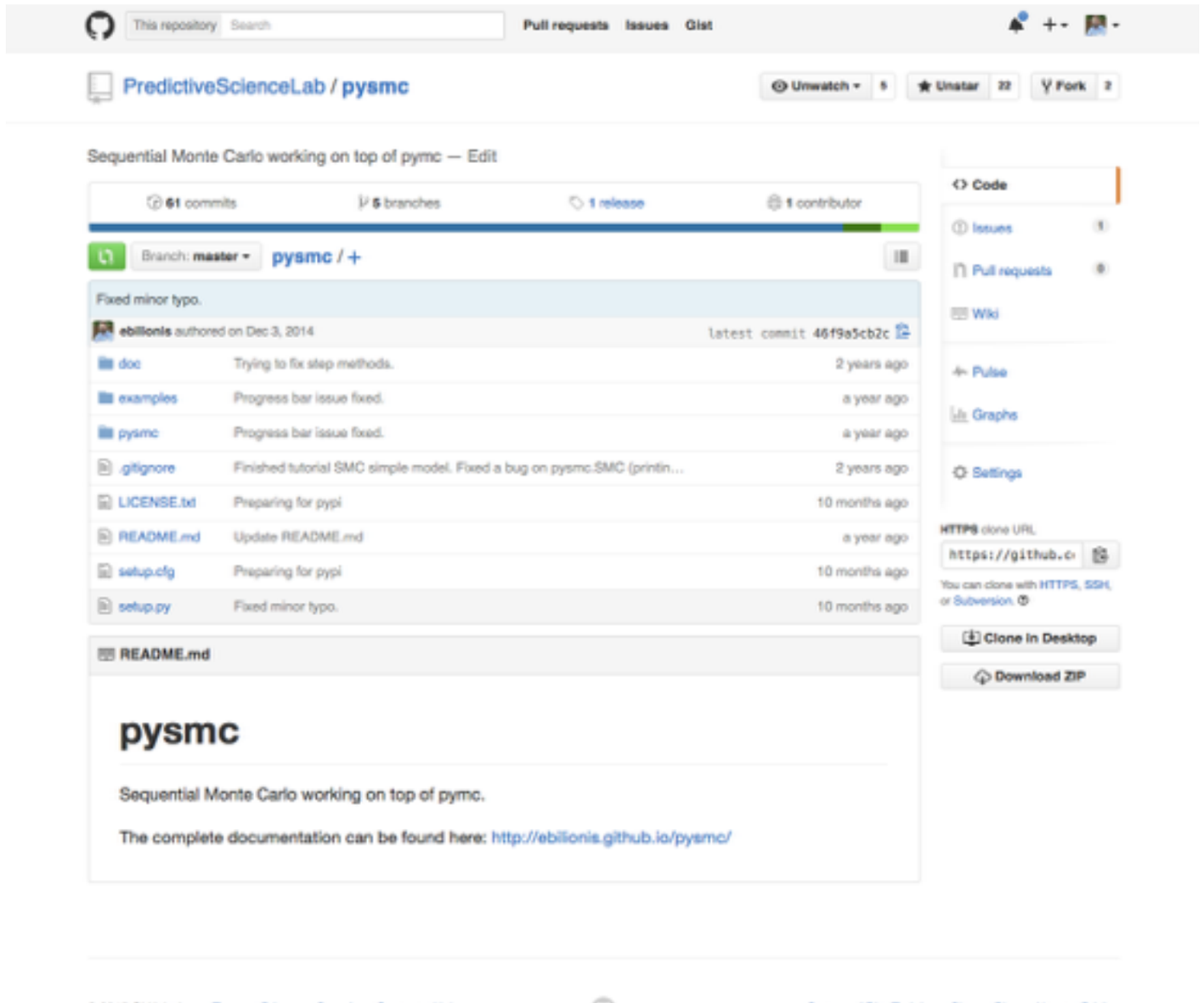
Kristensen, J., et al. (2013). "Relative entropy as model selection tool in cluster expansions." Physical Review B **87**(17).



(c)

PREDICTIVE SCIENCE LABORATORY

PYSMC CODE



PREDICTIVE
SCIENCE LABORATORY