Introduction to Bayesian Modelling

Alan Heavens,
Imperial College

June 20, 2022

Bayesian Deep Learning for Cosmology and Time Domain Workshop
APC Paris

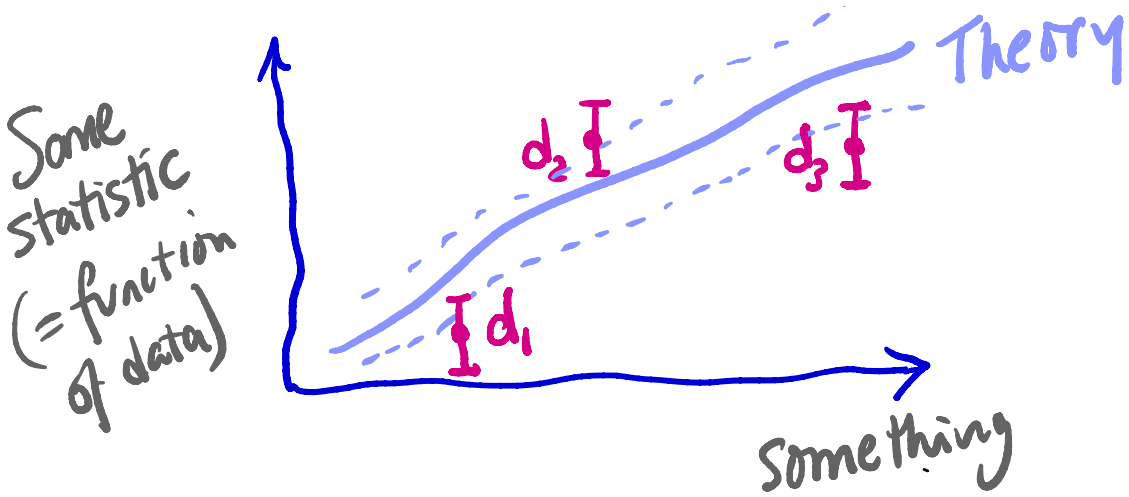
Overview

1. Inverse Problems
2. Parameter Inference
   - The posterior $p(parameters|data)$
   - How to set up a problem
   - Priors
3. Simple Bayesian Analysis
4. Bayesian Hierarchical Models
   - Case study: straight line fitting with errors in x and y
5. Sampling
   - Markov Chain Monte Carlo (MCMC)
   - Metropolis-Hastings algorithm
     - Burn-in
     - Marginalisation from samples
     - Correlated samples
   - Gibbs sampling
   - Hamiltonian Monte Carlo

Inverse Problems

- Analysis problems are *inverse problems*: given some data, we want to infer something about the process that generated the data
- Generally harder than predicting the outcome, given a physical process
- The latter is called *forward modelling*, or a *generative model*
- Typical classes of problem:
  - *Parameter inference*
  - *Model comparison*

How do we do science?

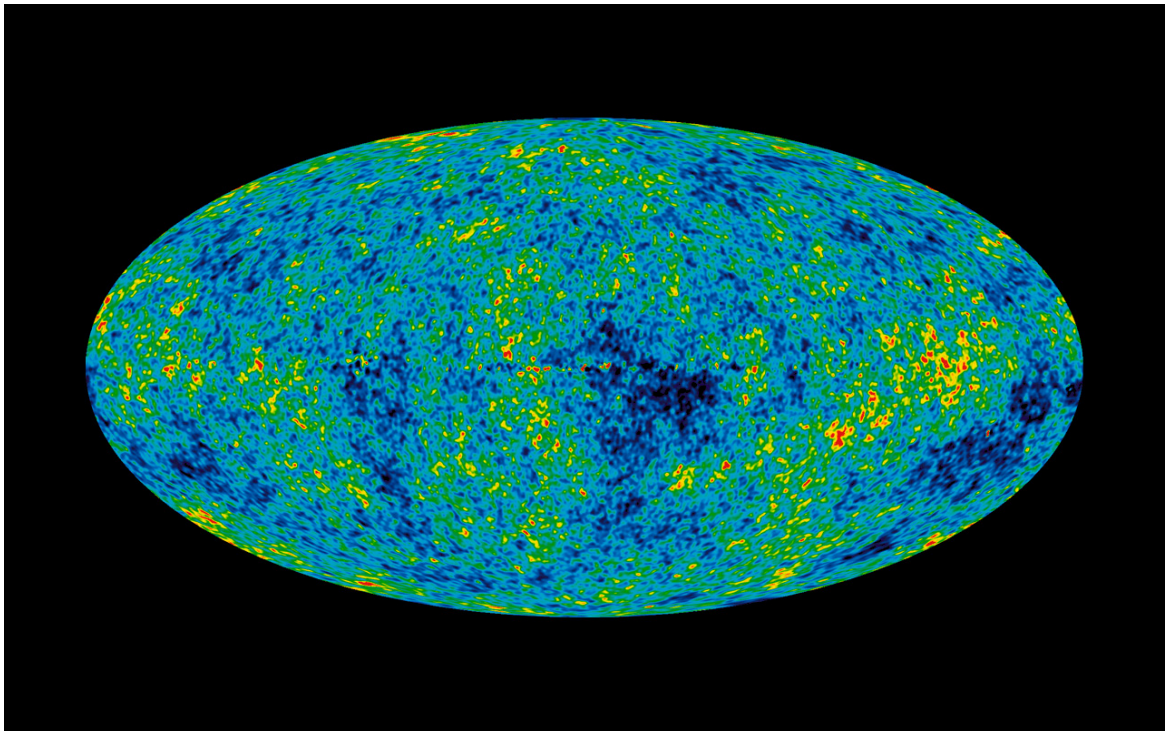


We need to know the *sampling distribution* (often called the likelihood)

$$p(\{d_1, d_2, d_3\}|\mathit{Theory}, \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ represents model parameters.
We *may* know it (gaussian, Poisson etc), but it may be complex (selection effects, complicated physics).
Without it how can we do science?

