Modeling of Data

Physics 113 – 02/25/21

General Problem setup

- Given a set of observations, want to summarize data by fitting a model f
- Model depends on a set of adjustable parameters $\theta_1, \theta_2, \ldots \theta_k$
- Models can come from underlying theory to explain the observations or they can be simply used to interpolate or extrapolate the observations.

$$\{y_i, \mathbf{x_i}\}_{i=1}^{N} \qquad\qquad y = f(\mathbf{x}|\theta_1, \theta_2, \ldots \theta_k)$$

- Approach: Define ’Merit function’ that measures agreement between observations and model with particular set of parameters.
- The parameters of the model are adjusted to find an extremum (’best fit’) of the merit function -> Optimization!

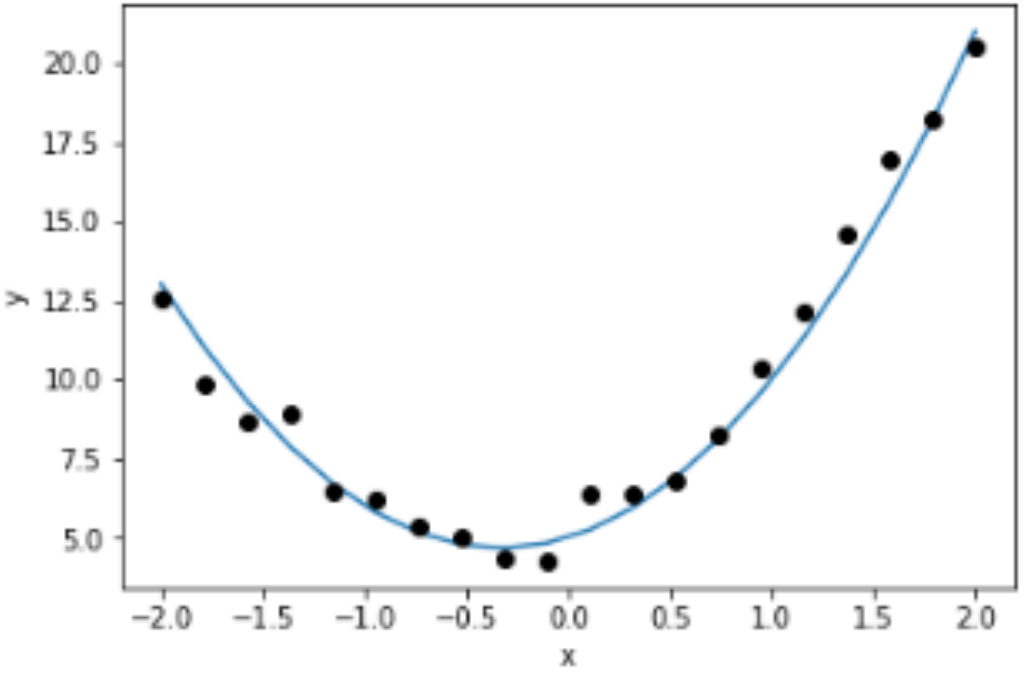
Example (Simple)

