Fitting a model to data

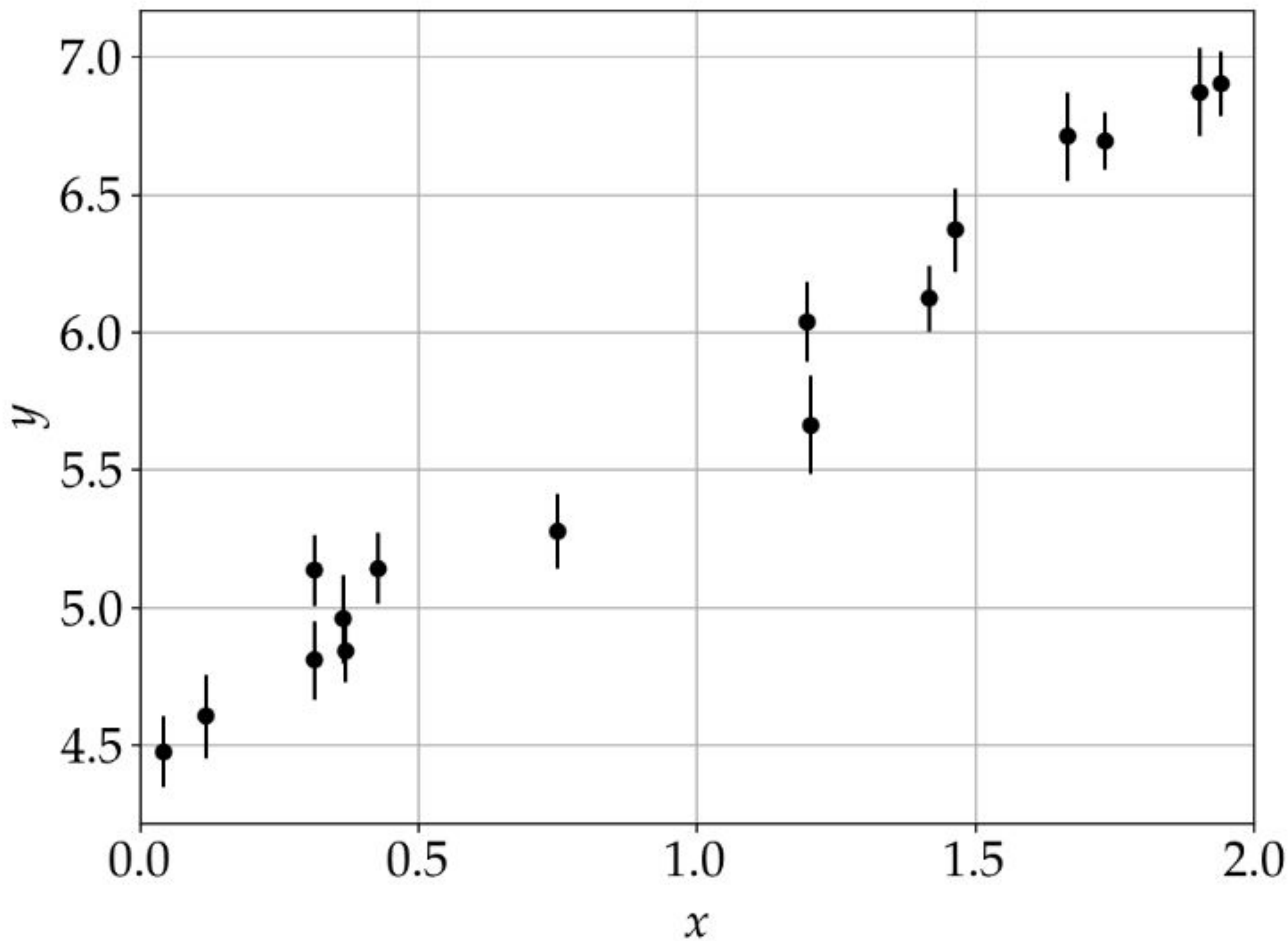
(Based on Hogg, Bovy & Lang (2010))

Fran Bartolić

fbartolic

Outline

1. Standard practice
2. The likelihood function
3. Bayes’ formula
4. Metropolis-Hastings MCMC
5. Where to go from here


7.0
6.5
6.0
5.5
5.0
4.5
y
0.0
0.5
1.0
1.5
2.0
x

Assumptions

- N independent data points $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$
- uncertainties $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$
- perfectly measured $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ values
- we have reason to believe that the data was generated from a straight line $y = mx + b$

Standard practice

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{\Phi} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{bmatrix}, \mathbf{C} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X}$$

Standard practice

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \mathbf{\Phi} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{bmatrix}, \mathbf{C} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{\Phi X} \rightarrow \mathbf{X} = \left[\mathbf{\Phi}^\tau \mathbf{C}^{-1} \mathbf{\Phi}\right]^{-1} \left[\mathbf{\Phi}^\tau \mathbf{C}^{-1} \mathbf{Y}\right]$$

Standard practice

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_N \end{bmatrix}, \mathbf{C} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} b \\ m \end{bmatrix}$$