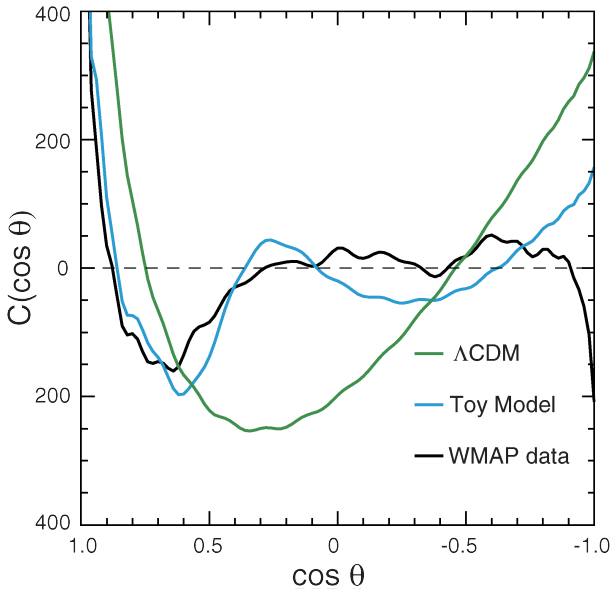
Case study: WMAP Cosmic Microwave Background Data



Typically we compress the data into some 'summary statistics', such as the correlation function of the temperature values, or the power spectrum.

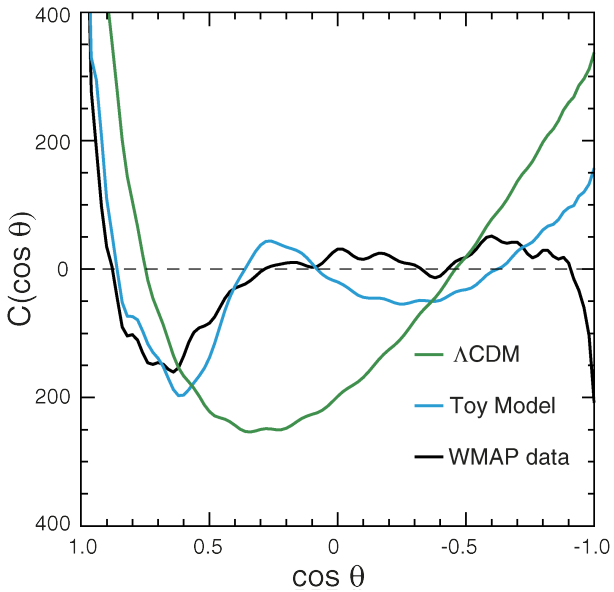
WMAP Cosmic Microwave Background Data

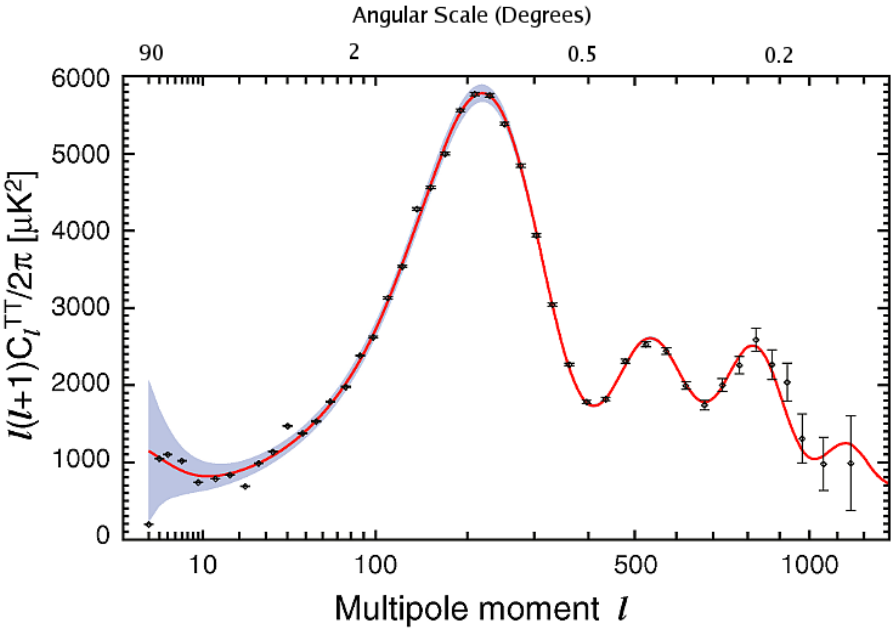
ΛCDM fits WMAP data well:



WMAP Cosmic Microwave Background Data

ΛCDM fits WMAP data well:





Notation

- **Data** d; **Model** M; Model **parameters** θ
- Rule 1: write down what you want to know
- Usually, it is the probability distribution for the parameters, given the data, and assuming a model.
- It is the **Posterior**: $p(\theta|\mathrm{d}, M)$
- To compute it, we use Bayes theorem:

$$p(\theta|d, M) = \frac{p(d|\theta, M)p(\theta|M)}{p(d|M)}$$

- where the **Likelihood** is $\mathcal{L}(\mathrm{d}|\theta) = p(\mathrm{d}|\theta, M)$
- and the **Prior** is $\pi(\theta) = p(\theta|M)$
- $p(\mathrm{d}|M)$ is the **Bayesian Evidence**, which is important for Model Comparison, but not for Parameter Inference.
- Dropping the M dependence

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{p(d)}$$

It is all probability

The Posterior

Everything is focussed on getting at the whole posterior $p(\theta|\mathrm{d})$. *Not* just a point estimate of 'best-fit' parameters.

Computing the posterior

$p(\theta|\mathrm{d}) \propto \mathcal{L}(\theta)\,\pi(\theta)$.

We need to analyse the problem:

What are the data, d?
What is the model for the data?
What are the model parameters?
What is the likelihood function $\mathcal{L}(\theta)$? Do we even know it?
What is the prior $\pi(\theta)$?

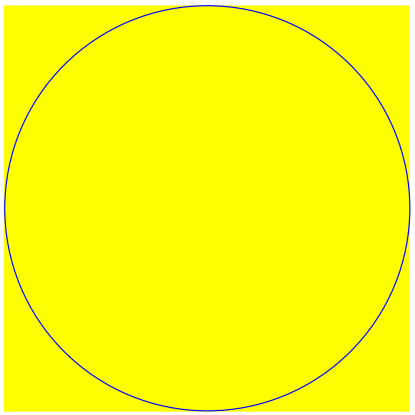
Priors

- Bayesian: prior = (usually) the state of knowledge before the new data are collected.
- For parameter inference, the prior becomes unimportant as more data are added and the likelihood dominates. (For model comparison, the prior remains important.)
- Issues: One usually wants an 'uninformative' prior, but what does this mean?
- Typical choices: $\pi(\theta) =$ constant (for location parameters); $\pi(\theta) \propto 1/\theta$ (for scale parameters) - so-called Jeffreys prior (by Astronomers)