- A 1 dimensional quadratic model

$$y = f(x|a,b,c) = ax^2 + bx + c$$

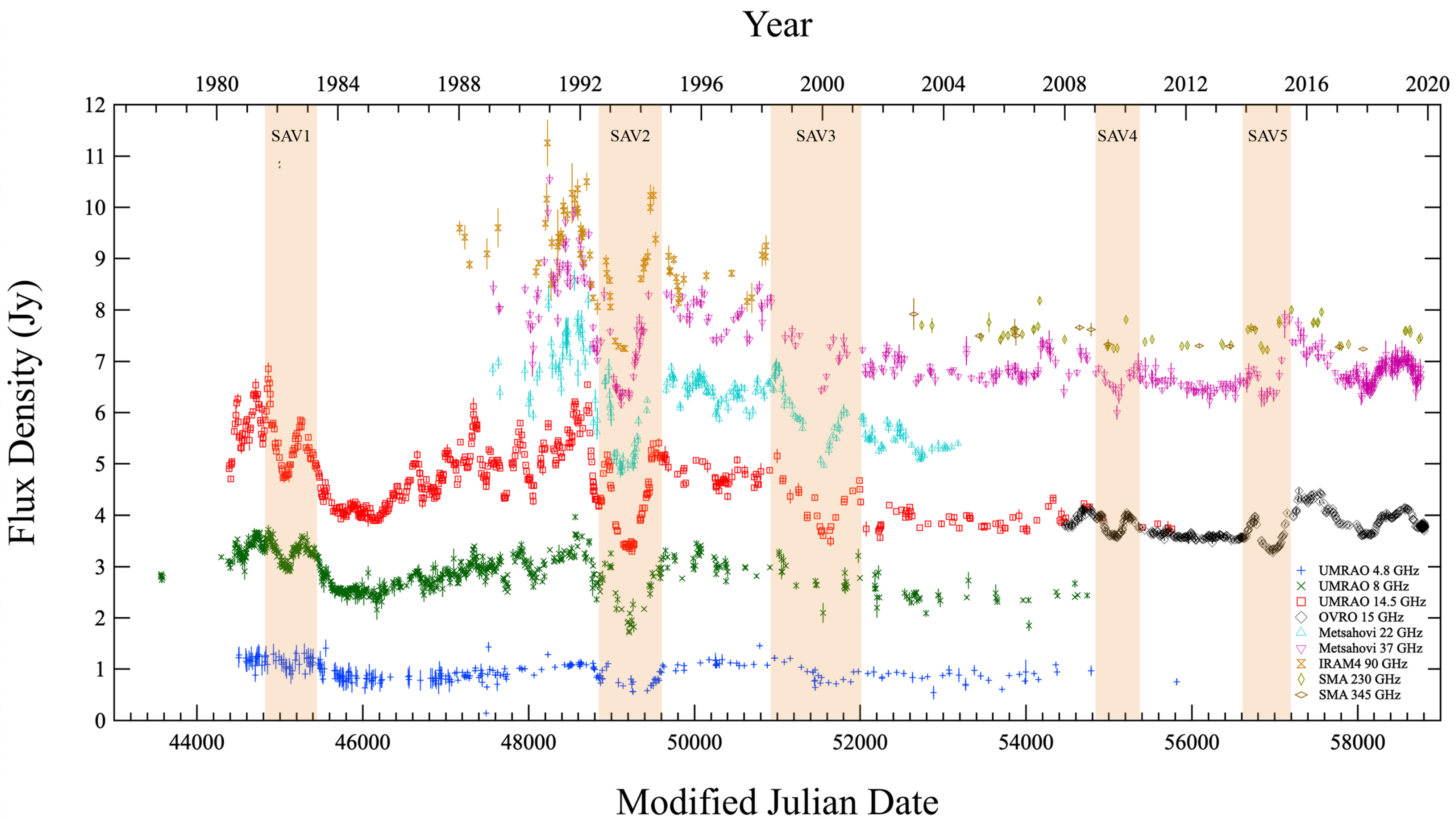
- A multidimensional quadratic model

$$y = f(x_1, x_2|a,b,c) = ax_1^2 + bx_2 + c$$



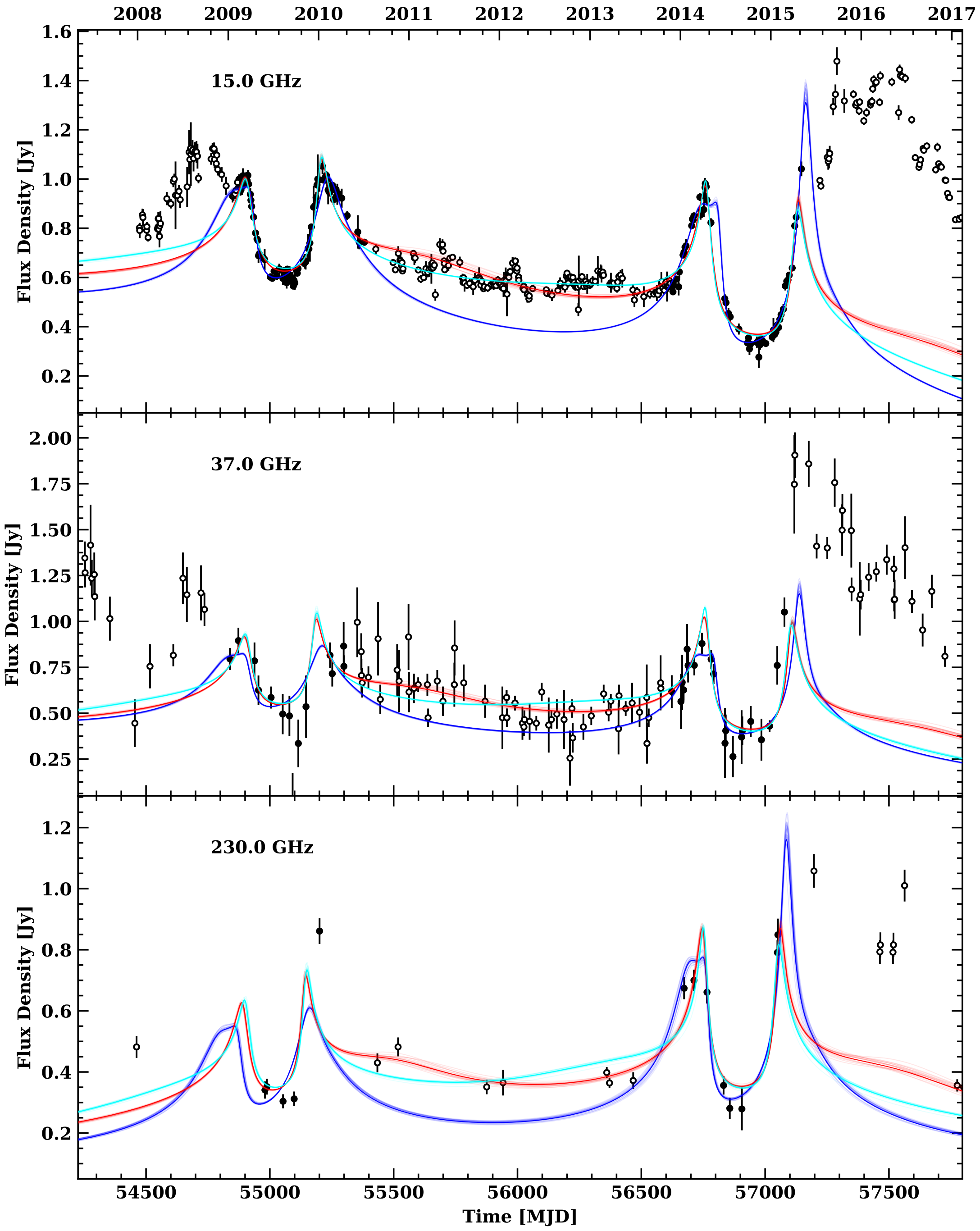
Example (Advanced)

- Fitting a gravitational lens magnification pattern to PKS 1413+135



$$y = f(t|t_0, t_E, q, s, \alpha, \gamma, \kappa, \theta, \beta_0, \beta_E)$$

- Very different models, same fitting paradigm.



