Data don't speak for themselves

A quick introduction to Bayesian statistics

João P. Faria

February 15, 2016

Outline

- main differences between frequentist and Bayesian statistics
  - problems with p-values, confidence intervals and null hypothesis testing
- the basic rules of probability theory
- how to assign probability distributions
  - the role of priors
  - the likelihood
- the simplest models in Bayesian statistics
  - linear regression
  - beta-binomial model
  - hierarchical models

Outline

- main differences between frequentist and Bayesian statistics
  - ~~problems with p-values, confidence intervals and null hypothesis testing~~
- the basic rules of probability theory
- how to assign probability distributions
  - the role of priors
  - the likelihood
- the simplest models in Bayesian statistics
  - linear regression
  - ~~beta-binomial model~~
  - hierarchical models
- MCMC

Q/A

Introduction

brief historical sketch

- statistical inference was invented because of astronomy

  (this claim is not based on an in-depth search)

brief historical sketch

- statistical inference was invented because of astronomy
  (this claim is not based on an in-depth search)
- Tycho Brahe, Galileo, Legendre, Laplace, Gauss
  used and developed statistical methods

brief historical sketch

- statistical inference was invented because of astronomy
  (this claim is not based on an in-depth search)
- Tycho Brahe, Galileo, Legendre, Laplace, Gauss
  used and developed statistical methods
- $\sim$ 20th century, astronomers focused on least-squares
  techniques, and heuristic procedures

brief historical sketch

- statistical inference was invented because of astronomy

  (this claim is not based on an in-depth search)
- Tycho Brahe, Galileo, Legendre, Laplace, Gauss used and developed statistical methods
- $\sim$ 20th century, astronomers focused on least-squares techniques, and heuristic procedures
- "astrostatistics" emerged in the late 1990s – collaborations between astronomers and statisticians

  names like Babu, Feigelson, Gregory, Hobson

brief historical sketch

- statistical inference was invented because of astronomy
  (this claim is not based on an in-depth search)
- Tycho Brahe, Galileo, Legendre, Laplace, Gauss
  used and developed statistical methods
- $\sim$ 20th century, astronomers focused on least-squares
  techniques, and heuristic procedures
- "astrostatistics" emerged in the late 1990s – collaborations
  between astronomers and statisticians
  names like Babu, Feigelson, Gregory, Hobson
- statistics today is mostly Bayesian
  (many) astronomers are (very) sceptical of (sophisticated) statistics

frequentist and Bayesian

Consider n independent measurements of the same quantity, under identical conditions

Calculate their mean $\bar{x}$ and standard deviation σ

frequentist and Bayesian

Consider n independent measurements of the same quantity, under identical conditions

Calculate their mean $\bar{x}$ and standard deviation σ

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

frequentist and Bayesian

Consider n independent measurements of the same quantity, under identical conditions

Calculate their mean $\bar{x}$ and standard deviation σ

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

what you think this means:

$$p(\bar{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$$

frequentist and Bayesian

Consider n independent measurements of the same quantity, under identical conditions

Calculate their mean $\bar{x}$ and standard deviation σ

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

what you think this means:

$$p(\bar{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$$

what it actually means:

$$p(\mu - \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + \frac{\sigma}{\sqrt{n}}) = 68\%$$

frequentist and Bayesian

Consider n independent measurements of the same quantity, under identical conditions

Calculate their mean $\bar{x}$ and standard deviation σ

$$\mu = \bar{x} \pm \frac{\sigma}{\sqrt{n}}$$

what you want it to mean:

$$p(\bar{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}) = 68\%$$

what it actually means:

$$p(\mu - \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + \frac{\sigma}{\sqrt{n}}) = 68\%$$

frequentist and Bayesian

Here is the main difference:

- Frequentists assign a probability to what is *random*
- Bayesians assign a probability to what is *unknown*

Bayesian statistics

the rules of probability

the rules of probability density functions

the rules of probability density functions

normalisation

$$1 = \int p(a) da$$

factorisation of joint pdfs

$$p(a, b) = p(a)\, p(b|a)$$
$$p(a, b) = p(b)\, p(a|b)$$

Bayes' theorem

$$p(a|b) = \frac{p(a)\, p(b|a)}{p(b)}$$

marginalisation

$$p(a) = \int p(a, b) db$$

the rules of probability density functions

normalisation

$$1 = \int p(a|c)da$$

factorisation of joint pdfs

$$p(a, b|c) = p(a|c)\, p(b|a, c)$$
$$p(a, b|c) = p(b|c)\, p(a|b, c)$$

Bayes' theorem

$$p(a|b, c) = \frac{p(a|c)\, p(b|a, c)}{p(b|c)}$$

marginalisation

$$p(a|c) = \int p(a, b|c)db$$

Bayes' theorem

$$p(a|b,c) = \frac{p(a|c)\,p(b|a,c)}{p(b|c)}$$

Bayes' theorem

$$p(\theta|b,c) = \frac{p(\theta|c)\,p(b|\theta,c)}{p(b|c)}$$

Bayes' theorem

$$p(\theta|D, c) = \frac{p(\theta|c)\, p(D|\theta, c)}{p(D|c)}$$

Bayes' theorem

$$p(\theta|D,\mathcal{I}) = \frac{p(\theta|\mathcal{I})\,p(D|\theta,\mathcal{I})}{p(D|\mathcal{I})}$$

Bayes' theorem

$$p(\theta|D,\mathcal{I}) = \frac{p(\theta|\mathcal{I})\,p(D|\theta,\mathcal{I})}{p(D|\mathcal{I})}$$

θ - parameters

D - data

$\mathcal{I}$ - assumed information (or hypotheses)

Bayes' theorem

$$p(\theta|D,\mathcal{I}) = \frac{p(\theta|\mathcal{I})\, p(D|\theta,\mathcal{I})}{p(D|\mathcal{I})}$$

θ - parameters

D - data

$\mathcal{I}$ - assumed information (or hypotheses)

$$p(\bar{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}) = 68\% \quad p(\mu - \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + \frac{\sigma}{\sqrt{n}}) = 68\%$$

Q/A

rules of probability

Assigning probabilities

assigning prior probabilities

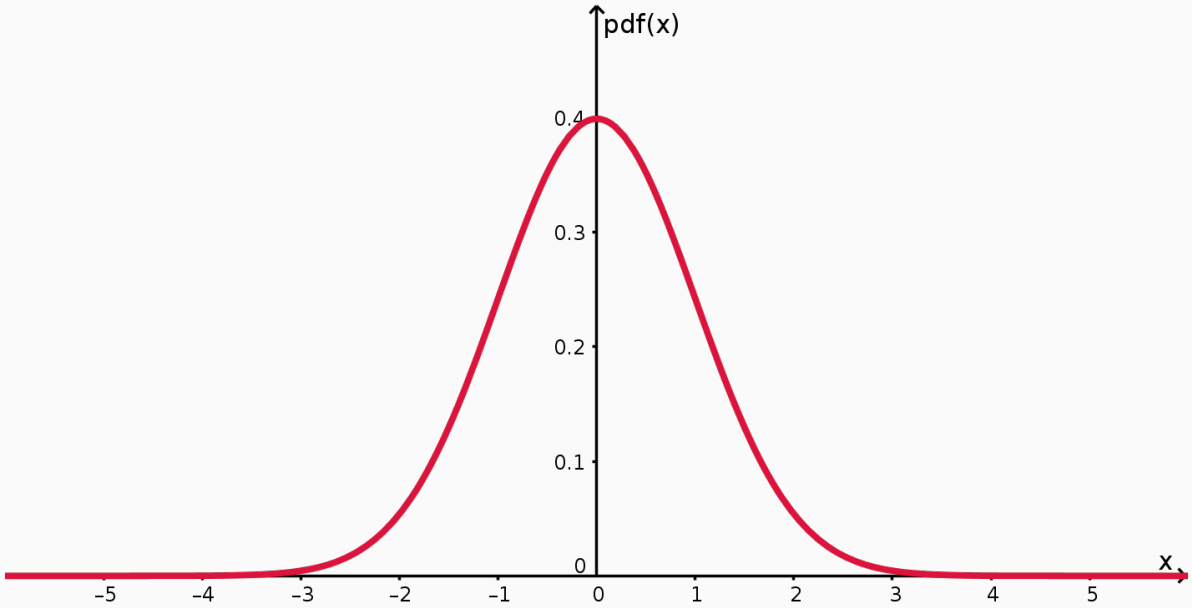
$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\textit{evidence}}$$

assigning prior probabilities

$$\text{posterior} = \frac{p(\theta|\mathcal{I}) \times \text{likelihood}}{\textit{evidence}}$$