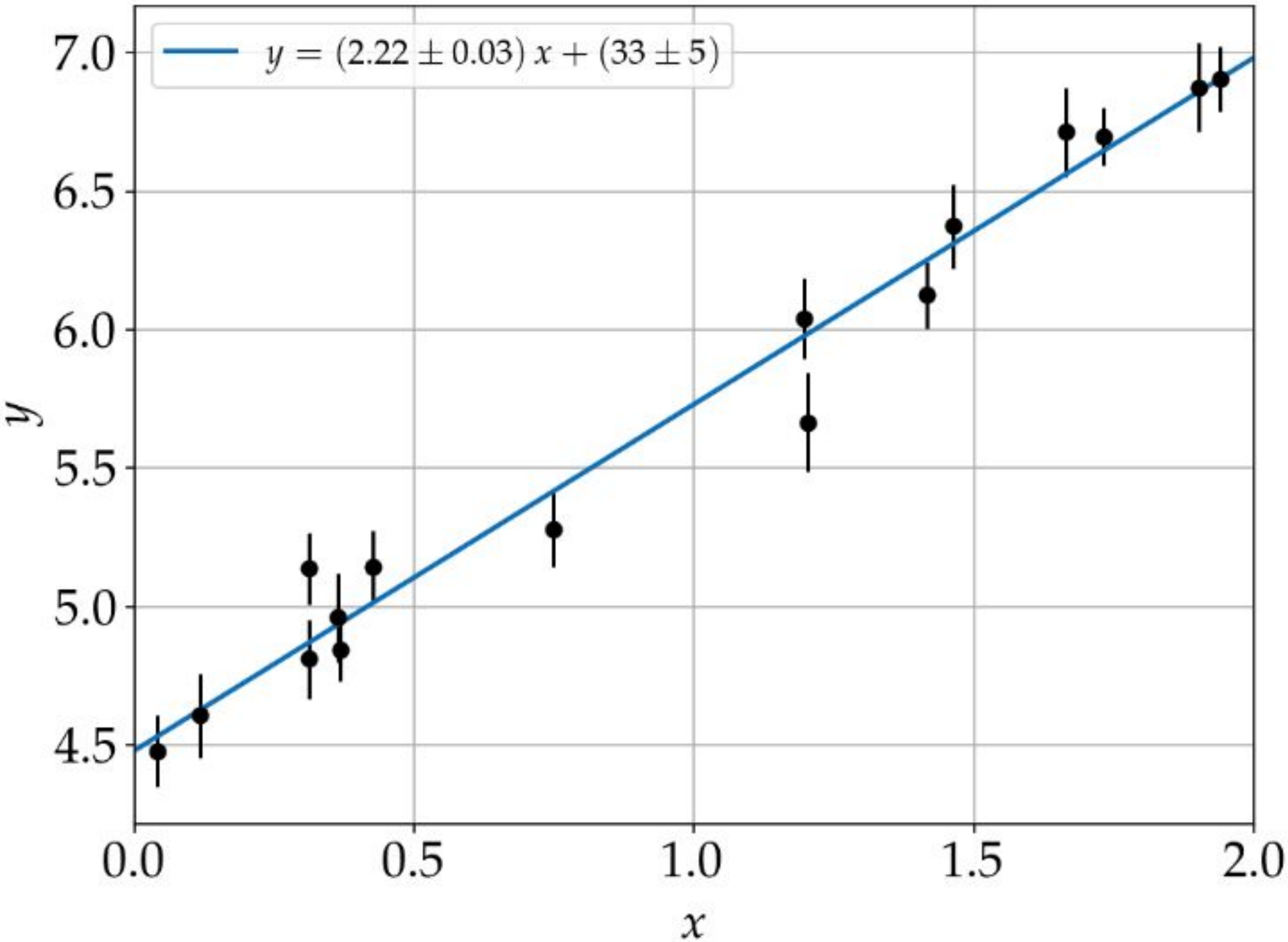
$$\mathbf{Y} = \boldsymbol{\Phi}\mathbf{X} \rightarrow \mathbf{X} = \left[\boldsymbol{\Phi}^\tau \mathbf{C}^{-1} \boldsymbol{\Phi}\right]^{-1} \left[\boldsymbol{\Phi}^\tau \mathbf{C}^{-1} \mathbf{Y}\right]$$

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_i^2} = [\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}]^\tau \mathbf{C}^{-1} [\mathbf{Y} - \boldsymbol{\Phi}\mathbf{X}]$$

Chi-squared

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_i^2}$$

- Dimensionless measure of distance between the model and the data
- Convex, there is a unique global minimum if assumptions hold


$y = (2.22 \pm 0.03)\, x + (33 \pm 5)$
7.0
6.5
6.0
5.5
5.0
4.5
y
0.0
0.5
1.0
1.5
2.0
x

Bayes' theorem

$$\underbrace{\mathrm{P}(\mathbf{\Theta}|\mathbf{D},\mathrm{M})}_{\text{}}\overset{\text{Posterior}}{} = \frac{\overset{\text{Likelihood}}{\mathrm{P}\left(\mathbf{D}|\mathbf{\Theta},\mathrm{M}\right)}\,\overset{\text{Prior}}{\mathrm{P}\left(\mathbf{\Theta}|\mathrm{M}\right)}}{\underset{\text{Evidence}}{\mathrm{P}\left(\mathbf{D}|\mathrm{M}\right)}}$$

$$\mathbf{\Theta} = \{m, b\} \quad \mathbf{D} = \{\mathbf{x}, \mathbf{y}, \boldsymbol{\sigma}\}$$

The Likelihood Function

$$p(y_i|x_i, \sigma_i, m, b) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{[y_i - mx_i - b]^2}{2\sigma_i^2}\right)$$