Outline

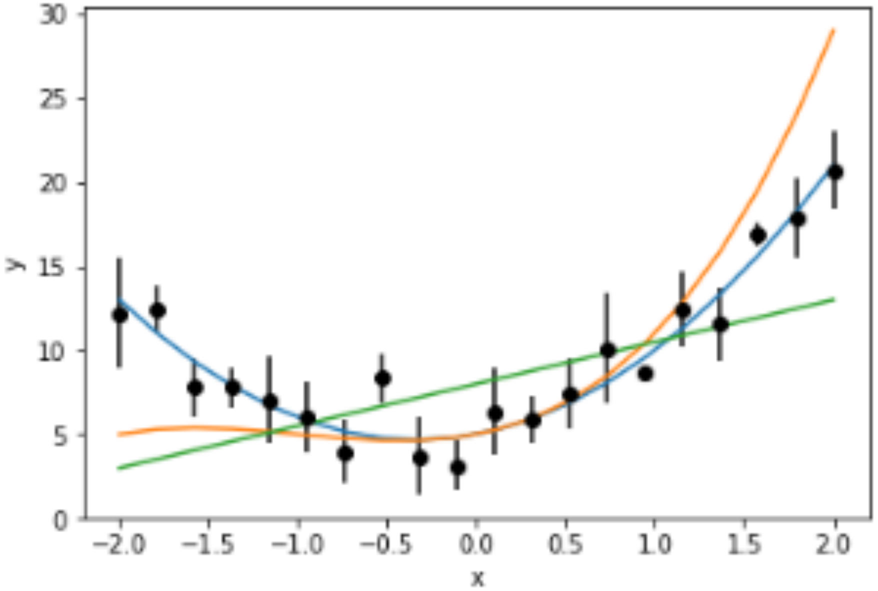
- Goodness of fit and merit functions
- Least squares
- Linear models
- Errors
- Regularization
- Nonlinear models
- Markov Chain Monte Carlo

Likelihood Function

- Probability of your data given your model

$$L(\theta) = p(\{y_i\}_{i=1}^{N} | \theta)$$

- Measured data always has some degree of uncertainty.
- Even if your model is perfectly specified, your likelihood will never be L = 1.



$$y = f(x|\theta) + \epsilon$$

Likelihood Function

- The likelihood function can be used as a merit function: the model parameters that best fit your data maximize the likelihood function.
- Its often easier minimize the log-likelihood function:

$$-\log L(\theta)$$

- The specific likelihood function used depends on the error distribution of your data. $y = f(x|\theta) + \epsilon$
- If the data points are independent and have Gaussian errors:

$$-\log L(\theta) \propto \sum_{i=1}^{N} \frac{\|y_i - f(x_i|\theta)\|^2}{\sigma_i^2}$$

Chi-Squared

$$\chi^2_{N-k} \sim \sum_{i=1}^{N} \frac{\|y_i - f(x_i|\theta)\|^2}{\sigma_i^2}$$

- Sum of squared residuals
- Follows a chi-squared distribution with (N-k) degrees of freedom
- A chi-squared distribution has mean (N-k) and variance 2(N-k)

- Reduced chi squared (mean 1, variance 2/(N-k)):

$$\frac{\chi^2_{N-k}}{N-k}$$

Summary

- To find the model parameters that best fit the data, we minimize the negative log-likelihood.
- Maximum Likelihood estimation.
- For Gaussian distributed, independent data, this means minimizing the sum of squared residuals (‘Least squares’).

$$\sum_{i=1}^{N} \frac{\|y_i - f(x_i|\theta)\|^2}{\sigma_i^2}$$

- Examples of non-gaussian situations: Counts in a detector (Poisson), resonance energy of rare particle (Cauchy).

Least squares with linear models

- Linear models are any models that are linear in the parameters to be estimated, e.g.

Linear

$$f(x|a,b) = ax + b$$

$$f(x|a,b,c) = ax^2 + bx + c$$

$$f(\mathbf{x}|a,b) = ax_1 + be^{x_2}$$

Non-Linear

$$f(x|a,b,c) = a + be^{x/c}$$

$$f(x|a,b) = a^2x + b^3$$

Least Squares with linear models

- Any linear model can be written in matrix form:

$$y = ax^2 + bx + c$$

$$y = \begin{bmatrix} x^2 & x & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

- If we have a set of observations, we can write a system of linear equations:

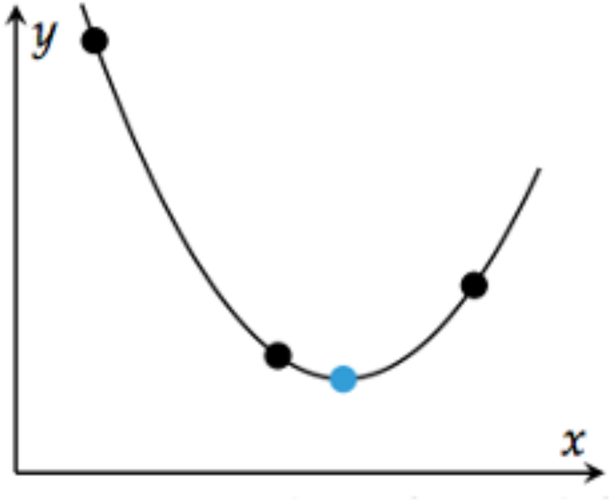
$$\{x_i, y_i\}_{i=1}^N$$

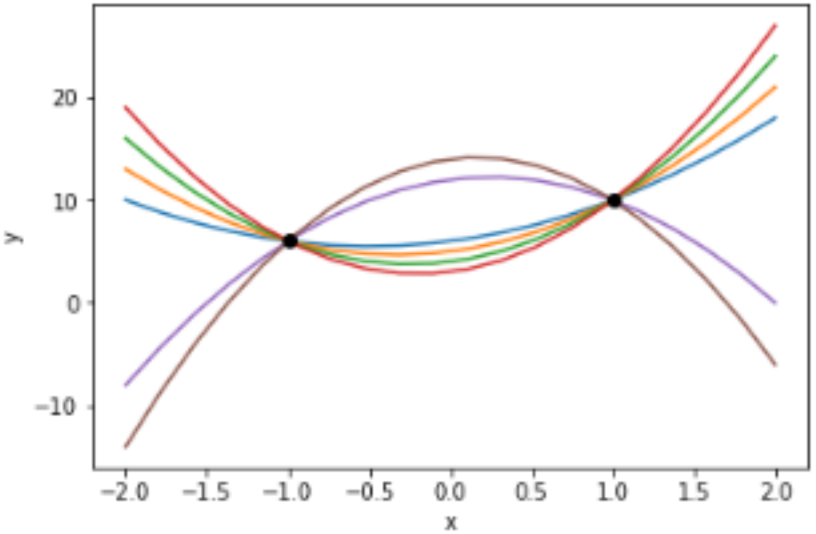
$$\underset{N \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}} = \underset{N \times 3}{\begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_N & 1 \end{bmatrix}} \underset{3 \times 1}{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}$$

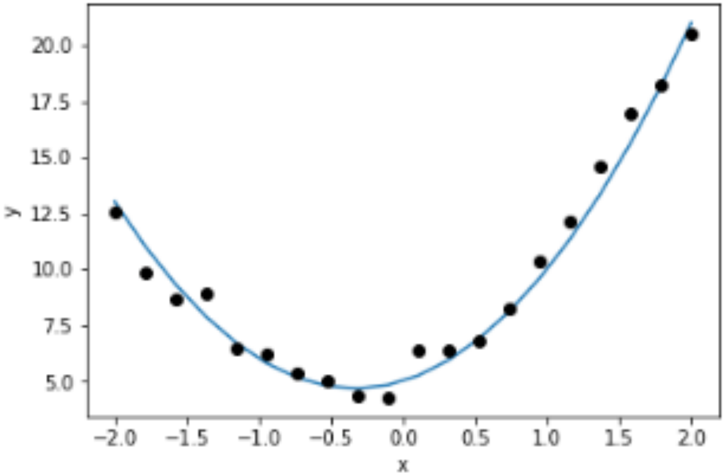
$$\mathbf{y} = X\theta$$

- If N == 3, then there can be only one exact solution for (a,b,c).
- If N < 3, there are many (infinite) exact solutions for (a,b,c).
- If N > 3, the system is underdetermined. There are likely no exact solutions for (a,b,c)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$







Minimizing the Merit function

$$\underset{\theta}{\text{minimize}}\ \|\mathbf{y} - X\theta\|^2 \equiv \underset{a,b,c}{\text{minimize}}\ \sum_{i=1}^{N}(y_i - f(x_i|a,b,c))^2 \equiv \underset{a,b,c}{\text{minimize}}\ \sum_{i=1}^{N}(y_i - (ax_i^2 + bx_i + c))^2$$

- For linear models, we can minimize the merit function analytically:

$$\|\mathbf{y} - X\theta\|^2 = \mathbf{y}^T\mathbf{y} - \theta^T X^T X\theta - 2\mathbf{y}^T X\theta$$

$$\frac{\partial}{\partial\theta}\|\mathbf{y} - X\theta\|^2 = -2X^T X\theta - 2X^T\mathbf{y} = 0$$

$$\boxed{\theta = (X^T X)^{-1} X^T \mathbf{y}}$$

What about with data errors?

- Just add in the weight matrix:

$$W = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & & & \\ & & \dots & 1/\sigma_N^2 \end{bmatrix}$$

$$\boxed{\theta = (X^T W X)^{-1} X^T W \mathbf{y}}$$

Errors on our model parameter estimates

- Now we have our parameter estimates, what are the errors on θ?