assigning prior probabilities

the Gaussian distribution

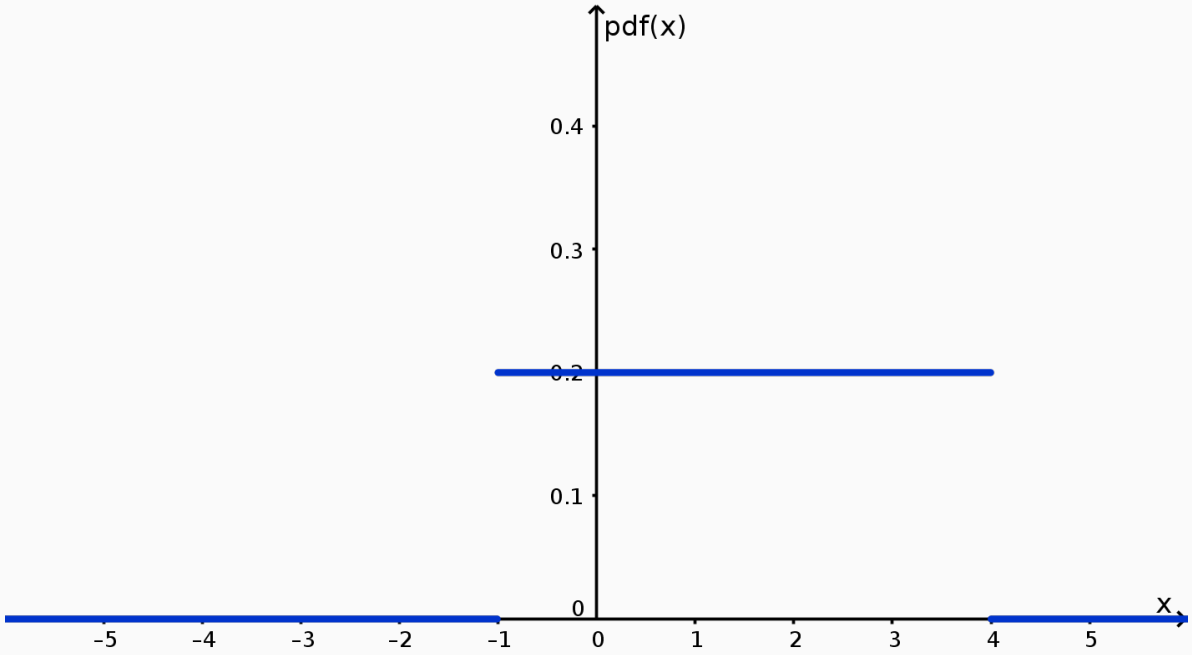


$$x \sim \mathcal{N}(\mu, \sigma^2)$$

$$\text{pdf}(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

assigning prior probabilities

the uniform distribution

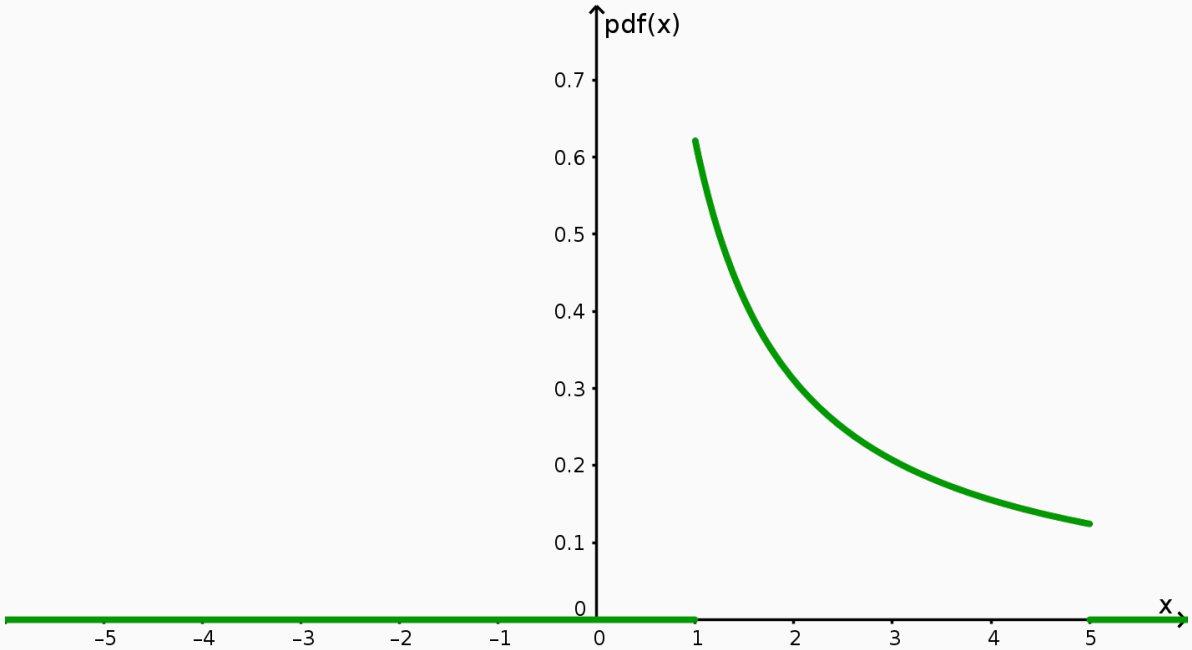


$$x \sim \mathcal{U}(a, b)$$

$$\text{pdf}(x|a, b) = \frac{1}{b - a}$$

assigning prior probabilities

the reciprocal (Jeffreys') distribution

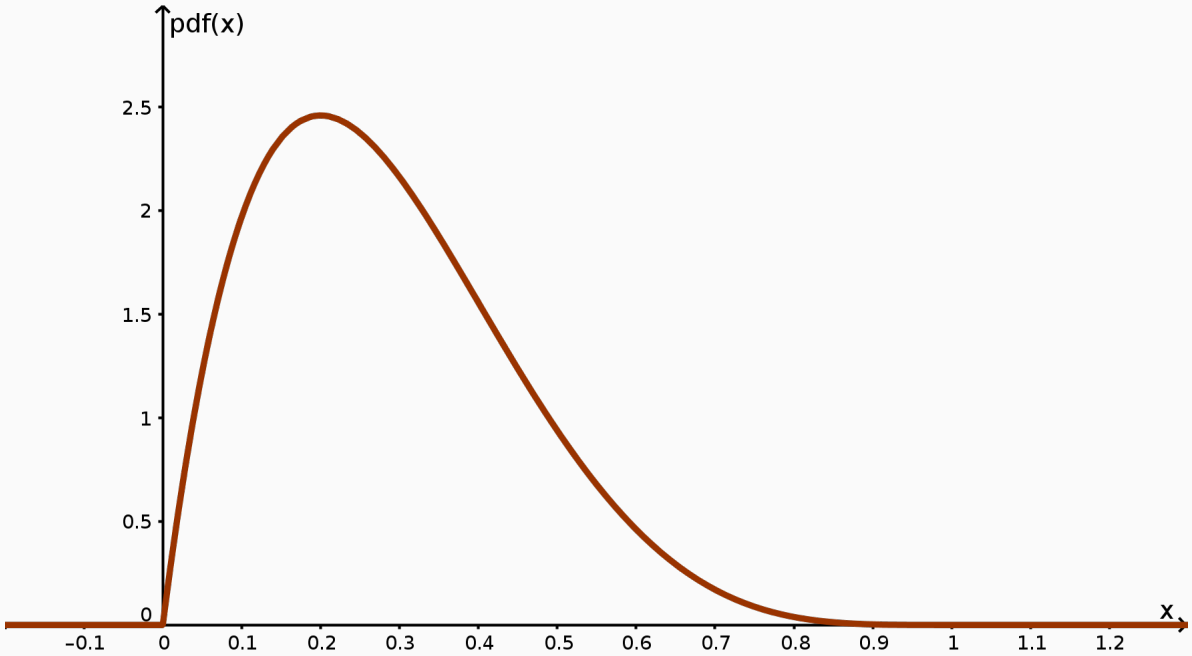


$$x \sim \mathcal{J}(a, b)$$

$$\text{pdf}(x|a, b) = \frac{1}{x\left[\log(b) - \log(a)\right]}$$

assigning prior probabilities

the Beta distribution

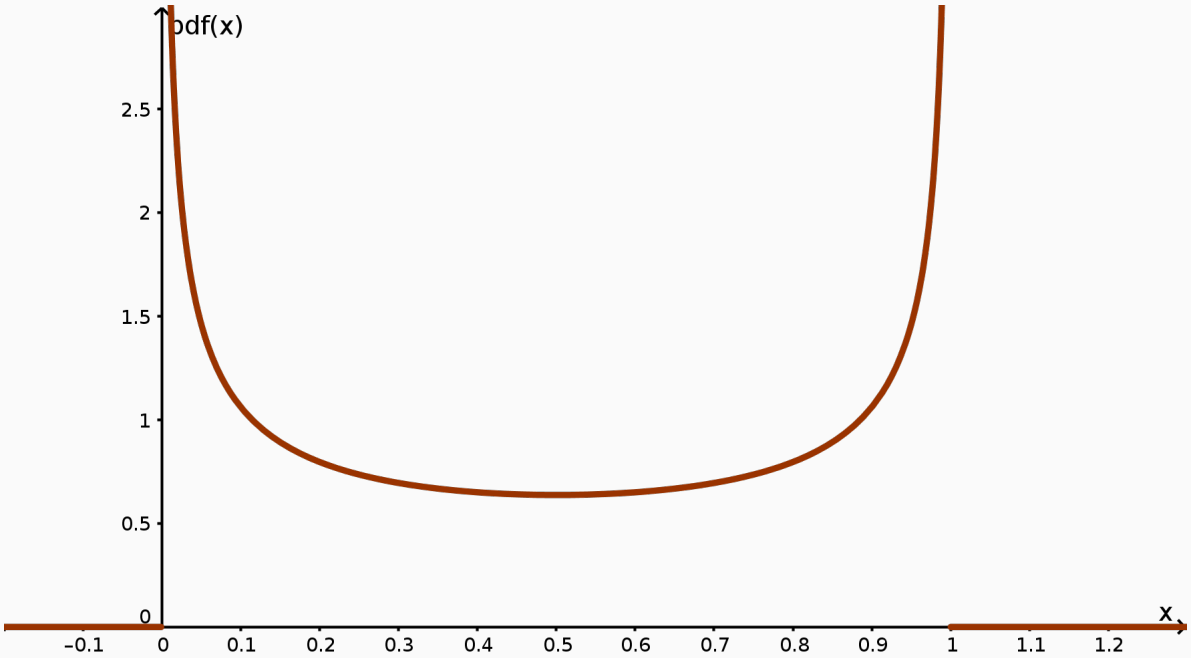


$$x \sim \mathcal{B}(\alpha, \beta)$$

$$\text{pdf}(x|a,b) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}$$

assigning prior probabilities

the Beta distribution



$$x \sim \mathcal{B}(\alpha, \beta)$$

$$\text{pdf}(x|a,b) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}$$

assigning prior probabilities

Use your Python:

```
from scipy.stats import *

norm(loc=0, scale=1).pdf(0.1)       # evaluate pdf at x=0.1
norm(loc=0.5, scale=0.2).rvs(100)   # get 100 random samples

uniform().logpdf(0.5)               # evaluate log pdf at x=0.5

beta(a=0.5, b=0.5).interval(1)      # the support (0, 1)
```

there are tons of distributions in Scipy!!

assigning prior probabilities

If you have two parameters and they are independent

$$p(\theta_1, \theta_2|\mathcal{I}) = p(\theta_1|\mathcal{I})\, p(\theta_2|\mathcal{I})$$

$$\log p(\theta_1, \theta_2|\mathcal{I}) = \log p(\theta_1|\mathcal{I}) + \log p(\theta_2|\mathcal{I})$$

assigning prior probabilities

If you have N parameters and they are independent

$$p(\theta_1, \ldots, \theta_N|\mathcal{I}) = \prod p(\theta_i|\mathcal{I})$$

$$\log p(\theta_1, \ldots, \theta_N|\mathcal{I}) = \sum \log p(\theta_i|\mathcal{I})$$

What is “prior information”

A mix of

- substantive knowledge,
- scientific conjectures,
- statistical properties,
- analytical convenience,
- disciplinary tradition,
- computational tractability.

Example: exoplanets

The radial-velocity signal produced by an orbiting planet depends

- on the orbital period $p(P|\mathcal{I}) \sim \mathcal{J}(1, 1000)$ days
- on the semi-amplitude $p(K|\mathcal{I}) \sim \mathcal{J}(0.1, 100)$ ms^{-1}
- on the eccentricity $p(e|\mathcal{I}) \sim \mathcal{U}(0, 1)$
- on the argument of periastron $p(\omega|\mathcal{I}) \sim \mathcal{U}(0, 2\pi)$
- on the phase of periastron $p(\chi|\mathcal{I}) \sim \mathcal{U}(0, 2\pi)$

Example: exoplanets

The radial-velocity signal produced by an orbiting planet depends

- on the orbital period `reciprocal(a=1,b=1000).pdf(P)`
- on the semi-amplitude `reciprocal(a=0.1,b=100).pdf(K)`
- on the eccentricity `uniform().pdf(e)`
- on the argument of periastron `uniform(scale=2*pi).pdf(w)`
- on the phase of periastron `uniform(scale=2*pi).pdf(X)`

Assigning prior probabilities

Take-home messages:

- take time to think about what you know
  (order of magnitude, likely value, scale *vs* position parameter)
- use simple distributions when possible
- start with small priors;
  increase them until you feel comfortable writing it in the paper
- use what other people use, if it makes sense
- see if the conclusions depend strongly on the prior
  and write that in the paper!