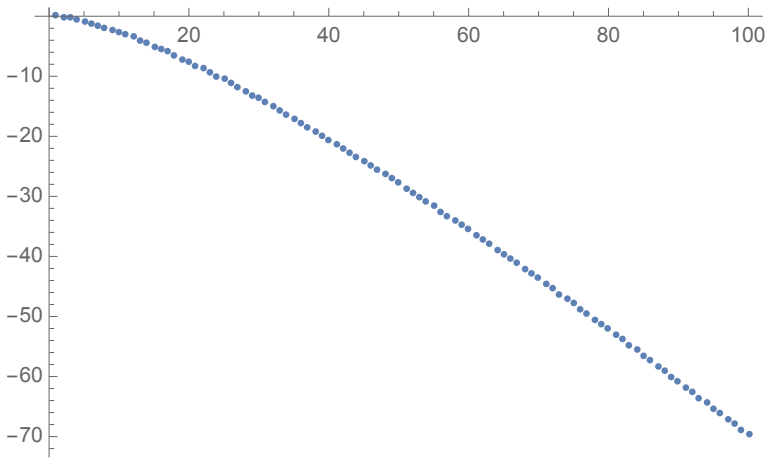
Uninformative prior

Flat prior? Seems natural, but consider this problem. Imagine cartesian coordinates in N dimensions, with the prior range being $(-\frac{1}{2}, \frac{1}{2})$ for all coordinates. The prior probability of being inside the N-sphere which just fits inside the prior volume is

$$\frac{\pi^{N/2}}{2^N \Gamma(1 + N/2)}$$



$\log_{10} p$ vs N



An apparently uninformative prior may be *highly informative* when viewed in a different way.

Simple Bayesian Analysis

Likelihood (sampling distribution) known. E.g. if gaussian,

$$p(d|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(d-\mu)^2}{2\sigma^2}\right]$$

If (as usual) the data $\mathbf{d}$ are multidimensional, they may be correlated, and if they are gaussian, we need the covariance matrix

$$\Sigma = \langle (\mathbf{d}-\boldsymbol{\mu})(\mathbf{d}-\boldsymbol{\mu})^T \rangle.$$

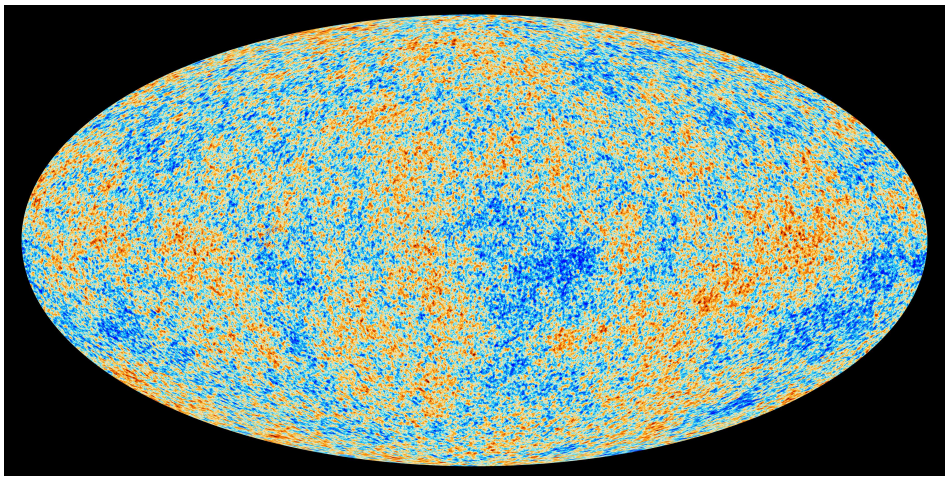
The sampling distribution/likelihood is then

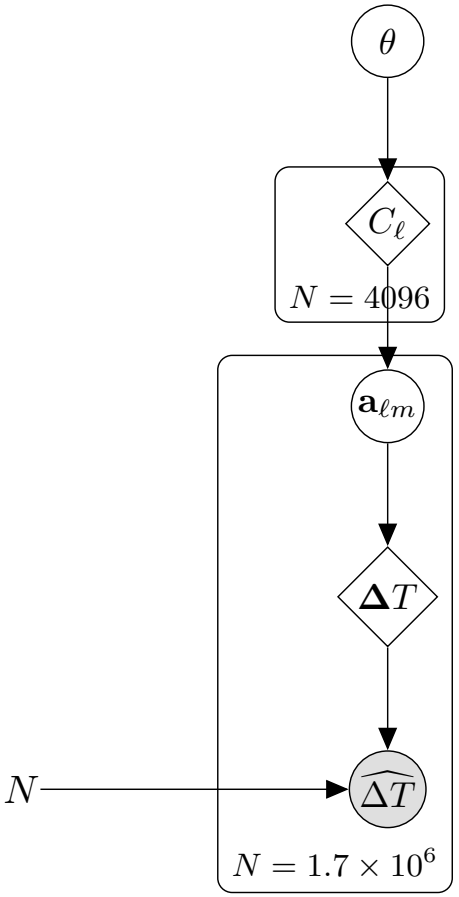
$$p(\mathbf{d}|\theta) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left[-\frac{1}{2}(\mathbf{d}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{d}-\boldsymbol{\mu})\right]$$

Are the data gaussian? Do you know the covariance matrix?

Bayesian Hierarchical Models (BHMs)

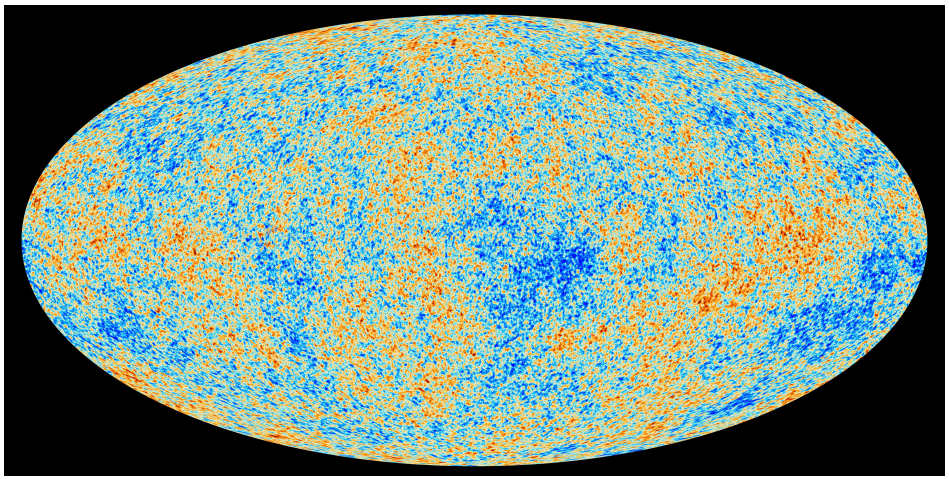
Real data can be more complex. E.g. Data = Planck pixel values, $\widehat{\Delta T}$.
What is the likelihood $p(\widehat{\Delta T}|\boldsymbol{\theta})$? Hard, but not impossible.