- For linear models with gaussian errors we can get these analytically:

$$\mathrm{Var}(\theta) = \Sigma = (X^T W X)^{-1}$$

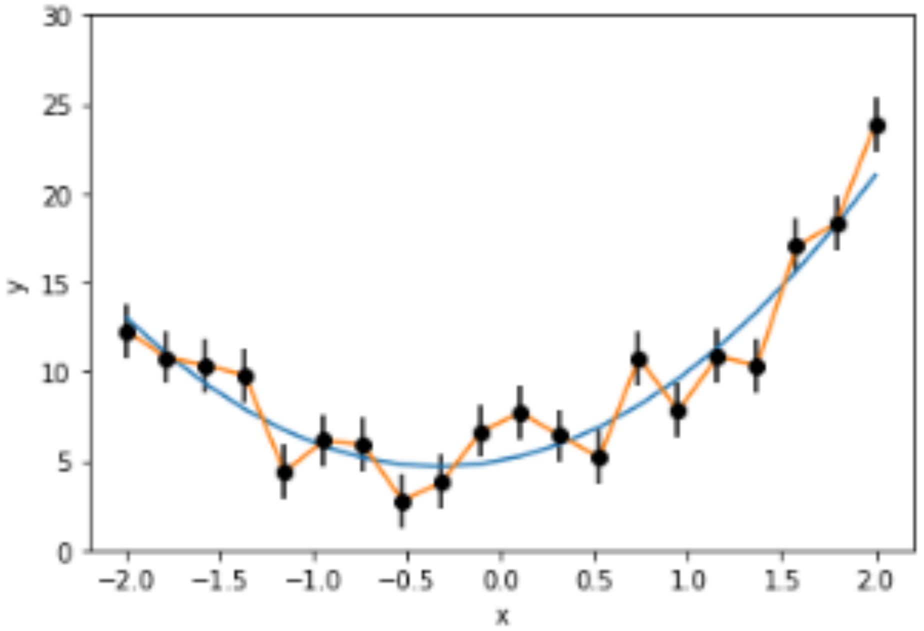
Linearization: the power of linear models

- Always guaranteed you have best solution
- All the information you want derived analytically (estimates, errors on estimates)
- For non-linear problems, you can always try to linearize:

$$f(\mathbf{x}) \approx f(x_0) + \nabla f(\mathbf{x_0})(\mathbf{x} - \mathbf{x_0})$$

- Many machine learning problems solved quite well with linear models.

Assessing goodness of fit



- Reduced Chi Squared

$$\frac{\chi^2_{N-k}}{N-k}$$

- Akaike Information criterion

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

- Bayesian Information Criterion

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

- Overfitting -> Useful in interpolation, bad for model fitting

Techniques to prevent overfitting

- Use a simpler model
- Collect more (independent) data
- Regularization:

$$\underset{\theta}{\text{minimize}}\ \|\mathbf{y} - X\theta\|^2$$

$$\big\downarrow$$

$$\underset{\theta}{\text{minimize}}\ \|\mathbf{y} - X\theta\|^2 + \mu\|\theta\|^2$$

$$\theta = (A^T A + \mu I)^{-1} A^T \mathbf{y}$$

Non-linear Models

- What if your model is non-linear (or the errors are non-Gaussian)?
- Now there is no analytic solution to the minimization problem:

$$\underset{\theta}{\text{minimize}} \; \|\mathbf{y} - f(\mathbf{x}|\theta)\|^2$$

- Can use minimization methods from the previous lecture!
- Gradient Descent, Newton’s method etc.
- There may be many local minima, difficult to guarantee that you have found the best solution.
- No analytic way to get errors on your parameter estimates.