- in your MCMC, set the log likelihood to 1 and check if the priors are ok

Likelihood (aka the sampling distribution)

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\textit{evidence}}$$

Likelihood (aka the sampling distribution)

$$\text{posterior} = \frac{\text{prior} \times p(D|\theta, \mathcal{I})}{\textit{evidence}}$$

Likelihood

$\mathcal{I}$:

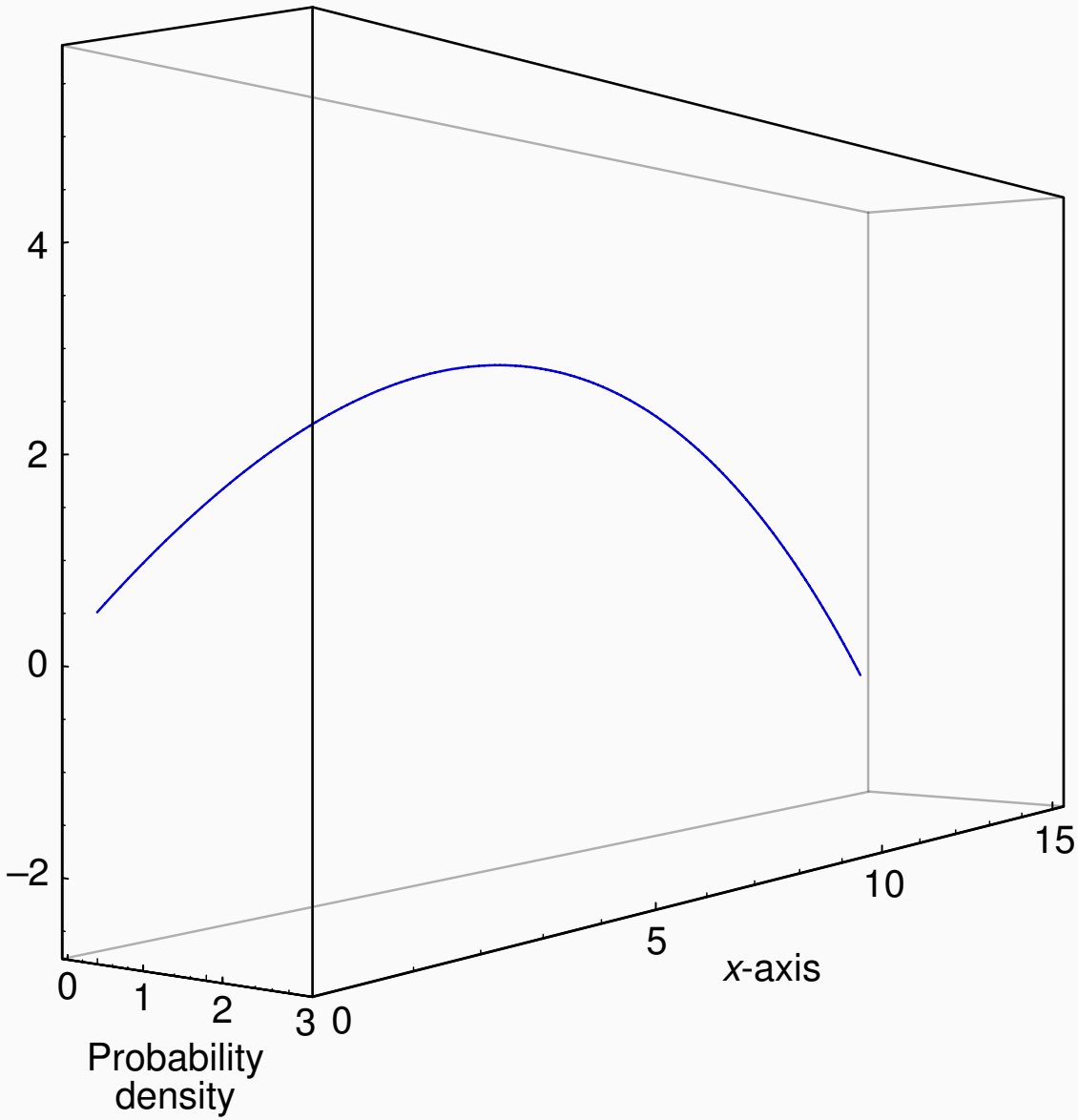
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

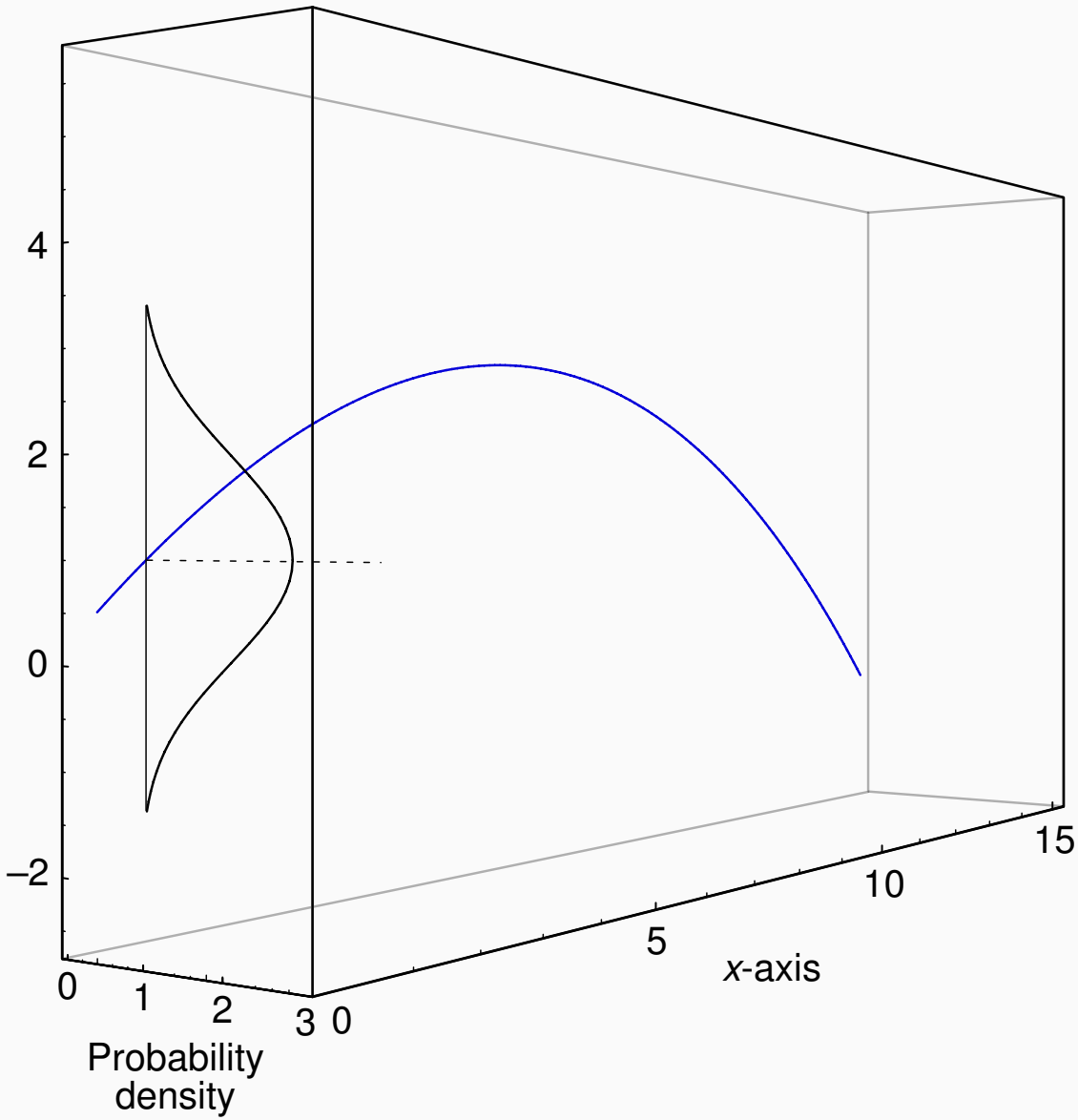
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

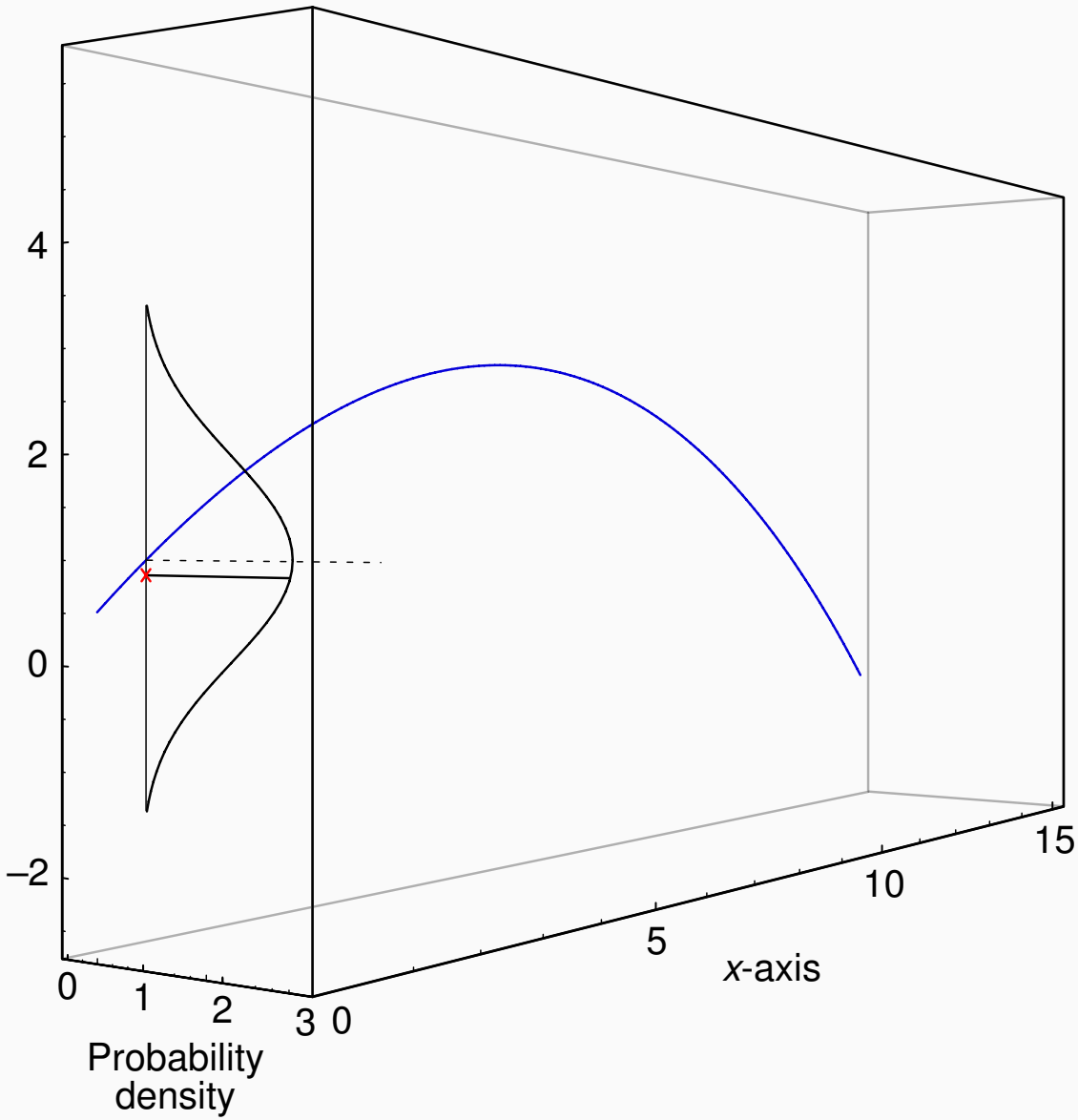
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