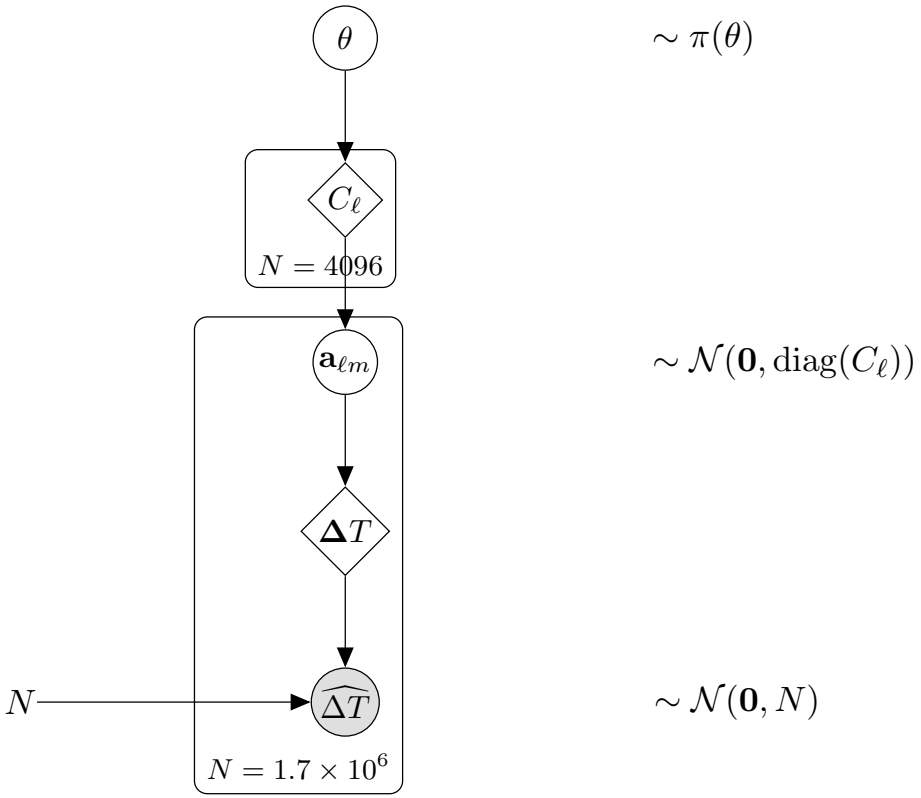




Bayesian Hierarchical Models (BHMs)

Real data can be more complex. E.g. Data = Planck pixel values, $\widehat{\Delta T}$.
What is the likelihood $p(\widehat{\Delta T}|\boldsymbol{\theta})$? Hard, but not impossible.





Bayesian Hierarchical Models, for more complex problems

If you can, this is how to do it

BHM

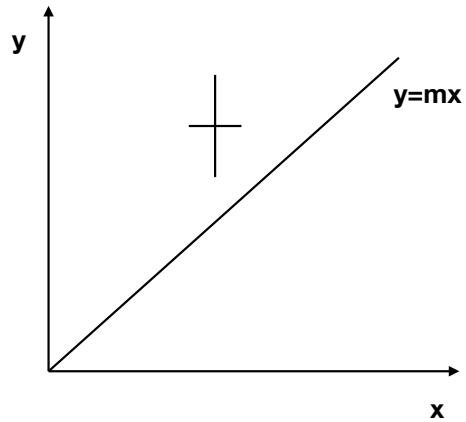
- We split the inference problem into steps, where the full model is made up of a series of sub-models
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- At each step ideally we will know the conditional distributions
- The aim is to build a complete model of the data
- Principled way to include systematic errors, selection effects (everything, really)

Case study: straight line fitting

- Let us illustrate with an example. We have a set of **data** pairs $(\hat{x}, \hat{y})$ of noisy measured values of x and y
- **Model:** $y = mx$
- **Parameter:** m.
- Complication: $\hat{x}$ and $\hat{y}$ are both noisy.
- How do we infer m?
- First, apply Rule 1: write down what you want to know.
- It is

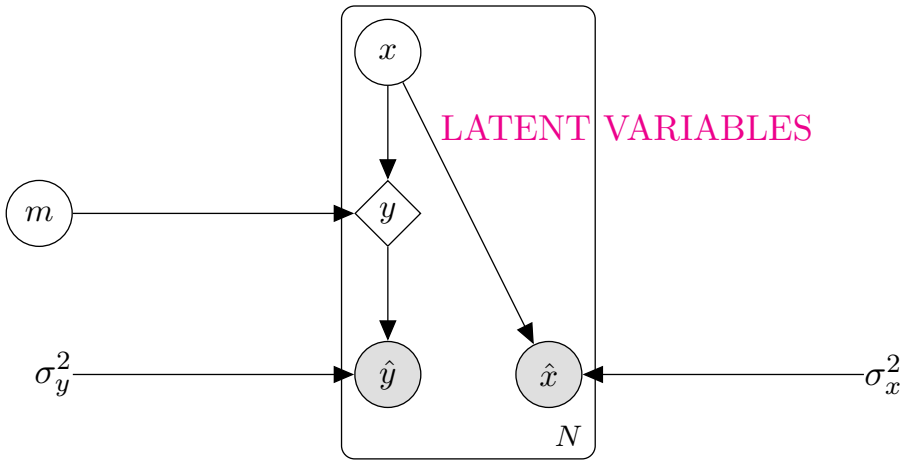
$$p(m|\hat{x}, \hat{y})$$

- We will take $\hat{x} = 10$, $\hat{y} = 15$, with independent unit gaussian errors.



Straight line fitting

How would you forward model it?



- Break problem into steps.
- There are extra unknowns in this problem (so-called **latent variables**), namely the *unobserved true values of* $\hat{x}$ *and* $\hat{y}$.
- The model connects the *true* variables. i.e.,

$$y = mx.$$

- The latent variables x and y are **nuisance parameters** - we are (probably) not interested in them, so we marginalise over them.

Hierarchical Bayes vs Ordinary Bayes

- Hierarchical Bayes:

$$p(m|\hat{x},\hat{y}) \propto p(\hat{x},\hat{y}|m)\, p(m)$$

- We do not know the likelihood $p(\hat{x},\hat{y}|m)$ directly, and we *introduce the latent variables and marginalise over them:*

$$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y},x,y|m)\, p(m)\, dx\, dy$$

- Let us now analyse the problem. Manipulating the last equation

$$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,y,\cancel{m})\, p(x,y|m)\, p(m)\, dx\, dy$$

Analysis

- $$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,y)\, p(y|x,m)\, p(x|\cancel{m})\, p(m)\, dx\, dy$$

  This splits the problem into a noise term, a theory term, and priors. We can write all of these down.