The Likelihood Function

$$p(y_i|x_i, \sigma_i, m, b) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{[y_i - mx_i - b]^2}{2\sigma_i^2}\right)$$

$$\mathcal{L} = \prod_{i=1}^{N} p(y_i|x_i, \sigma_i, m, b)$$

The Likelihood Function

$$p(y_i|x_i, \sigma_i, m, b) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{[y_i - mx_i - b]^2}{2\sigma_i^2} \right)$$

$$\mathcal{L} = \prod_{i=1}^{N} p(y_i|x_i, \sigma_i, m, b)$$

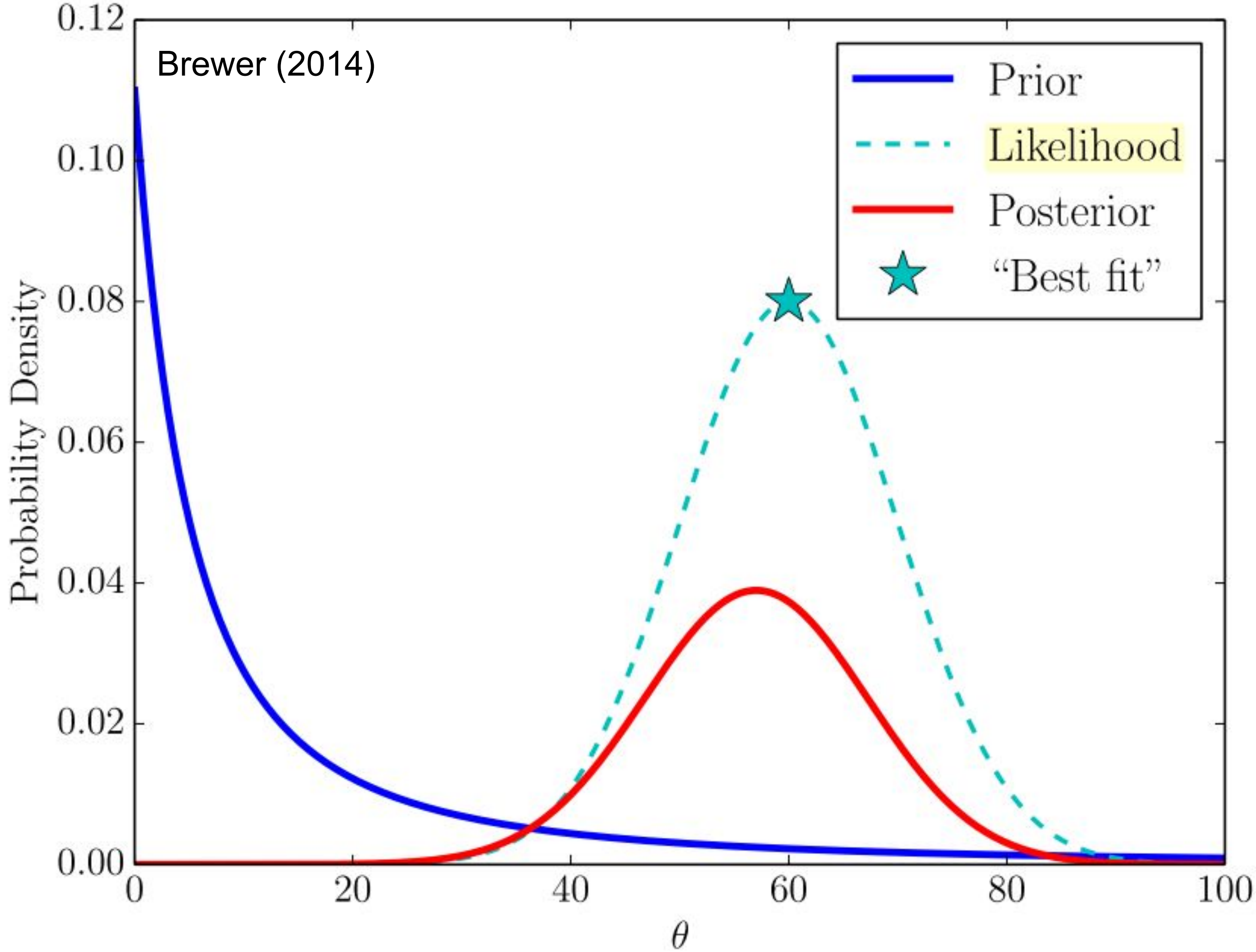
$$\ln \mathcal{L} = \text{const.} - \sum_{i=1}^{N} \frac{[y_i - mx_i - b]^2}{2\sigma_i^2} = \text{const.} - \frac{1}{2}\chi^2$$

Minimizing chi-squared = maximizing likelihood

Priors

- Depend only on the model parameters
- Represent prior knowledge about parameters, *before* we've seen any data
- No such thing as "uninformative prior", at best there is a "weakly informative prior"

Posterior



Brewer (2014)
Prior
Likelihood
Posterior
"Best fit"
Probability Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
θ

How to get samples from posteriors?

1. In all but the most trivial cases the posterior cannot be calculated analytically
2. We would like to have a way of obtaining samples from the posterior pdf

Monte Carlo Integration

$$p(\theta) \geq 0 \quad \text{for all } \theta, \qquad \int p(\theta)\mathrm{d}\theta = 1$$

$$\mathrm{E}_{p(\theta)}\left[\theta\right] \equiv \int \theta\, p(\theta)\mathrm{d}\theta$$

Monte Carlo Integration

$$p(\theta) \geq 0 \quad \text{for all } \theta, \qquad \int p(\theta)\mathrm{d}\theta = 1$$

$$\mathrm{E}_{p(\theta)}\left[\theta\right] \equiv \int \theta\, p(\theta)\mathrm{d}\theta$$

$$\mathrm{E}_{p(\theta)}\left[g(\theta)\right] \equiv \int g(\theta)\, p(\theta)\mathrm{d}\theta$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} g(\theta_k)$$

Why sample?