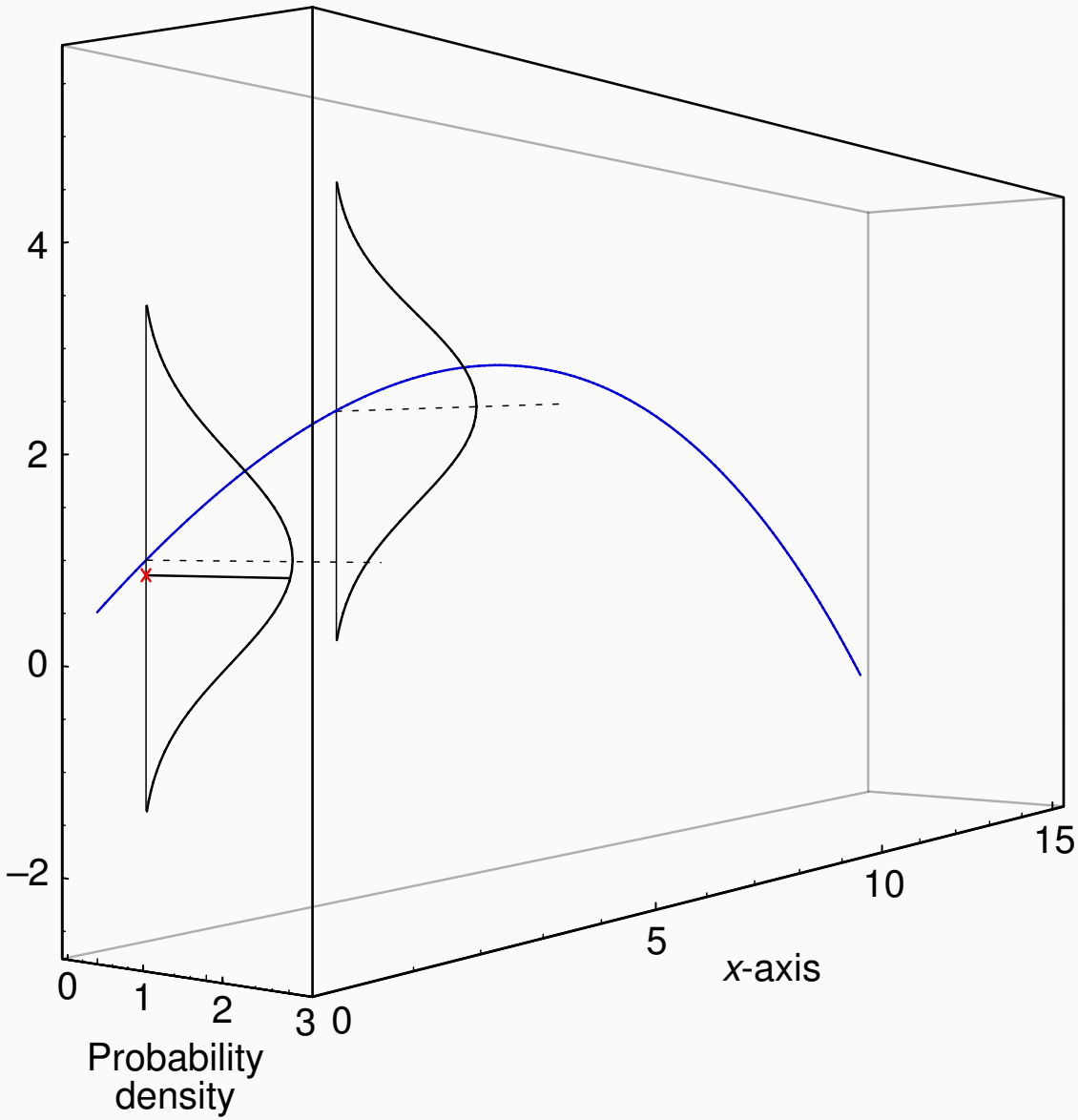
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

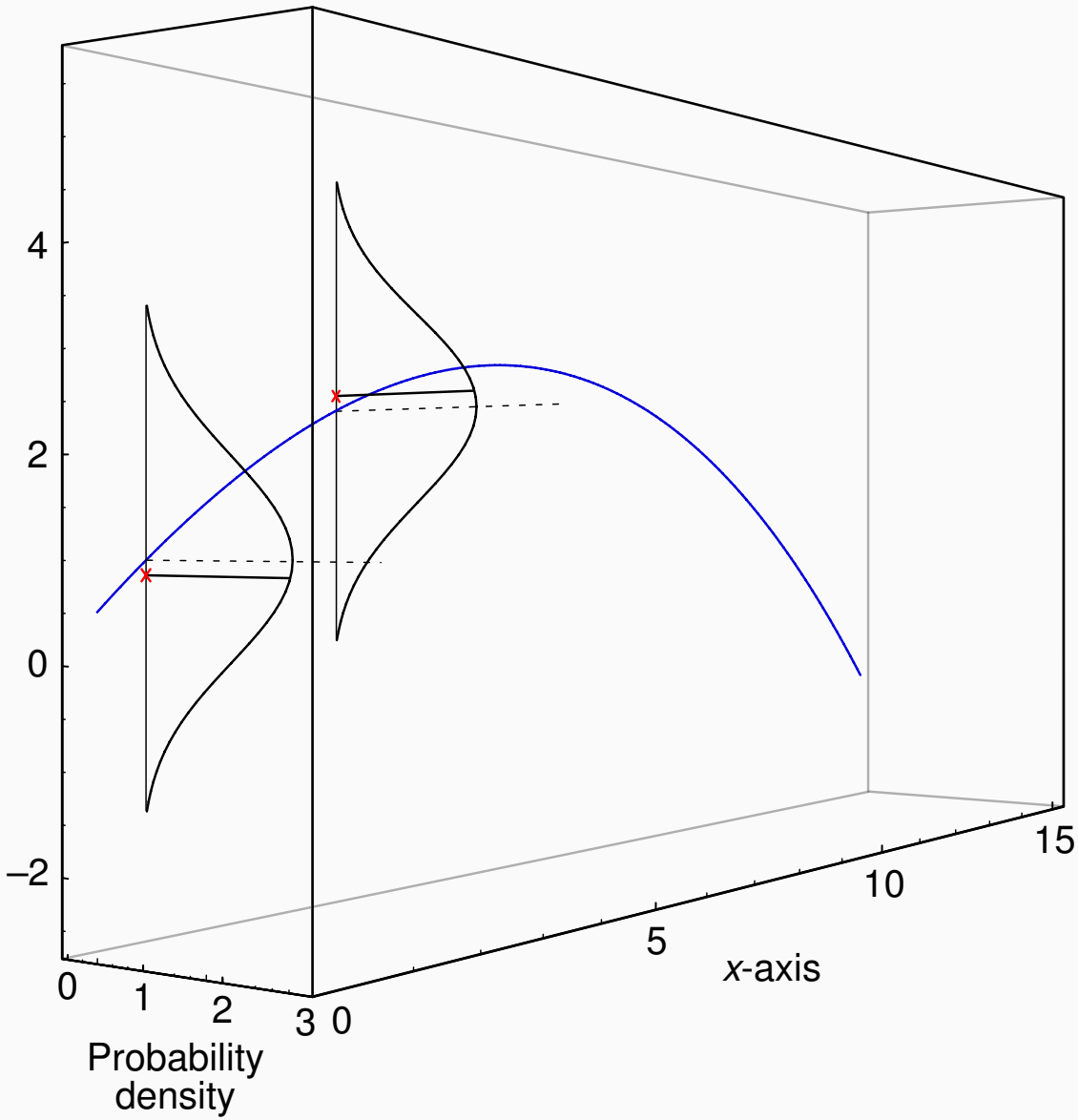
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

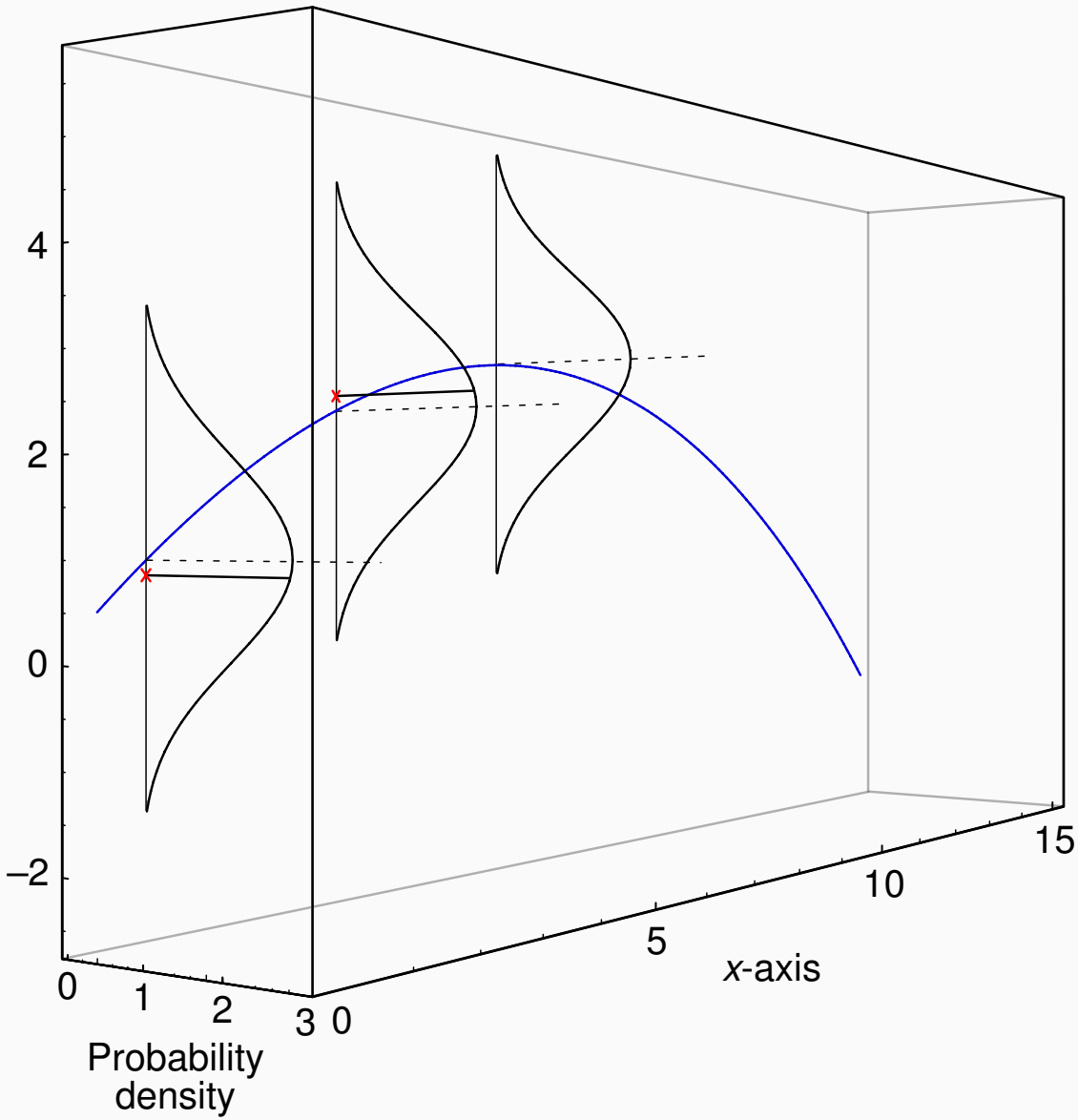
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