- Here, the theory is deterministic:

$$p(y|x,m) = \delta(y - mx)$$

  Integration over y is trivial with the Dirac delta function:

$$p(m|\hat{x},\hat{y}) \propto \int p(\hat{x},\hat{y}|x,mx)\, p(x)\, p(m)\, dx.$$

Choose some priors, and integrate, or sample from the joint distribution of m and x:

$$p(m,x|\hat{x},\hat{y}) \propto p(\hat{x},\hat{y}|x,mx)\, p(x)\, p(m)$$

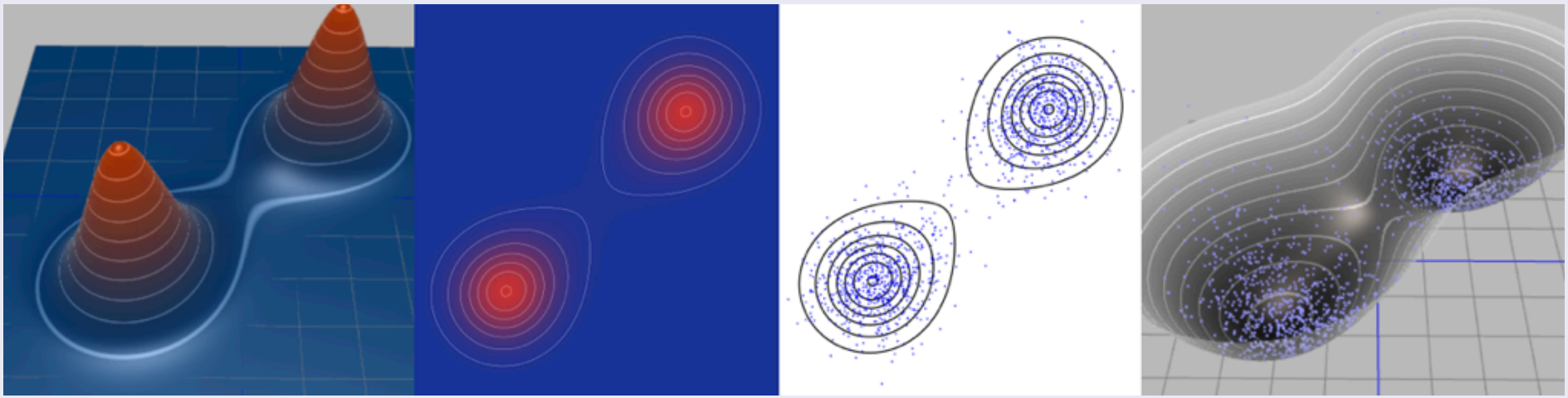
Sampling

The posterior is rarely a simple function, and evaluating it on a parameter grid can be prohibitively expensive with > 2 or 3 parameters.

MCMC

Standard technique is MCMC (Markov Chain Monte Carlo), where random steps are taken in parameter space, according to a proposal distribution, and accepted or rejected according to the Metropolis-Hastings algorithm. This gives a chain of samples of the posterior (or the likelihood), with an expected number density proportional to the posterior.

MCMC example



Sampling algorithms

There are several generic MCMC (Markov Chain Monte Carlo) algorithms, where random steps are taken in parameter space, according to a proposal distribution. We will concentrate on three common ones:

- Metropolis-Hastings
- Gibbs Sampling
- Hamiltonian Monte Carlo (HMC)

Goal: generate samples of the *target distribution* (usually the posterior or the likelihood), with an expected number density proportional to the posterior. This will be satisfied asymptotically if the algorithm satisfies *detailed balance*.

The target distribution need not be normalised, but it needs to be everywhere positive, and normalisable (i.e. the integral is finite).

Markov processes

Sequential process where new element depends only on the previous element.
The general algorithm is as follows:

- Choose a starting point $\boldsymbol{\theta}_0$. e.g. randomly from a prior.
- $\boldsymbol{\theta}_{s+1}$ generated from $\boldsymbol{\theta}_s$ by generating a trial point randomly from a *proposal distribution*, and which is either accepted or rejected (depending on the algorithm)[1]
- If accepted, trial becomes the next sample. If rejected, the previous sample is repeated.
- The chain is stopped at some point. There is no magic answer as to when to stop, but the main idea is to reach *convergence*. e.g. Gelman-Rubin test. Must do convergence tests!

[1] Some algorithms, such as Gibbs, may always accept, dependent on some factors.

Metropolis-Hastings algorithm

For low-dimensional problems. Draw from a *proposal distribution* to generate a new *proposed* sample $\boldsymbol{\theta}_{s+1}$

$$q(\boldsymbol{\theta}_{s+1}|\boldsymbol{\theta}_s) \tag{1}$$

Often this is a function of $|\boldsymbol{\theta}_{s+1} - \boldsymbol{\theta}_s|$, but it doesn't have to be, and a common choice is a gaussian centred on the previous sample in the chain. The algorithm specifies that the point is accepted with probability

$$\alpha = min\left[1, \frac{\rho(\boldsymbol{\theta}_{s+1})}{\rho(\boldsymbol{\theta}_s)} \frac{q(\boldsymbol{\theta}_{s+1}|\boldsymbol{\theta}_s)}{q(\boldsymbol{\theta}_s|\boldsymbol{\theta}_{s+1})}\right]. \tag{2}$$

As a rule of thumb, an acceptance rate of ~ 0.3 is usually efficient.

Burn-in and marginalisation

Throw away exploratory phase: e.g. find first sample within some factor (e.g. 10) of the highest value, and discard all the previous samples.

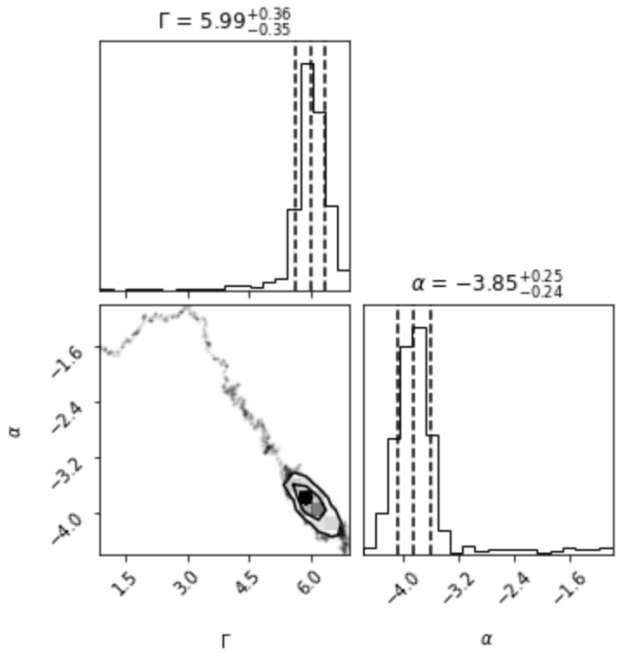


Figure: 2D posterior for LHC background parameters.