- Because it enables calculation of integrals
- Integrals are involved in calculating means, medians and quantiles but not the mode!
- Sampling also enables trivial marginalization of nuisance parameters

Metropolis-Hastings Markov Chain Monte Carlo

Given a function $f(\theta)$ and a sample θ_k of parameter θ:

1. Draw proposal θ' from the proposal pdf $q(\theta'|\theta_k)$ which we know how to sample from

2. Draw a uniform random number $r \sim \mathcal{U}(0,1)$

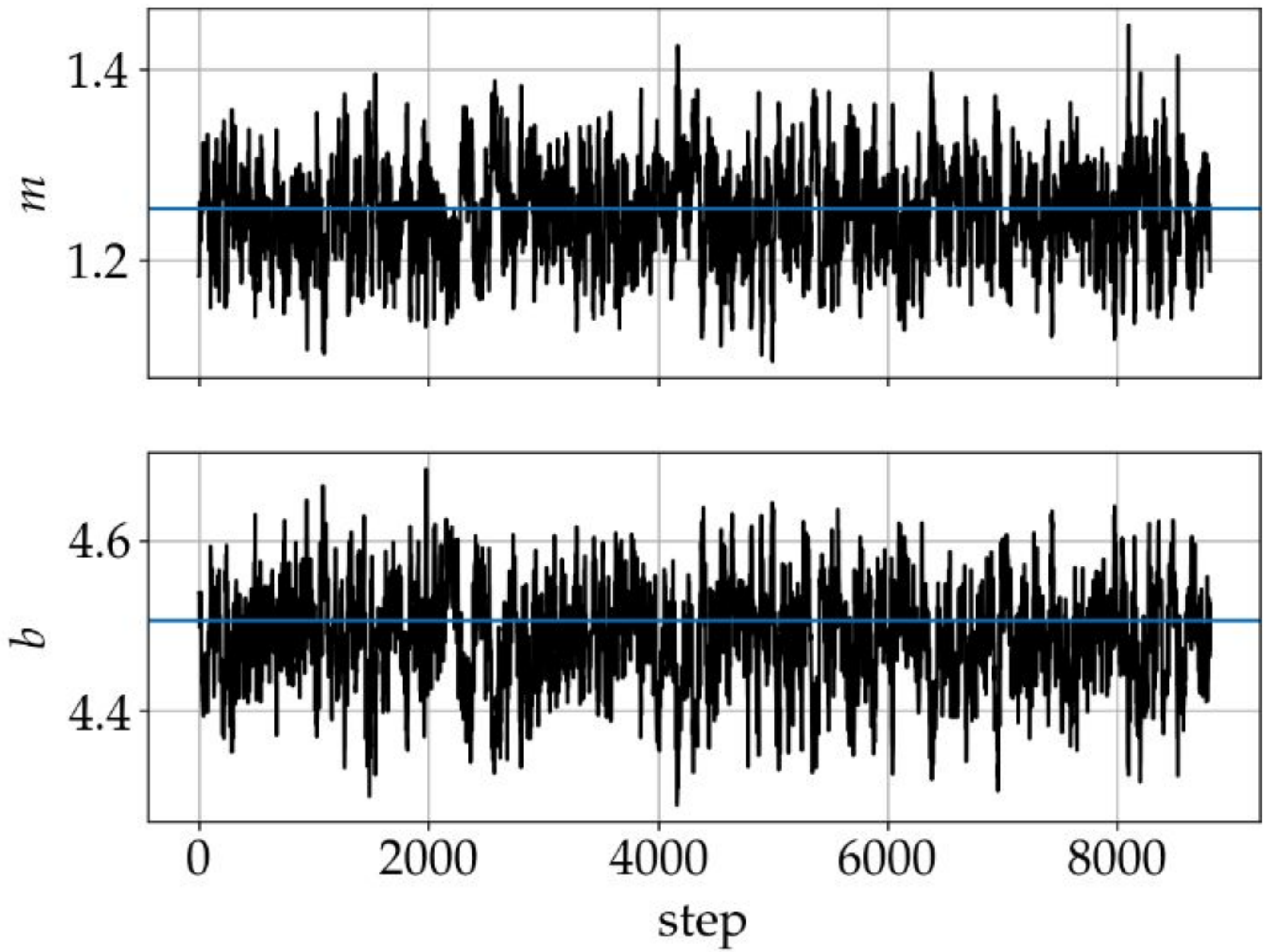
3. If $f(\theta')/f(\theta_k) > r$ accept the proposal and set $\theta_{k+1} \leftarrow \theta_k$, otherwise reject the proposal and set $\theta_{k+1} \leftarrow \theta_k$