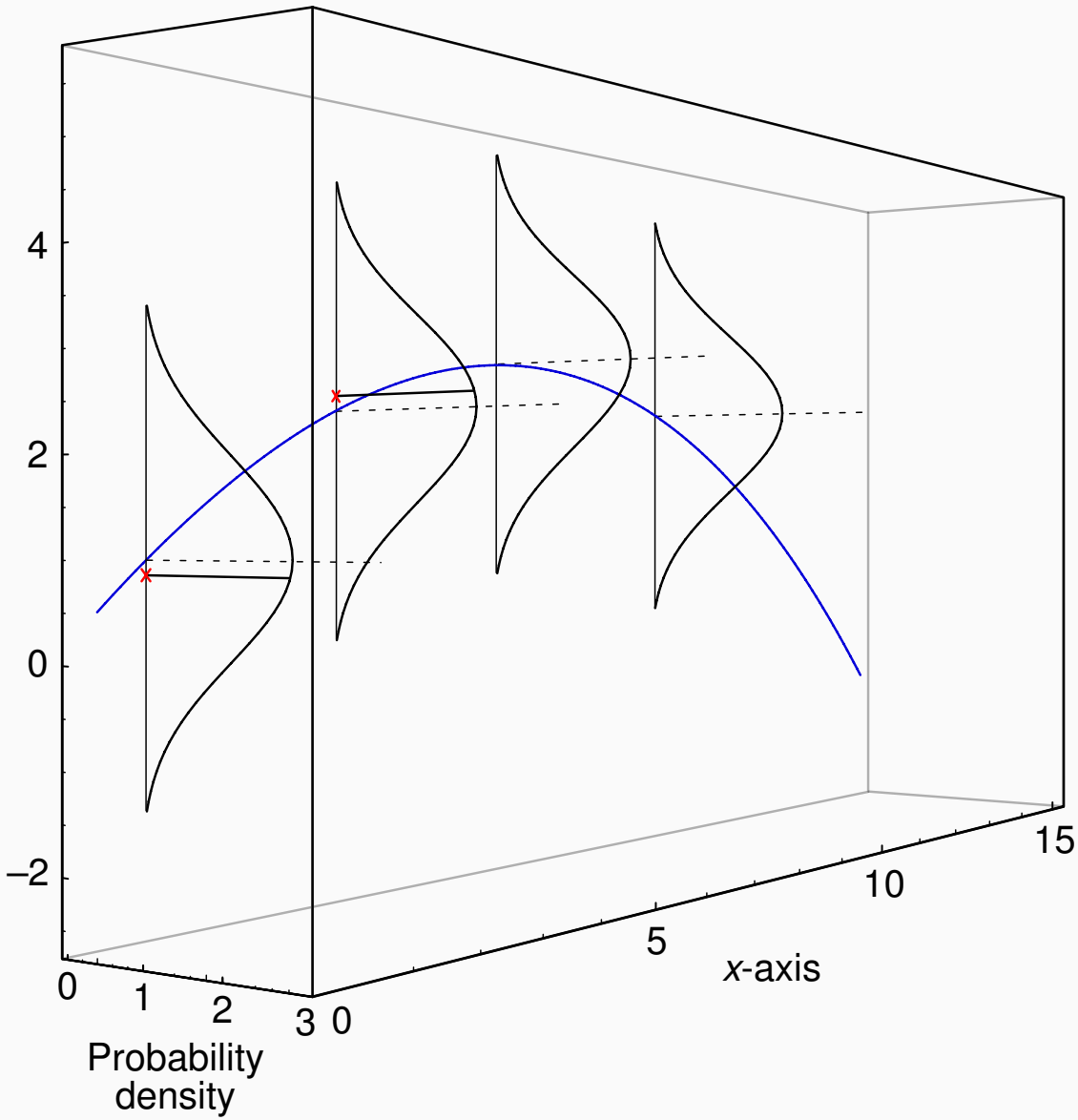
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

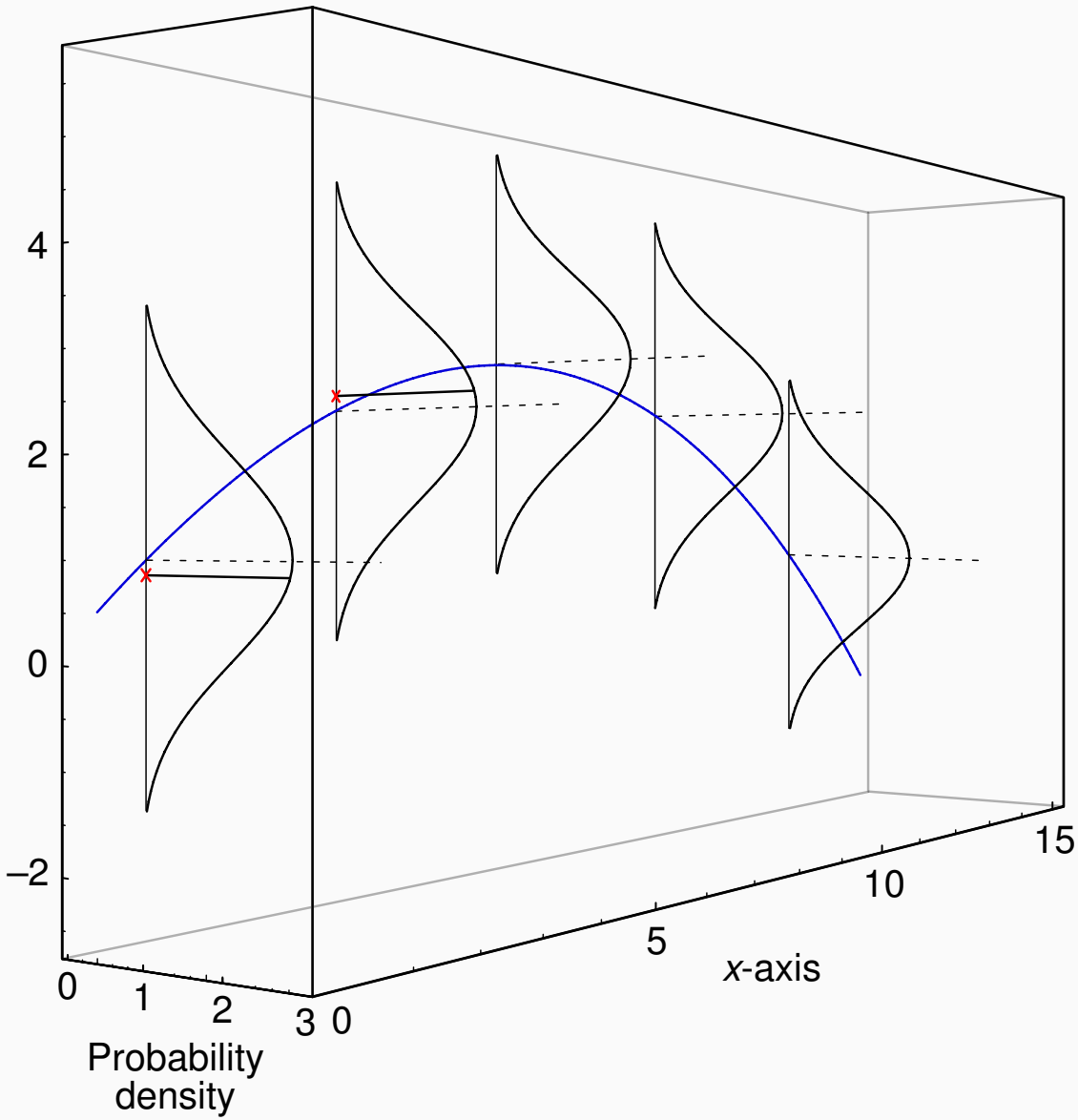
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

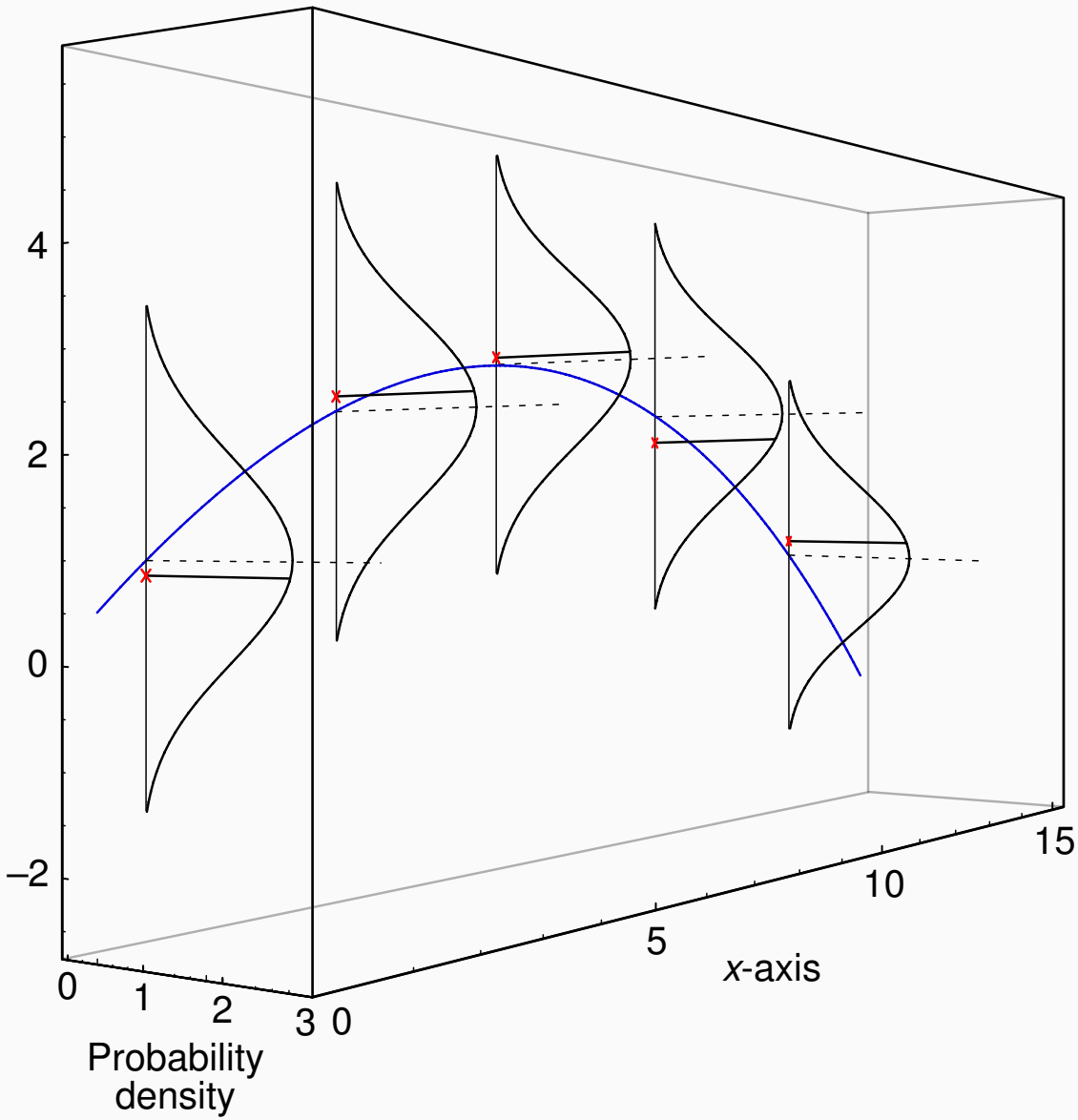
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