Marginalisation from samples

Trivial. Each sample has values for all of the parameters. If you want the distribution of θ_1, simply ignore the values of the other parameters.

Correlated samples

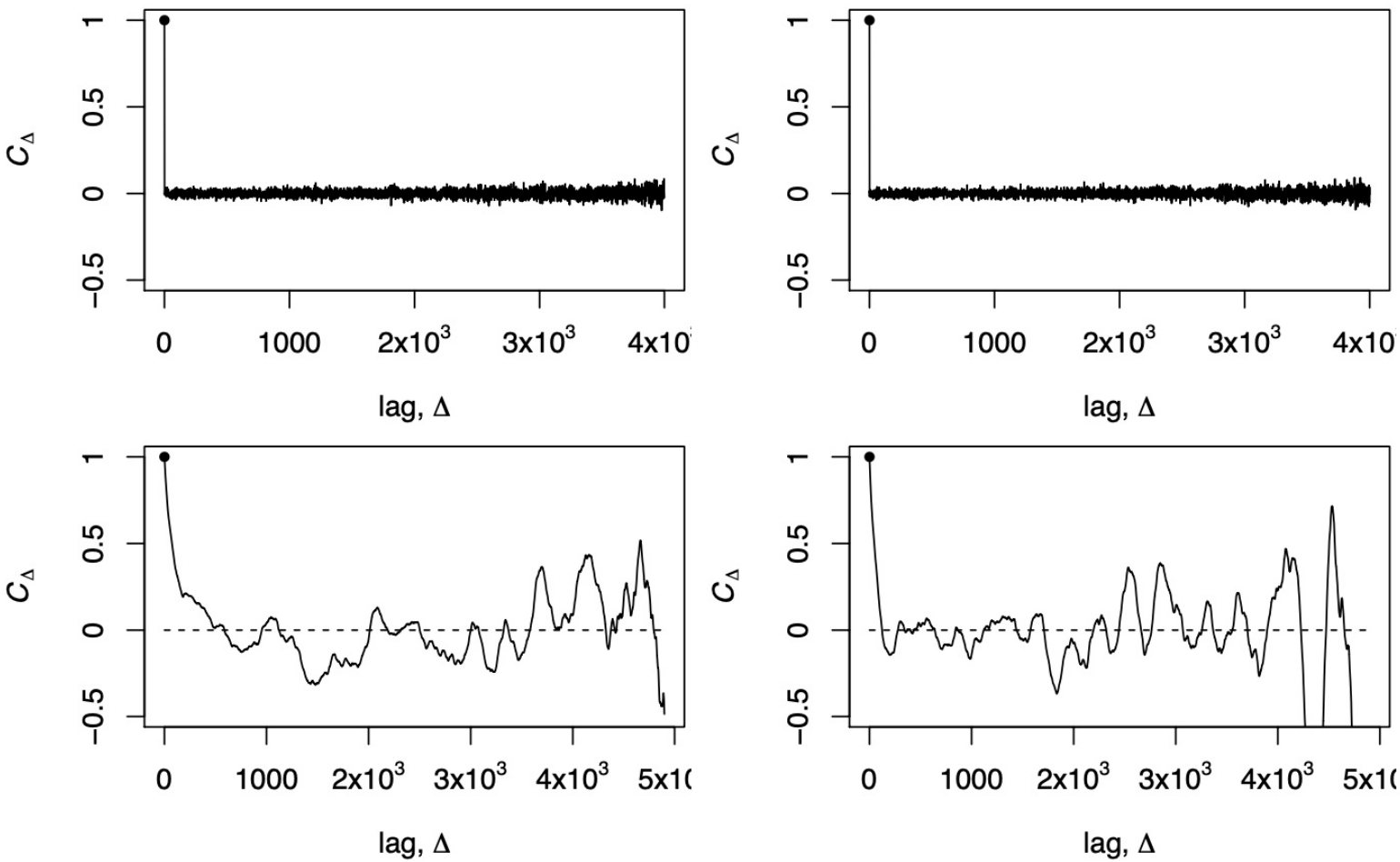


Figure: Correlation coefficient of samples for uncorrelated samples (top) and badly-correlated samples (bottom). From D. Mortlock.

Gibbs sampling

Powerful *if the conditional distributions are known*. Algorithm:

- $\theta_1^{s+1} \sim p(\theta_1 | \theta_2^s, \theta_3^s, \ldots, \theta_n^s)$
- $\theta_2^{s+1} \sim p(\theta_2 | \theta_1^{s+1}, \theta_3^s, \ldots, \theta_n^s)$
- etc ...

Repeat, randomizing (or reversing) the order.

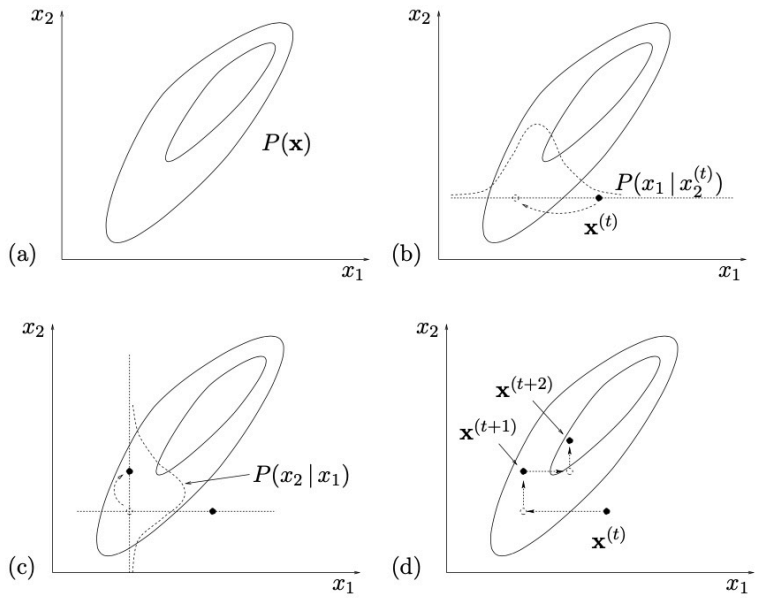


Figure: From Mackay (2003). Slow if target is highly correlated.