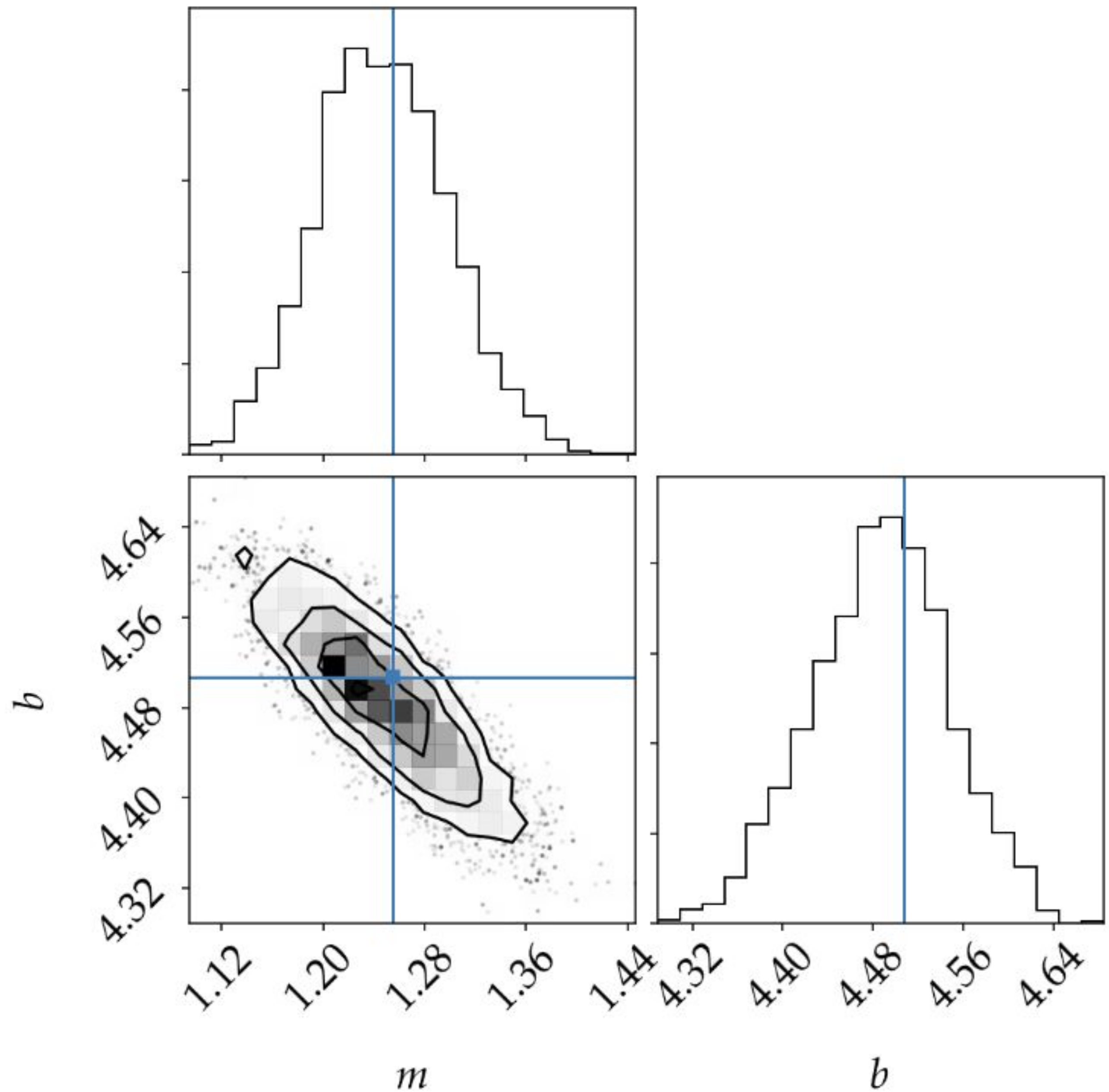
4. Repeat until you have enough samples.

Why does this work?

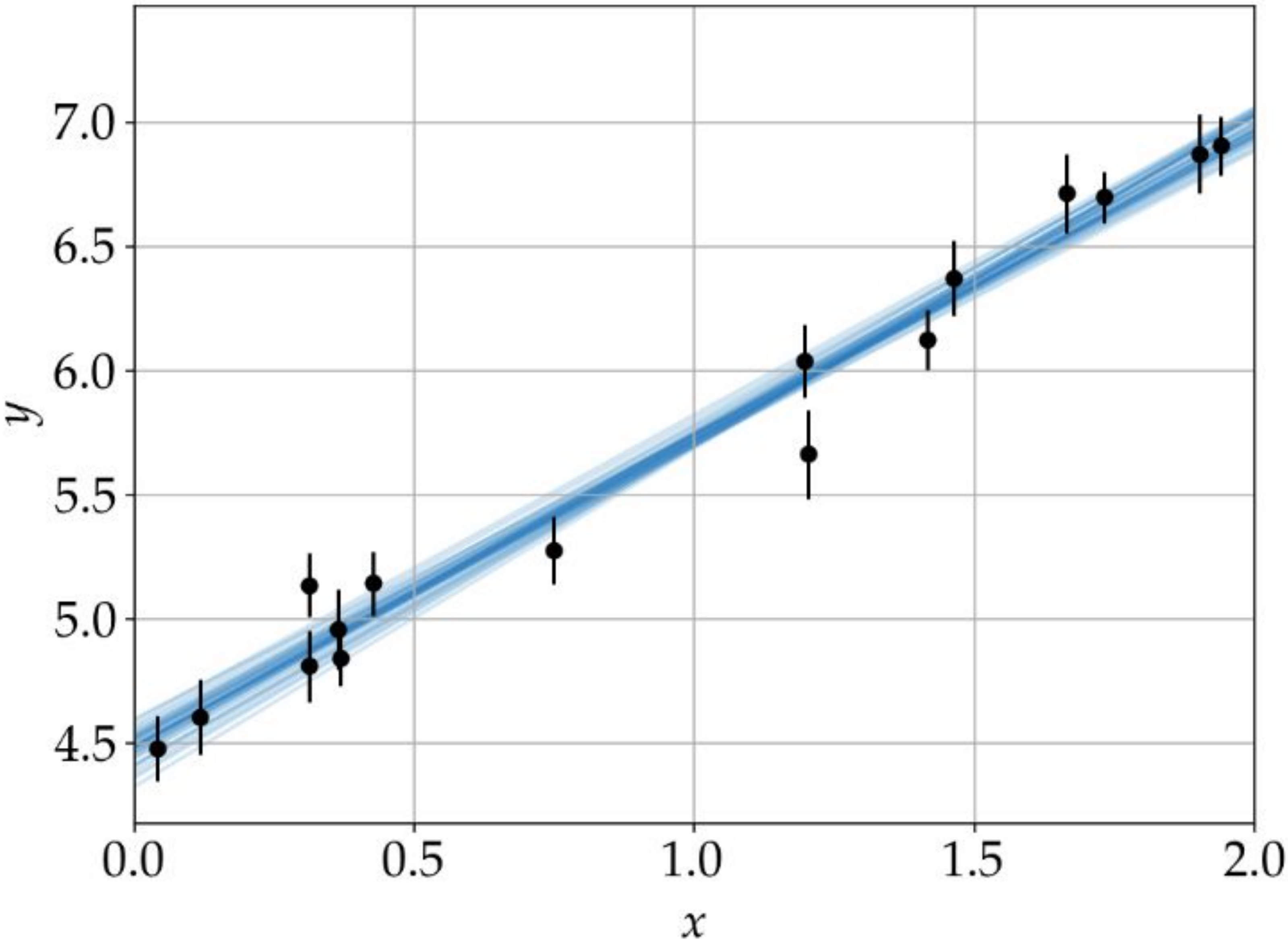
- The MH algorithm defines a sequence of states called a Markov Chain
- In a Markov Chain each state θ_{k+1} depends only on the previous state θ_k
- Markov chains have a stationary probability distribution over all states and in this case the stationary distribution happens to be the posterior we're interested in