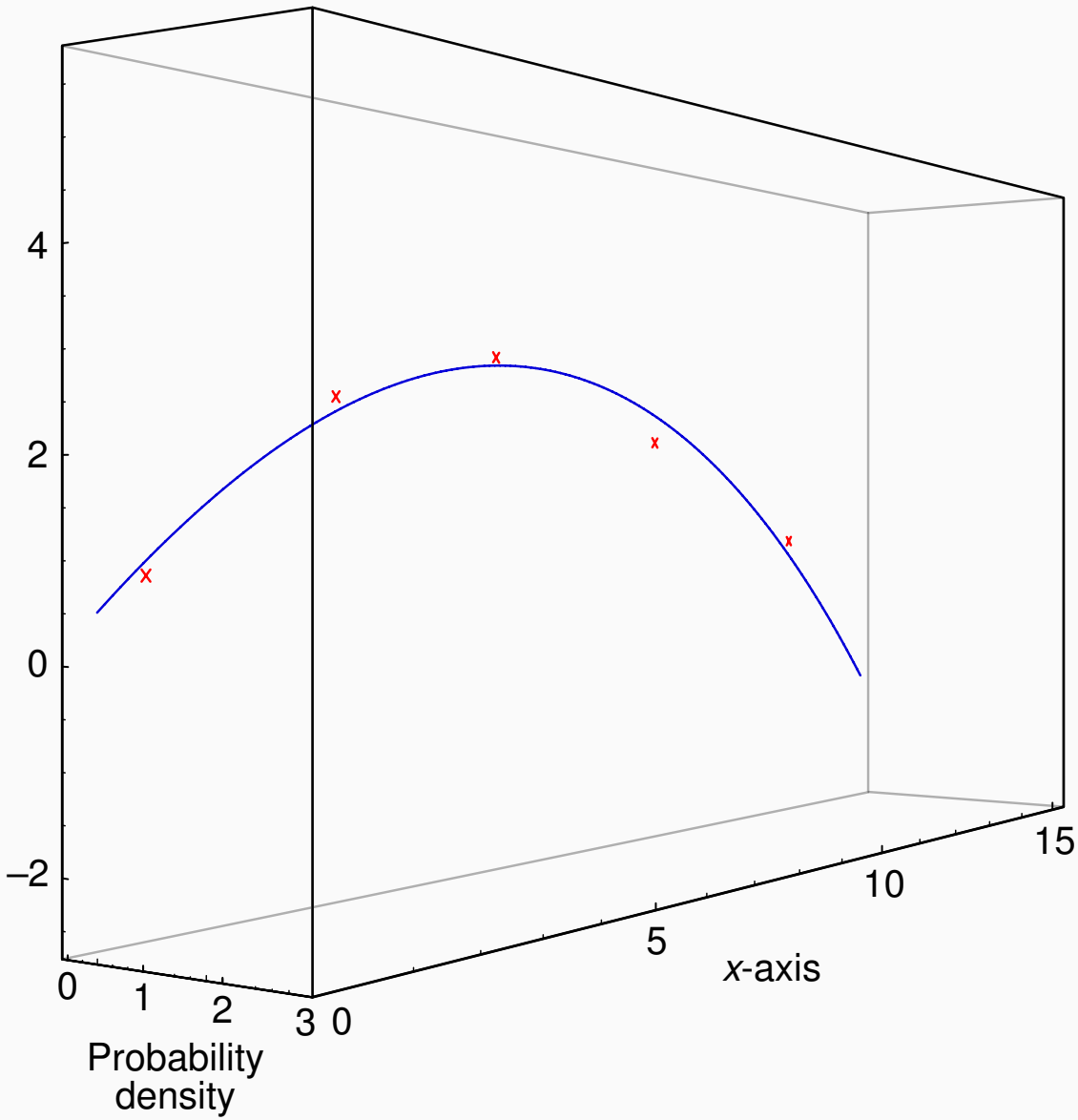
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

σ

$\theta = \{A_1, A_2, A_3\}$

$p(D|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma)$



Likelihood

$\mathcal{I}$:

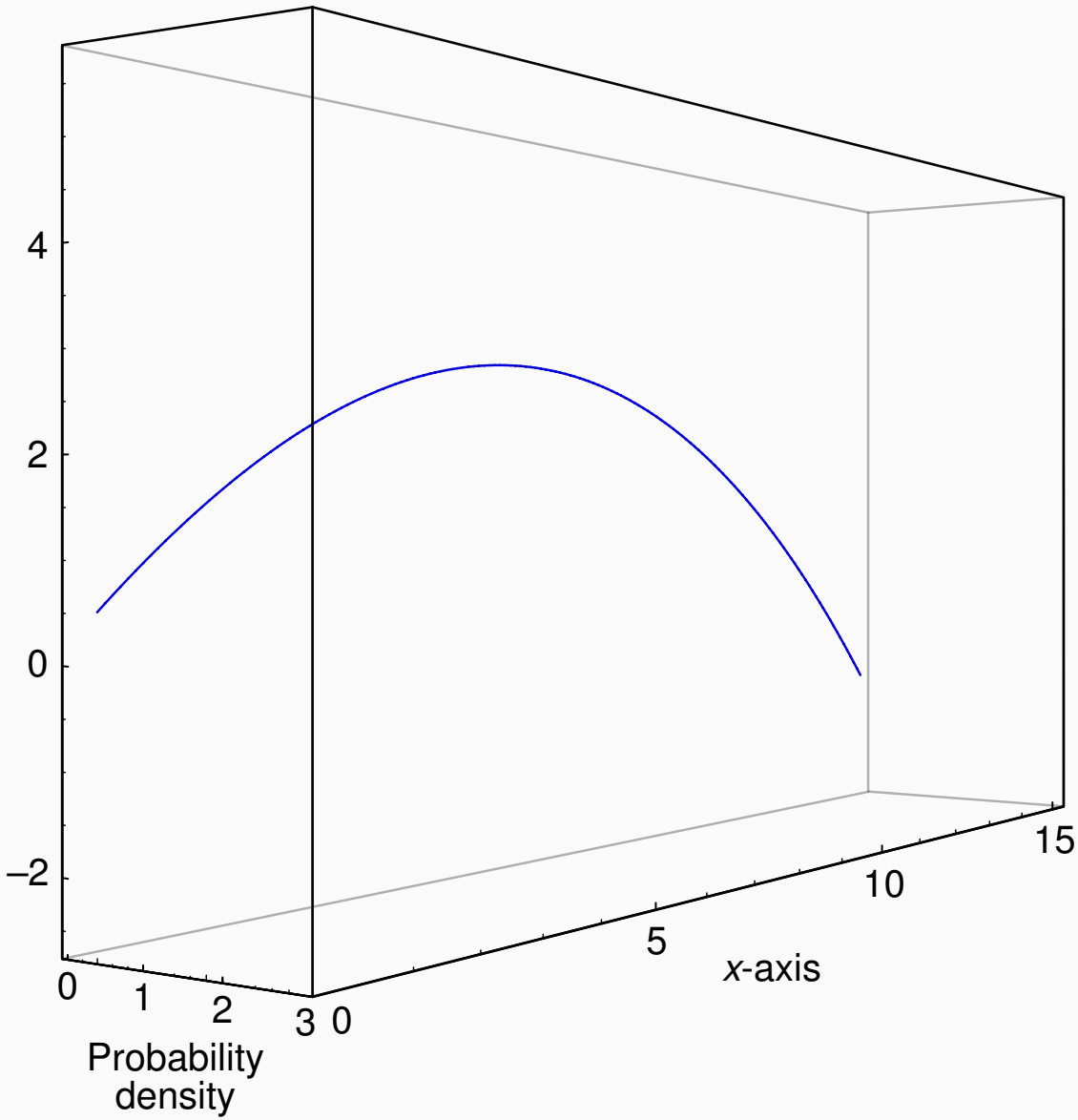
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

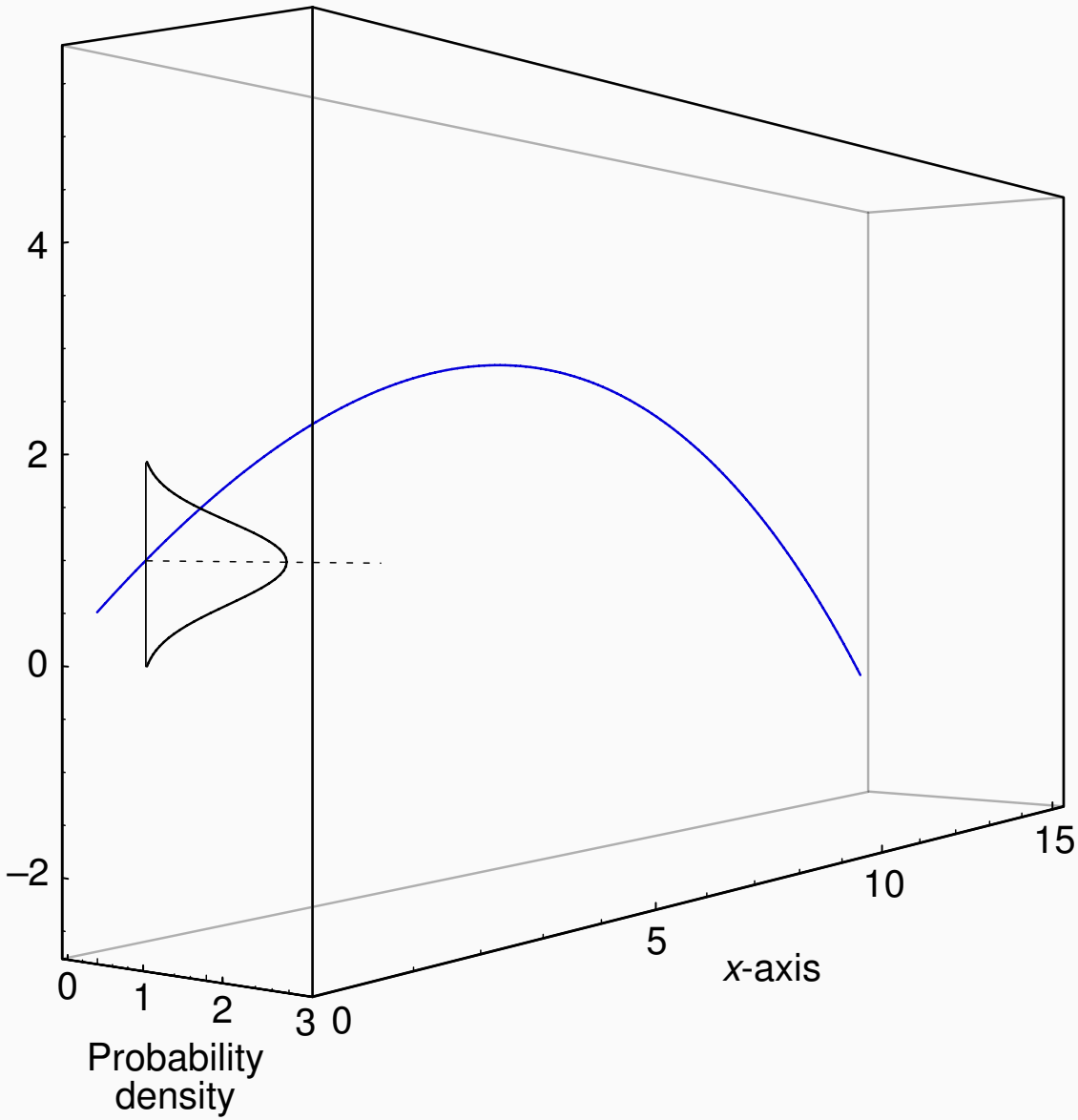
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