Hamiltonian Monte Carlo

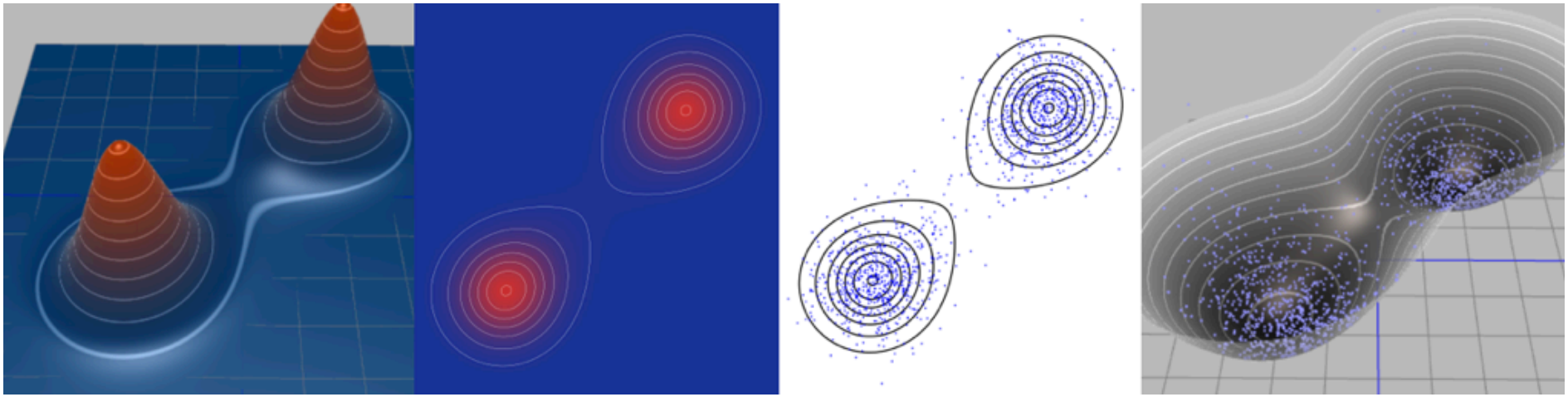


Figure: Credit: Alex Rogozhnikov

HMC defines a potential $U(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta})$, where $p(\theta)$ is the target distribution. Think of $\boldsymbol{\theta}$ as a position vector. Define a kinetic energy

$$K(\mathbf{u}) = \frac{1}{2}\mathbf{u} \cdot \mathbf{u} \tag{3}$$

where $\mathbf{u}$ is a momentum, drawn randomly, e.g. $u_i \sim \mathcal{N}(0, \sigma^2)$.
The Hamiltonian (energy) is conserved:

$$H(\theta, \mathbf{u}) = U(\theta) + K(\mathbf{u}) \tag{4}$$

Hamiltonian Monte Carlo

Define new target distribution in the $2n$-dimensional parameter space:

$$T(\theta, u) = \exp[-H(\theta, u)]. \tag{5}$$

HMC explores this phase space using Hamilton's equations:

$$\begin{aligned} \dot{\theta}_i &= \frac{\partial H}{\partial u_i} = u_i \\ \dot{u}_i &= -\frac{\partial H}{\partial \theta_i} = \frac{\partial \ln p}{\partial \theta_i} \end{aligned} \tag{6}$$

Solve numerically, e.g. leapfrog (symmetric forward-back, to satisfy detailed balance).
Integrate for a while[2], a new proposed sample is generated, and accepted or rejected[3], then a new random momentum is generated.

[2] How long? e.g. until trajectory turns round: No U-turn (NUTS)
[3] H is not quite conserved, because of numerical integration.

Hamiltonian Monte Carlo

Full HMC algorithm is (from Hajian 2006):

1: initialize $\boldsymbol{\theta}_{(0)}$
2: for i $= 1$ to $N_{samples}$
3: $\quad \mathbf{u} \sim \mathcal{N}(0, 1)$ (Normal distribution)
4: $\quad (\boldsymbol{\theta}^*_{(0)}, \mathbf{u}^*_{(0)}) = (\boldsymbol{\theta}_{(i-1)}, \mathbf{u})$
5: $\quad$ for j $= 1$ to N
6: $\quad\quad$ make a leapfrog move: $(\boldsymbol{\theta}^*_{(j-1)}, \mathbf{u}^*_{(j-1)}) \rightarrow (\boldsymbol{\theta}^*_{(j)}, \mathbf{u}^*_{(j)})$
7: $\quad$ end for
8: $\quad (\boldsymbol{\theta}^*, \mathbf{u}^*) = (\boldsymbol{\theta}_{(N)}, \mathbf{u}_{(N)})$
9: $\quad$ draw $\alpha \sim$ Uniform(0,1)
10: $\quad$ if $\alpha < \min\{1, e^{-(H(\boldsymbol{\theta}^*, \mathbf{u}^*) - H(\boldsymbol{\theta}, \mathbf{u}))}\}$
11: $\quad\quad \boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}^*$
12: $\quad$ else
13: $\quad\quad \boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}_{(i-1)}$
14: end for

Gibbs sampling of errors in x and y problem.

- Find the conditional distributions, and sample from m and x in a random order, to sample $p(m, x|\hat{x}, \hat{y})$, and marginalise over x. Here, $\hat{x} = 10$, $\hat{y} = 15$, and both have gaussian errors with unit variance.

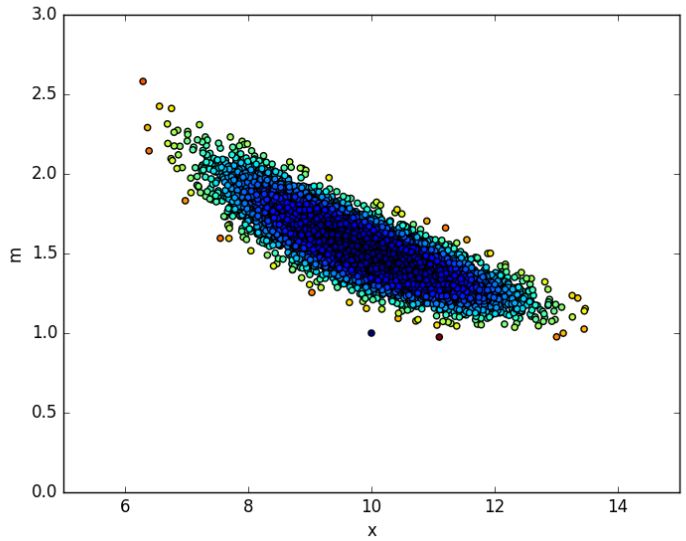


Figure: Gibbs sampling of the latent variable x, and the slope m.