Posterior samples





b
4.64
4.56
4.48
4.40
4.32
1.12
1.20
1.28
1.36
1.44
m
4.32
4.40
4.48
4.56
4.64
b



7.0
6.5
6.0
5.5
5.0
4.5
y
0.0
0.5
1.0
1.5
2.0
x

How to tell if the chains converged to the posterior?

- You can't really :(
- Samples $\{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots\}$ generated from a Markov process are (by construction) not independent
- One can measure this with an integrated correlation time

$$\mathrm{E}_{p(\theta)}\left[g(\theta)\right] \approx \frac{1}{K}\sum_{k=1}^{K} g(\theta_k) \qquad \sigma^2 = \frac{\tau_f}{K}\mathrm{Var}\left[g(\theta)\right]$$

$\tau_f =$ number of steps before the chain "forgets" where it started

$N/\tau_f =$ effective sample size

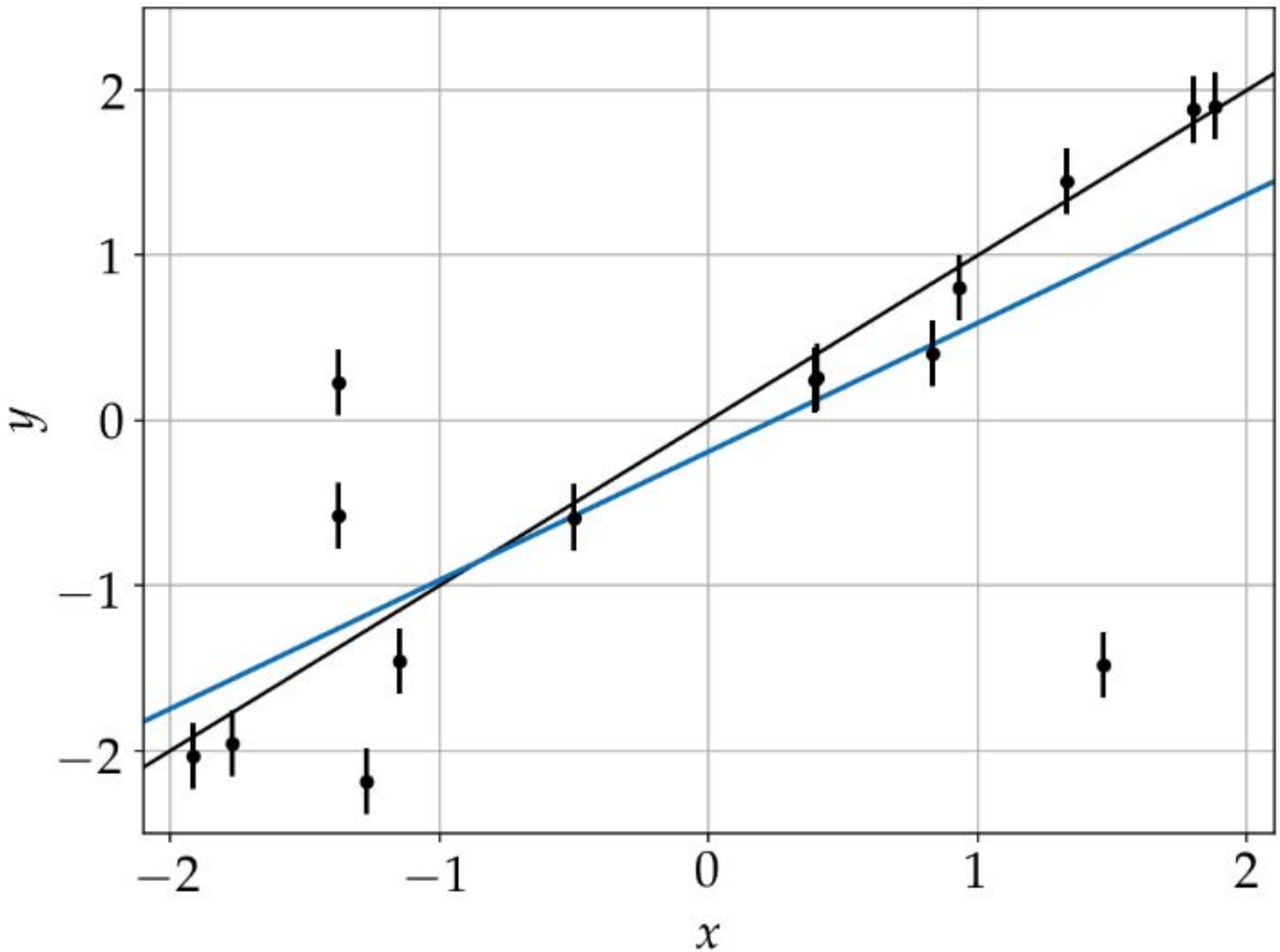
Other samplers

- Gibbs samplers
- Affine invariant ensamble samplers
- Hamiltonian Monte Carlo
- Nested Sampling