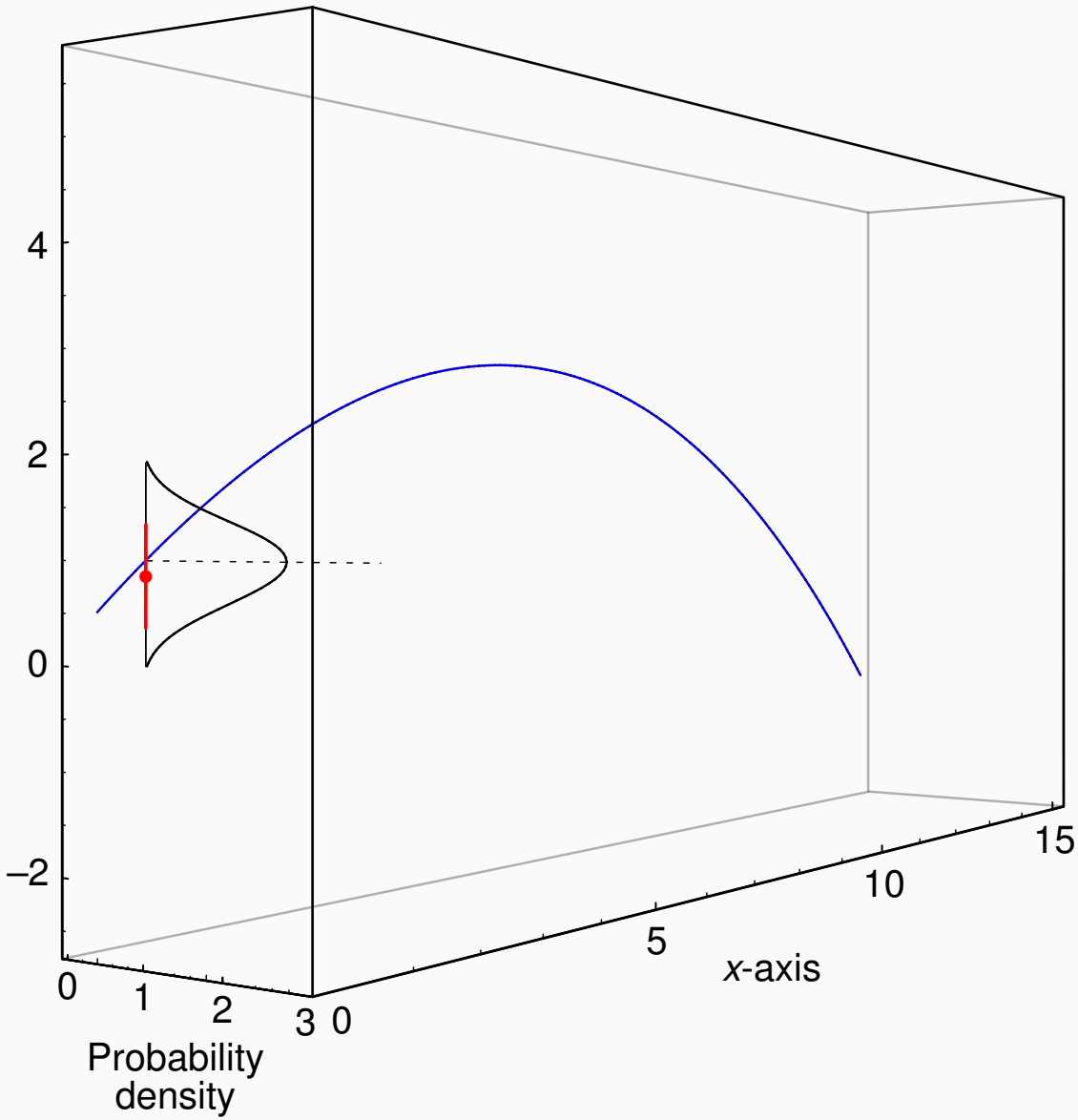
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

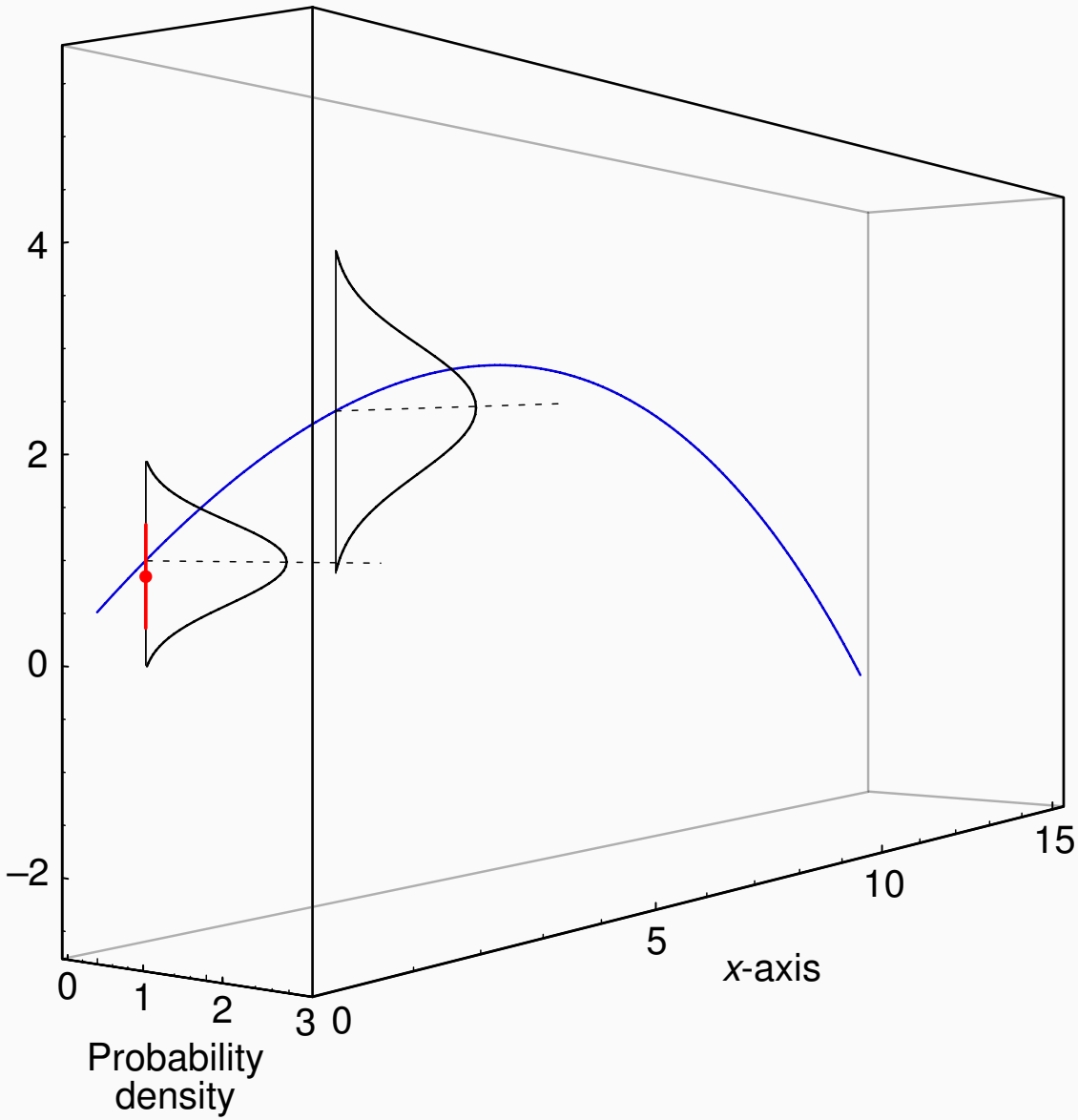
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

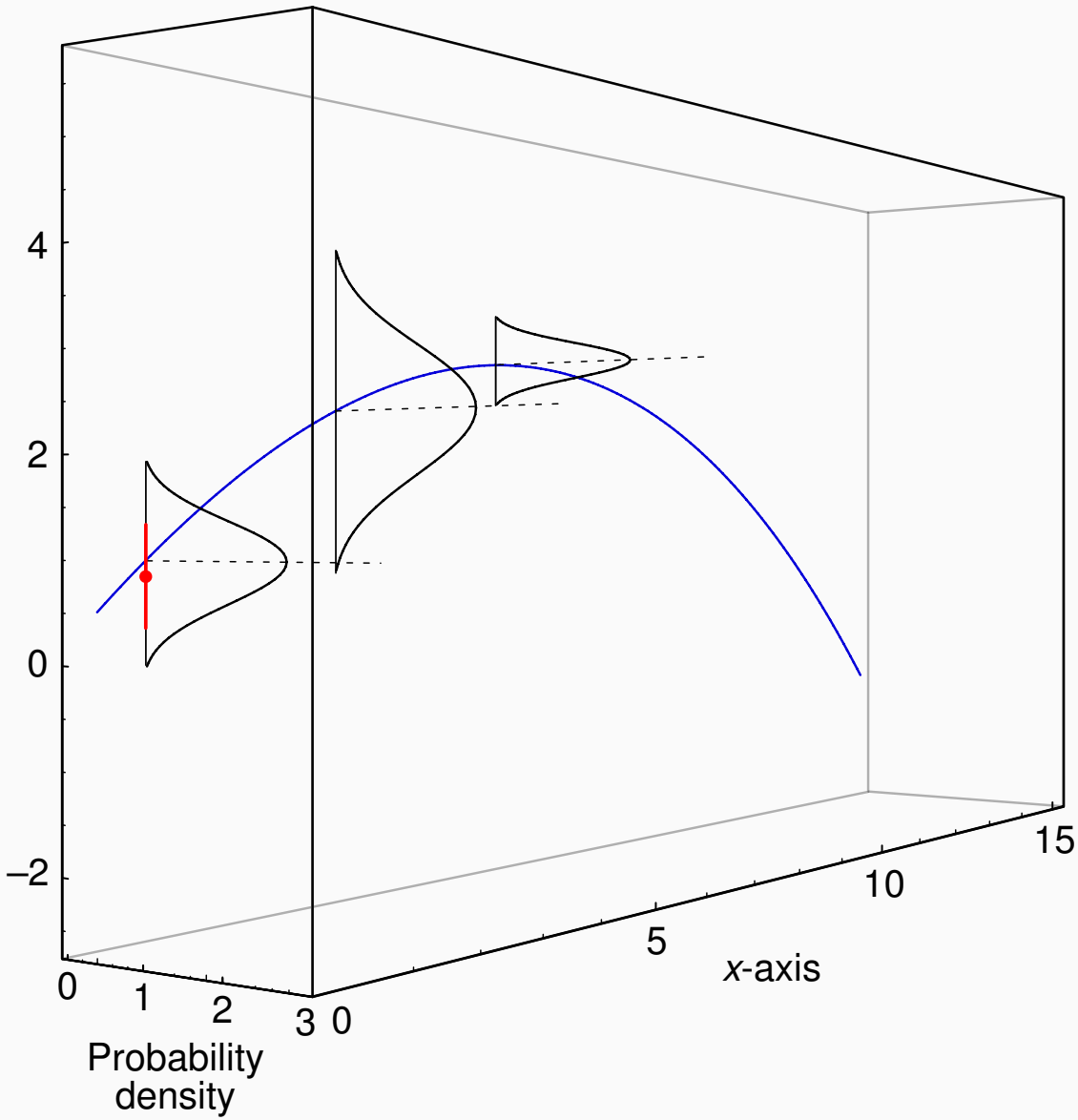
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