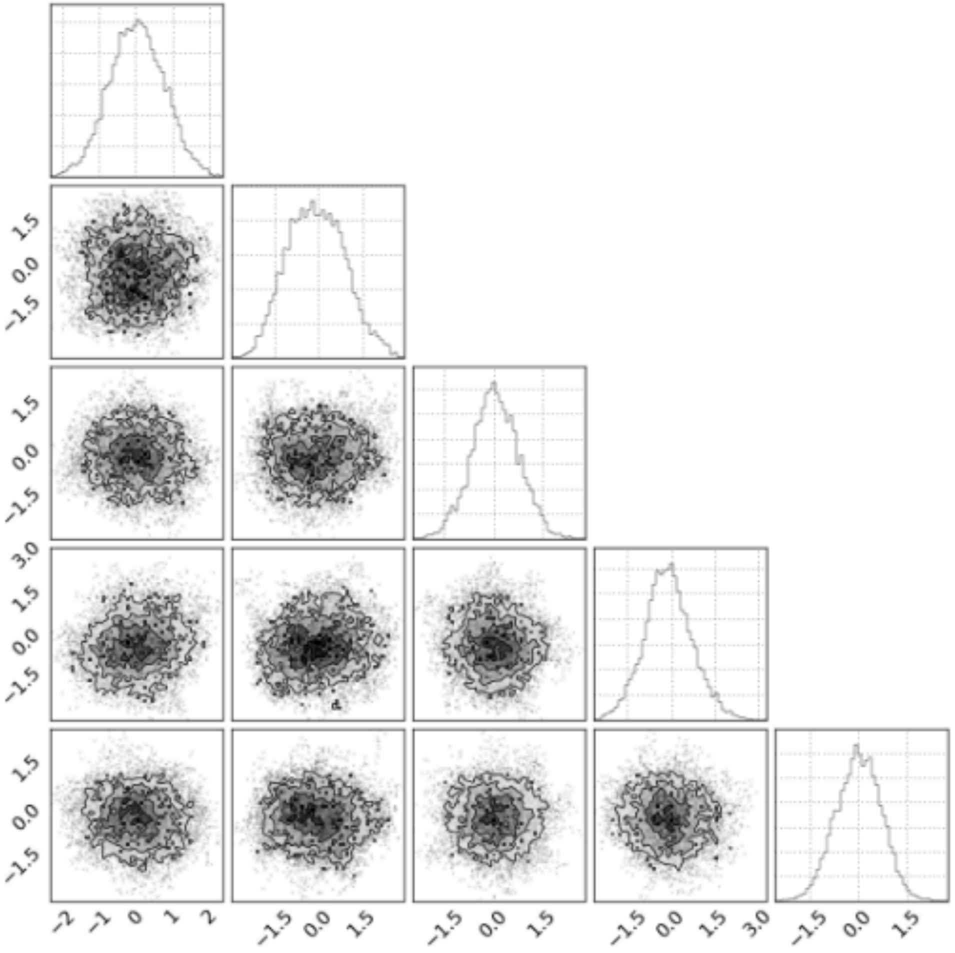
Markov Chain Monte Carlo

- Stochastic Optimization method
- No gradients or Hessians required

- Randomly samples theta space to (hopefully) give full posterior distribution for theta.
- Popular in Astrophysics

Posterior Likelihood Prior

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta)$$



Useful to use log probabilities

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta)$$

$$\log p(\theta|\mathbf{x}) \propto \log p(\mathbf{x}|\theta) + \log p(\theta)$$

- An uninformative prior is often used

$$\log p(\theta|\mathbf{x}) \propto \log p(\mathbf{x}|\theta)$$

- For model fitting with Gaussian errors:

$$\log p(\theta|\mathbf{x}) \propto -\|\mathbf{y} - f(\mathbf{x}|\theta)\|^2$$

How can we sample from $p(\theta|\mathbf{x})$?

- Metropolis-Hastings Algorithm
- Use a Markov Chain model: Probability of selecting next point only depends on current and previous point.

1. Set Gaussian proposal distribution: $q(\theta_2|\theta_1)$

2. Starting at θ_1 draw a candidate point from q, θ_{2c}

3. Accept point with probability $\alpha(\theta_1, \theta_{2c}) = \min\big(1, \frac{p(\theta_{2c}|\mathbf{x})}{p(\theta_1|\mathbf{x})}\big) = \min\big(1, \frac{p(\mathbf{x}|\theta_{2c})}{p(\mathbf{x}|\theta_1)}\big)$

Otherwise reject and $\theta_2 = \theta_1$

4. Repeat.

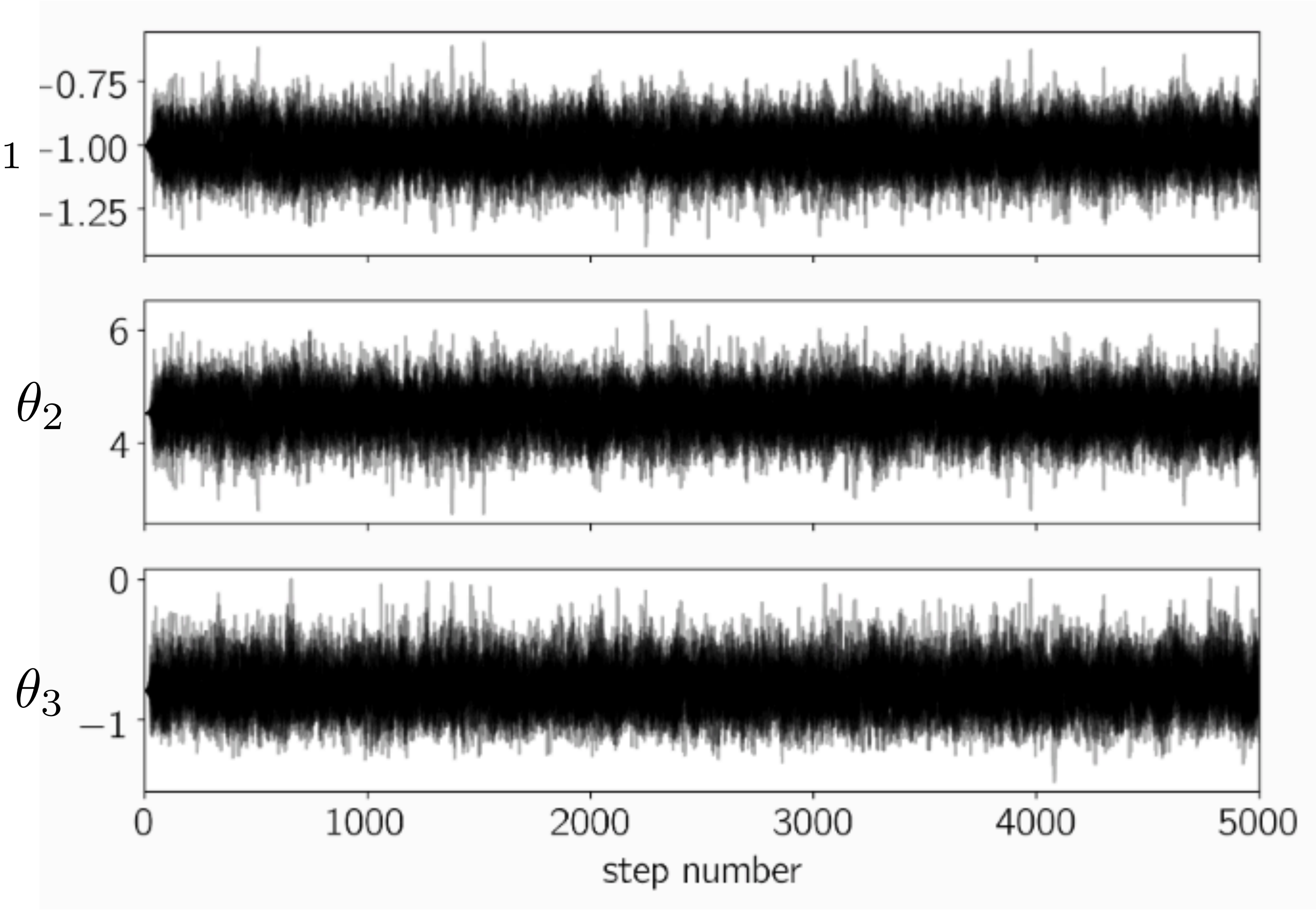
In our case

- Uniform prior and $\log p(\theta|\mathbf{x}) \propto -\|\mathbf{y} - f(\mathbf{x}|\theta)\|^2$
- This means:

$$\alpha(\theta_1, \theta_{2c}) = \min\big(1, \frac{e^{-\|\mathbf{y}-f(\mathbf{x}|\theta_{2c})\|^2}}{e^{-\|\mathbf{y}-f(\mathbf{x}|\theta_1)\|^2}}\big) = \min\big(1, e^{\|\mathbf{y}-f(\mathbf{x}|\theta_1)\|^2-\|\mathbf{y}-f(\mathbf{x}|\theta_{2c})\|^2}\big)$$

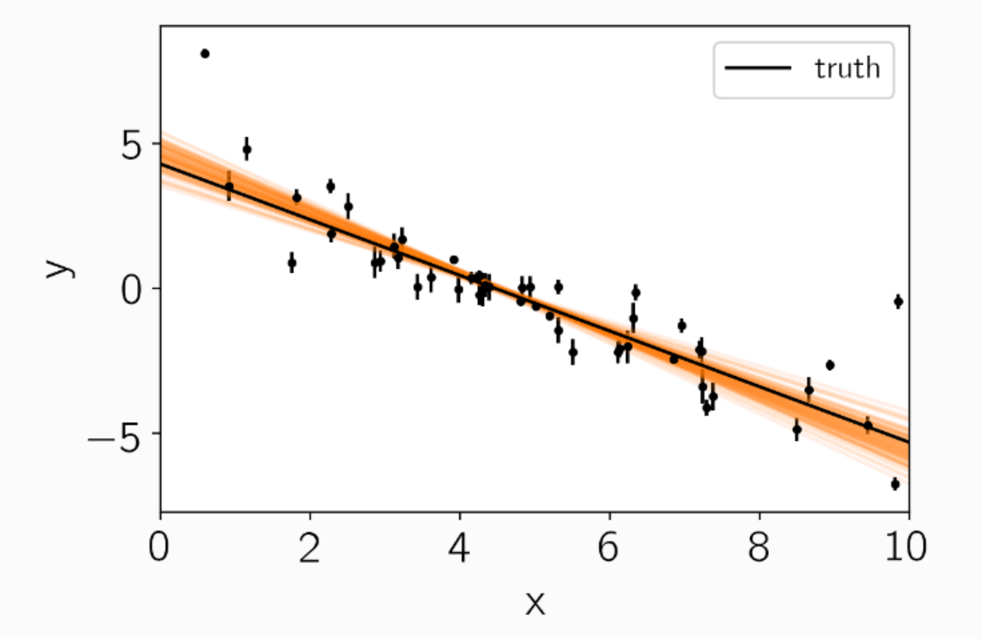
How many steps before convergence?

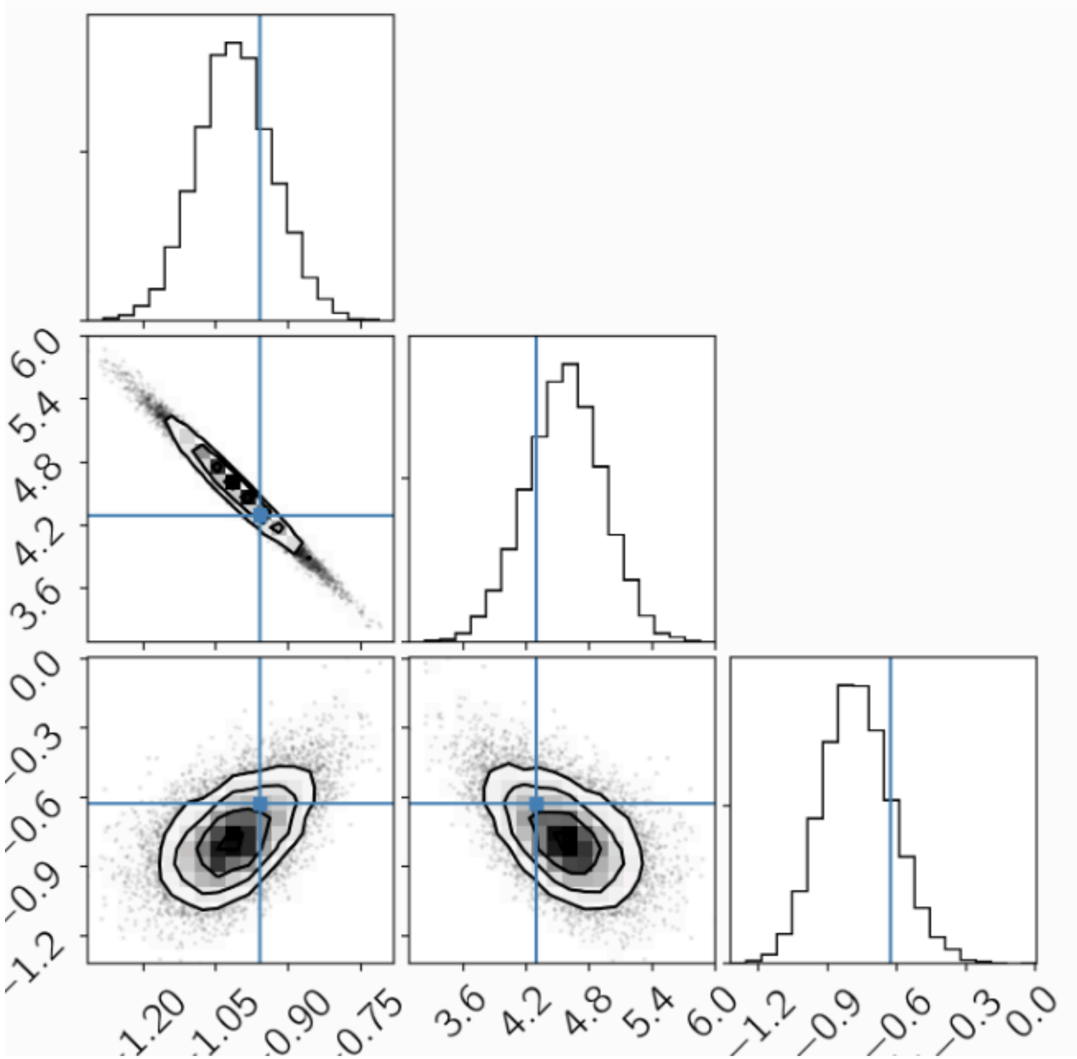
- Open Question
- Usually have a burn in period
- Advisable to thin points to reduce correlation