Freely available packages

- emcee (http://dfm.io/emcee/current/)
- Stan (http://mc-stan.org/)
- Dnest4 (https://github.com/eggplantbren/DNest4)

Extending the model to account for outliers



Extending the model to account for outliers

$$p(y_i|x_i, \sigma_i, m, b, \mu_b, \sigma_b^2, q_i) = \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{[y_i - mx_i - b]^2}{2\sigma_i^2}\right)\right]^{q_i} \left[\frac{1}{\sqrt{2\pi(\sigma_i^2 + \sigma_b^2)}} \exp\left(-\frac{[y_i - \mu_b]^2}{2(\sigma_i^2 + \sigma_b^2)}\right)\right]^{1-q_i}$$

Extending the model to account for outliers

$$p(y_i|x_i, \sigma_i, m, b, \mu_b, \sigma_b^2, q_i) = \left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{[y_i - mx_i - b]^2}{2\sigma_i^2}\right)\right]^{q_i} \left[\frac{1}{\sqrt{2\pi(\sigma_i^2 + \sigma_b^2)}} \exp\left(-\frac{[y_i - \mu_b]^2}{2(\sigma_i^2 + \sigma_b^2)}\right)\right]^{1-q_i}$$

$$P(q_i) = \begin{cases} P_b & \text{if } q_i = 0 \\ 1 - P_b & \text{if } q_i = 1 \end{cases}$$

$$p(y_i|x_i, \sigma_i, m, b, \mu_b, \sigma_b^2) = \sum_{\{q_i\}} \prod_{i=1}^{N} p(q_i) p(y_i|x_i, \sigma_i, m, b, \mu_b, \sigma_b^2, q_i)$$

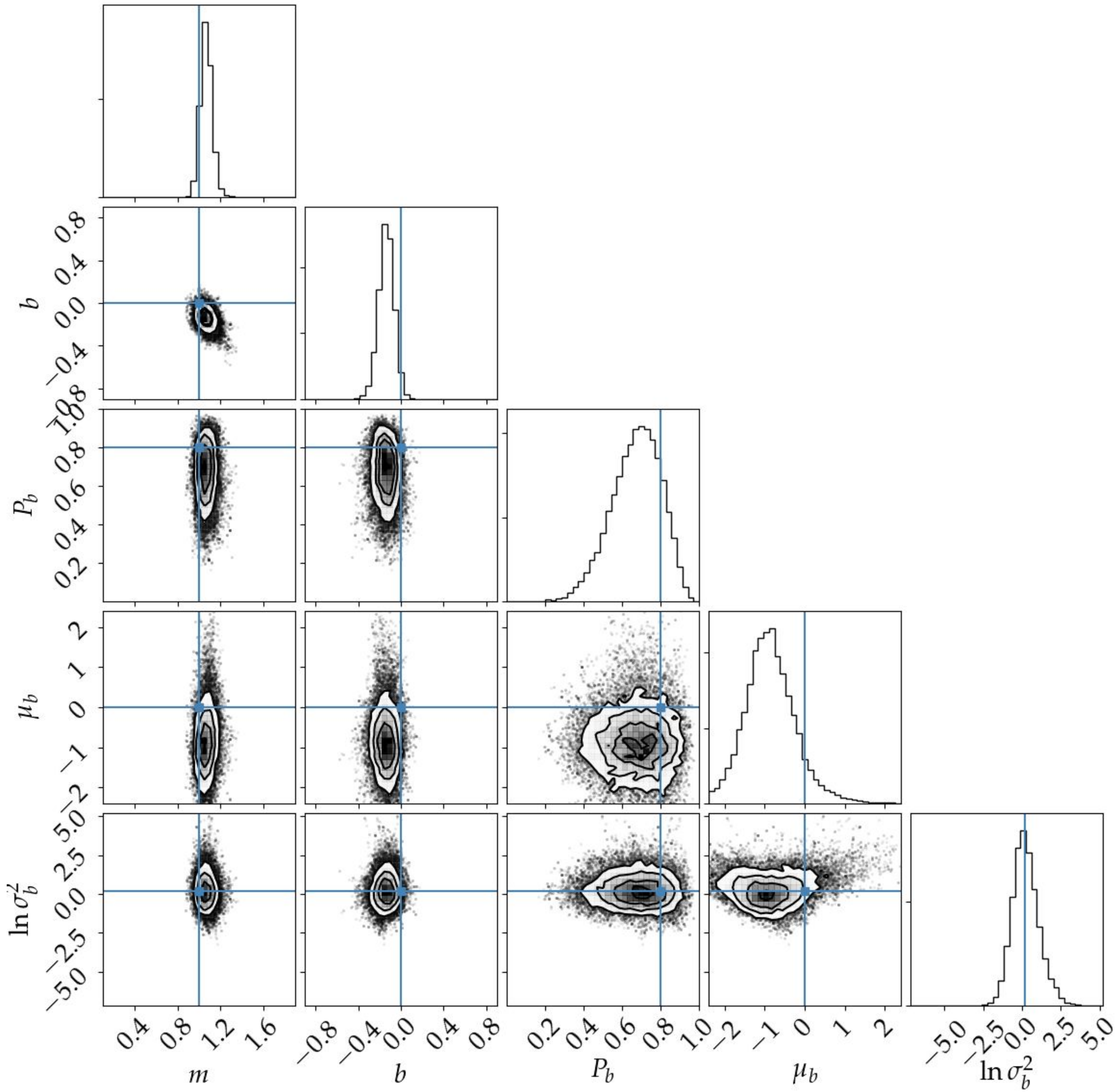
Extending the model to account for outliers

$$p(y_i|x_i,\sigma_i,m,b,\mu_b,\sigma_b^2,q_i) = \left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{[y_i - mx_i - b]^2}{2\sigma_i^2}\right)\right]^{q_i}\left[\frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_b^2)}}\exp\left(-\frac{[y_i - \mu_b]^2}{2(\sigma_i^2+\sigma_b^2)}\right)\right]^{1-q_i}$$