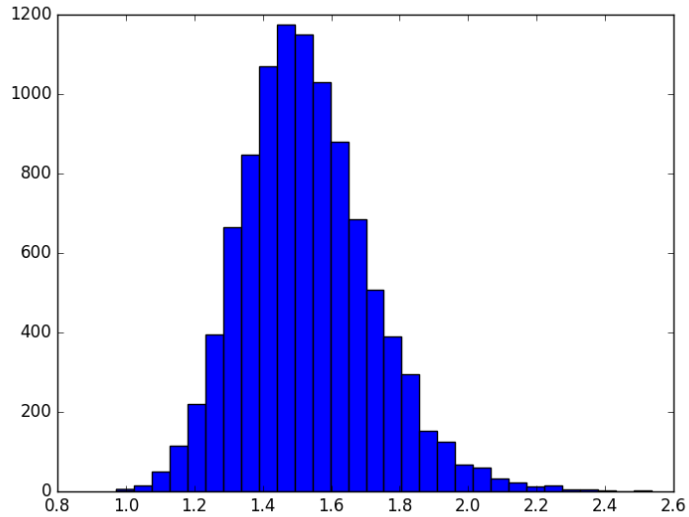


Figure: Gibbs sampling of the slope m.

Question: is this the most probable slope?

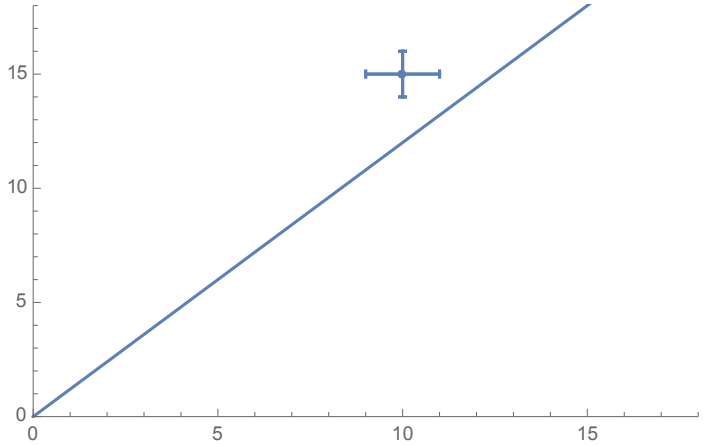


Figure: Noisy data

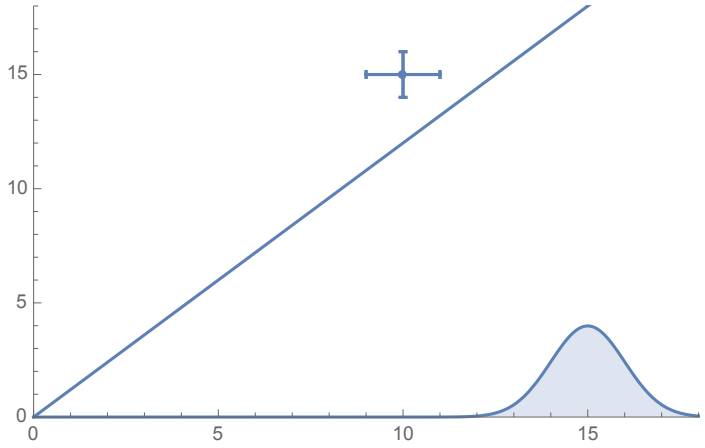


Figure: Yes! - there is a prior on x . . .

Sampling in *very* high dimensions

- **Metropolis-Hastings**: Good for small (maybe up to ~ 10) dimensions. Fails in very high dimensions, since it is very hard to devise a proposal distribution that does not always reject
- **Gibbs sampling**: can work if conditional distributions are known
- **Hamiltonian Monte Carlo**: can work well in very high dimensions, if model is differentiable (e.g. using Stan, jax, tensorflow)

Conclusions

- Standard way for non-trivial parameter inference is to sample the posterior, using MCMC
- More complex problems may be tackled with BHMs, usually with HMC, where autodifferentiation (jax, tensorflow probability) is amazing (see Junpeng Lao's lecture)
- If we can't construct an explicit likelihood, likelihood-free (or simulation-based) inference can be used (LFI/SBI).
    - *Data choice:* Massive data compression may well be necessary, e.g. using the MOPED algorithm, score compression, or IMNN (See Ben Wandelt's lecture)
- As forward modelling techniques, BHMs and LFI can include systematics and fully propagate errors.
- Neural Networks (see François Lanusse's lecture) can also be useful in approximating the distribution of samples, e.g. with Normalising Flows.

Some books for further reading

- D. Silvia & J. Skilling: Data Analysis: a Bayesian Tutorial (CUP) P. Saha: Principles of Data Analysis. (Capella Archive) `http://www.physik.uzh.ch/~psaha/pda/pda-a4.pdf`
- T. Loredo: Bayesian Inference in the Physical s `http://www.astro.cornell.edu/staff/loredo/bayes/`
- M. Hobson et al: Bayesian Methods in Cosmology (CUP)
- D. Mackay: Information Theory, Inference and Learning Algorithms. (CUP) `http://www.inference.phy.cam.ac.uk/itprnn/book.pdf`
- A. Gelman et al: Bayesian Data Analysis (CRC Press)

More details: straight line fit with errors in x and y

- Data: $\hat{x} = 10$, $\hat{y} = 15$.
- Choose some priors, and sample from the joint distribution of m and x:

$$p(m, x|\hat{x}, \hat{y}) \propto p(\hat{x}, \hat{y}|x, mx)\, p(x)\, p(m)$$

- For uniform priors, the joint distribution is (for HMC):

$$p(\hat{x}, \hat{y}|x, mx)\, p(x)\, p(m) \propto \exp\left[-\frac{(\hat{x} - x)^2}{2}\right] \exp\left[-\frac{(\hat{y} - mx)^2}{2}\right]$$

- For Gibbs, the conditional distributions are, for m given x:

$$p(m|x, \hat{x}, \hat{y}) \sim \mathcal{N}\left(\frac{\hat{y}}{x}, \frac{1}{x^2}\right)$$

- The conditional distribution of x given m is another normal distribution (in x now):

$$p(x|m, \hat{x}, \hat{y}) \sim \mathcal{N}\left(\frac{\hat{x} + m\hat{y}}{1 + m^2}, \frac{1}{1 + m^2}\right)$$