Corner Plot

- Once finished, you will end up with the sampled posterior distribution.
- Can represent as a corner plot.





Example

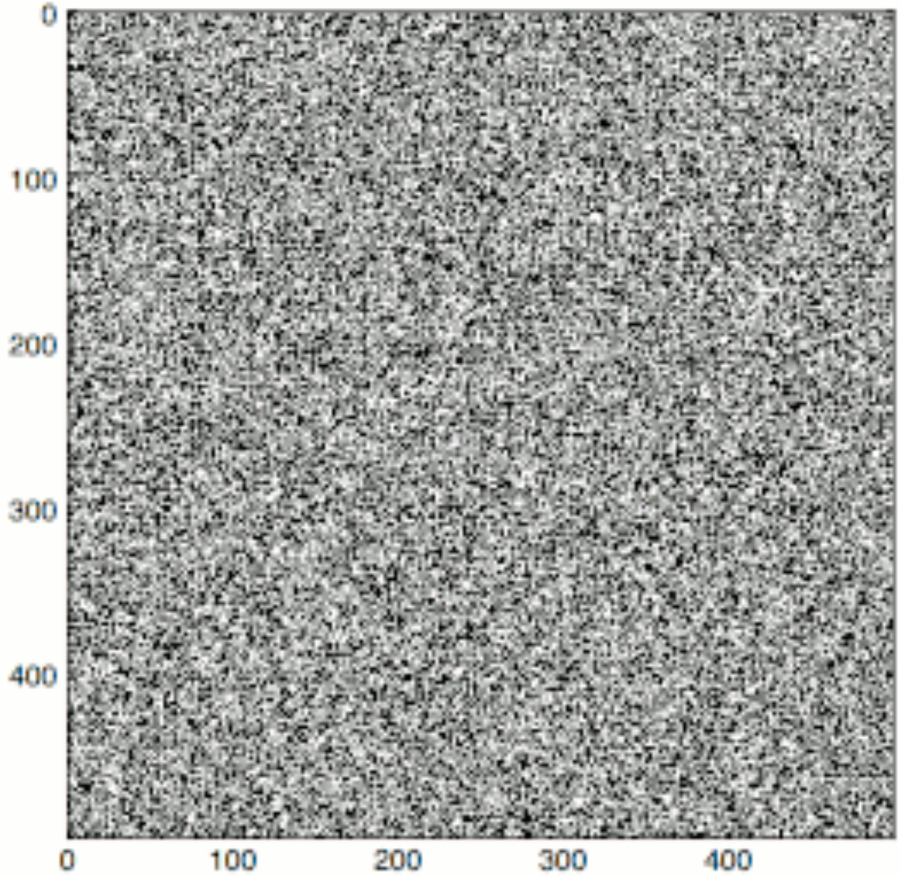
- https://github.com/chi-feng/mcmc-demo#ref-2

Example 2D Ising model

$$H(\sigma) = -J \sum_{\langle i\ j\rangle} \sigma_i \sigma_j - h \sum_j \sigma_j$$

$$\alpha(\theta_1, \theta_{2c}) = \min\Big(1, \frac{e^{-\beta H(\theta_{2c})}/Z}{e^{-\beta H(\theta_1)}/Z}\Big)$$

$$\alpha(\theta_1, \theta_{2c}) = \min\Big(1, e^{-\beta(H(\theta_{2c}) - H(\theta_1))}\Big)$$



Hamiltonian MCMC

- Uses gradient information (so often not possible to use)
- Typically much more efficient than standard Metropolis-Hastings

$$H(\theta, \mathbf{p}) = U(\theta) + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}$$

$$U(\theta) = -\log p(\theta|\mathbf{x})$$