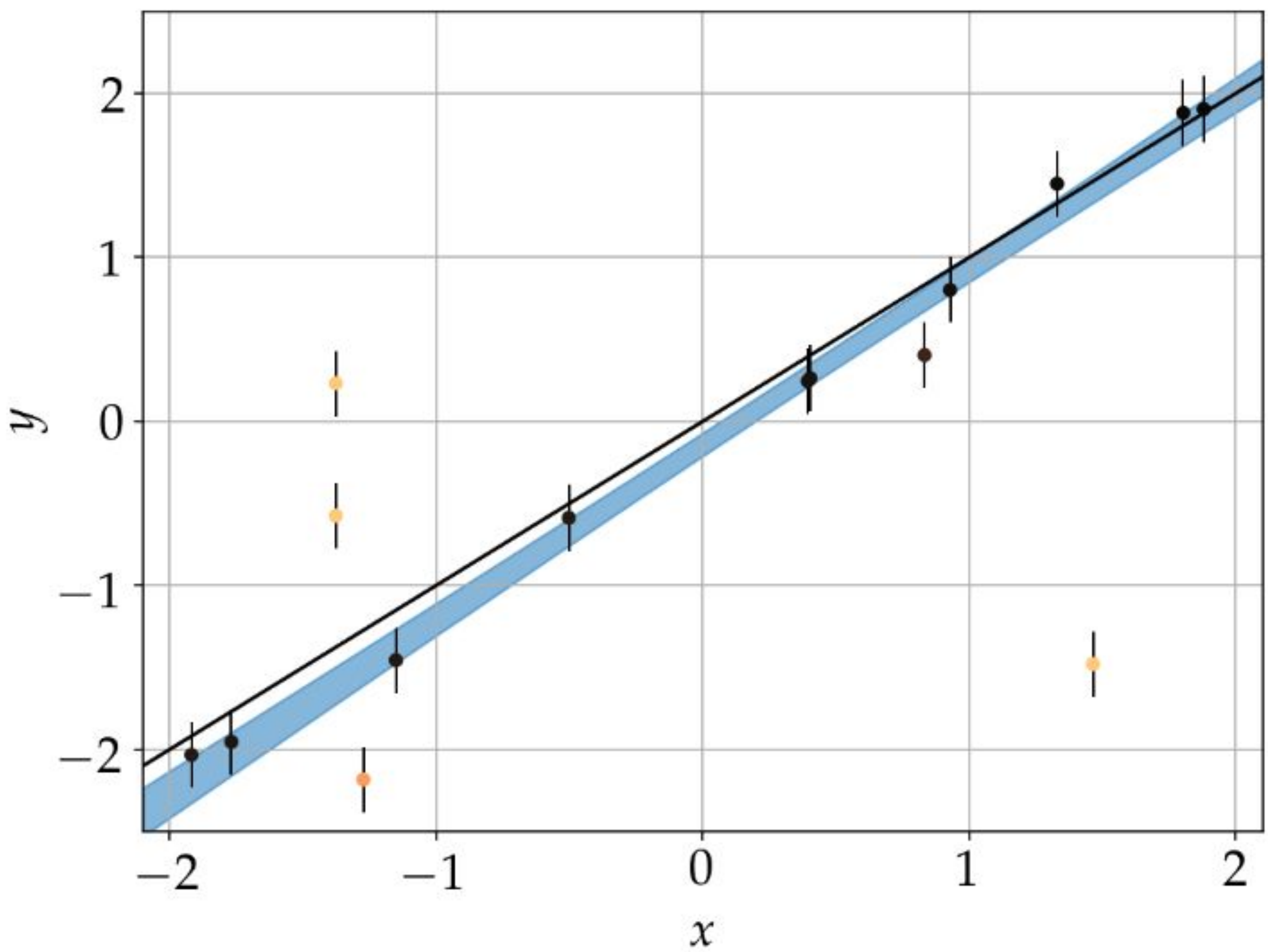
$$P(q_i) = \begin{cases} P_b & \text{if } q_i = 0 \\ 1 - P_b & \text{if } q_i = 1 \end{cases}$$

$$p(y_i|x_i,\sigma_i,m,b,\mu_b,\sigma_b^2) = \sum_{\{q_i\}}\prod_{i=1}^{N} p(q_i)p(y_i|x_i,\sigma_i,m,b,\mu_b,\sigma_b^2,q_i)$$

$$\mathcal{L} = \prod_{i=1}^{N}\left[P_b\mathcal{N}(mx_i+b,\sigma_i) + (1-P_b)\mathcal{N}(\mu_b,\sigma_b)\right]$$



m
b
P_b
μ_b
$\ln \sigma_b^2$


2
1
0
−1
−2
y
−2
−1
0
1
2
x

Where to go from here

- Check out the code on the Code&Cake GitHub repo
- If your models are simple you might not need anything fancy, Metropolis-Hastings MCMC is fine
- If you’re dealing with very high dimensional spaces, check out Hamiltonian Monte Carlo
- If your working on simulations and can’t evaluate your model quickly, look into Aproximate Bayesian Computation (ABC) methods

Thank you for attention!

fbartolic