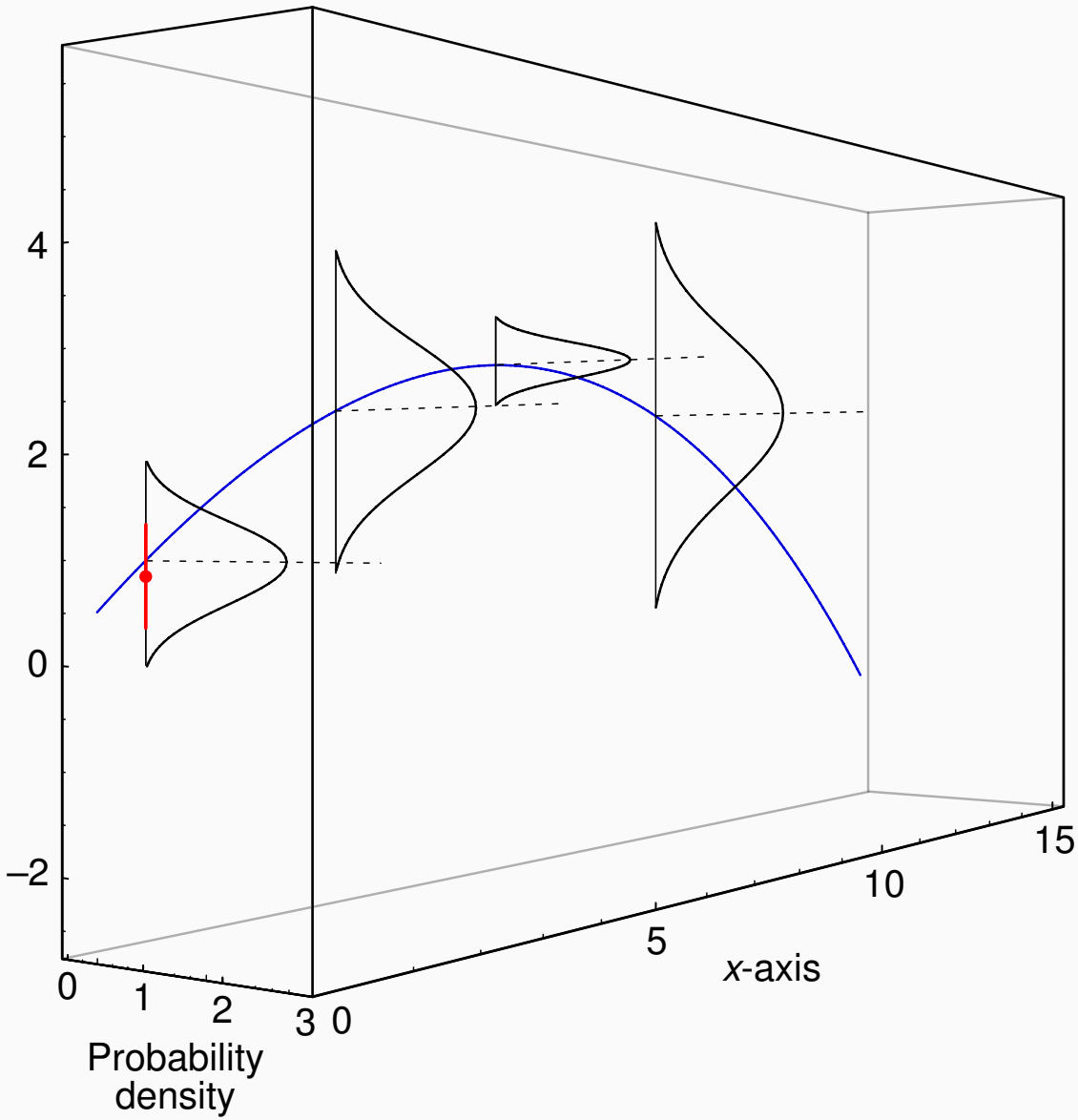
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

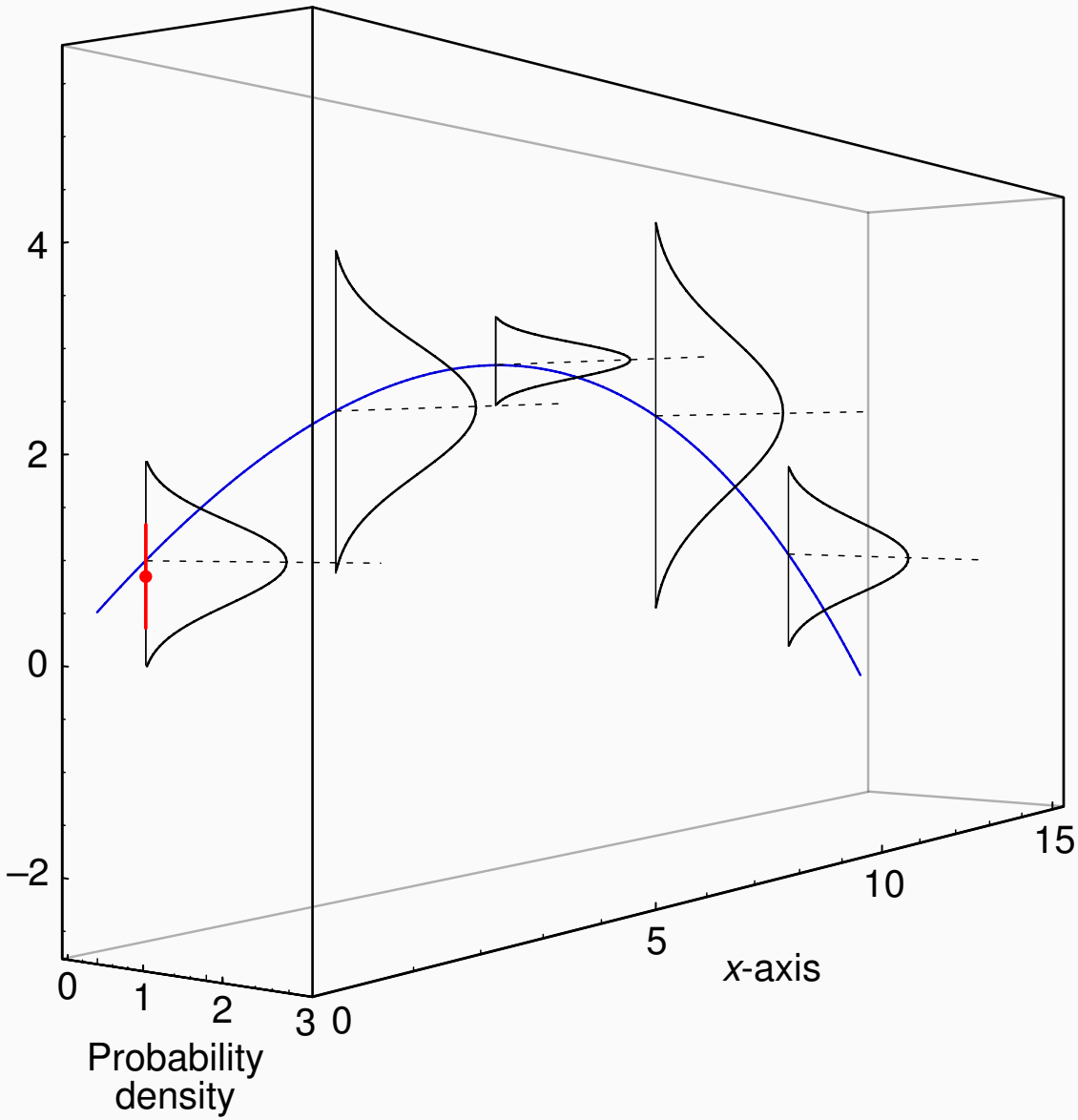
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

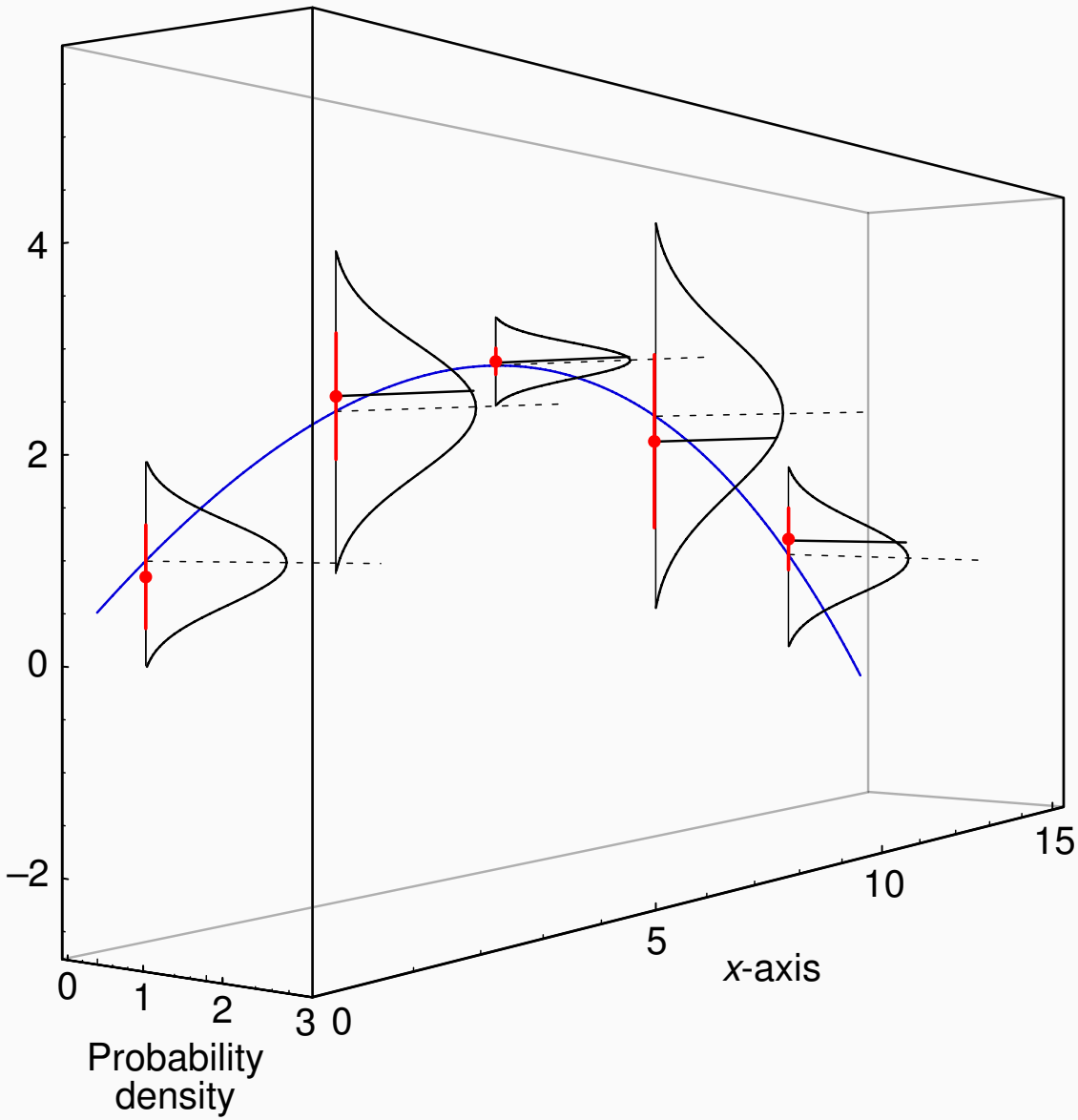
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

$\mathcal{I}$:

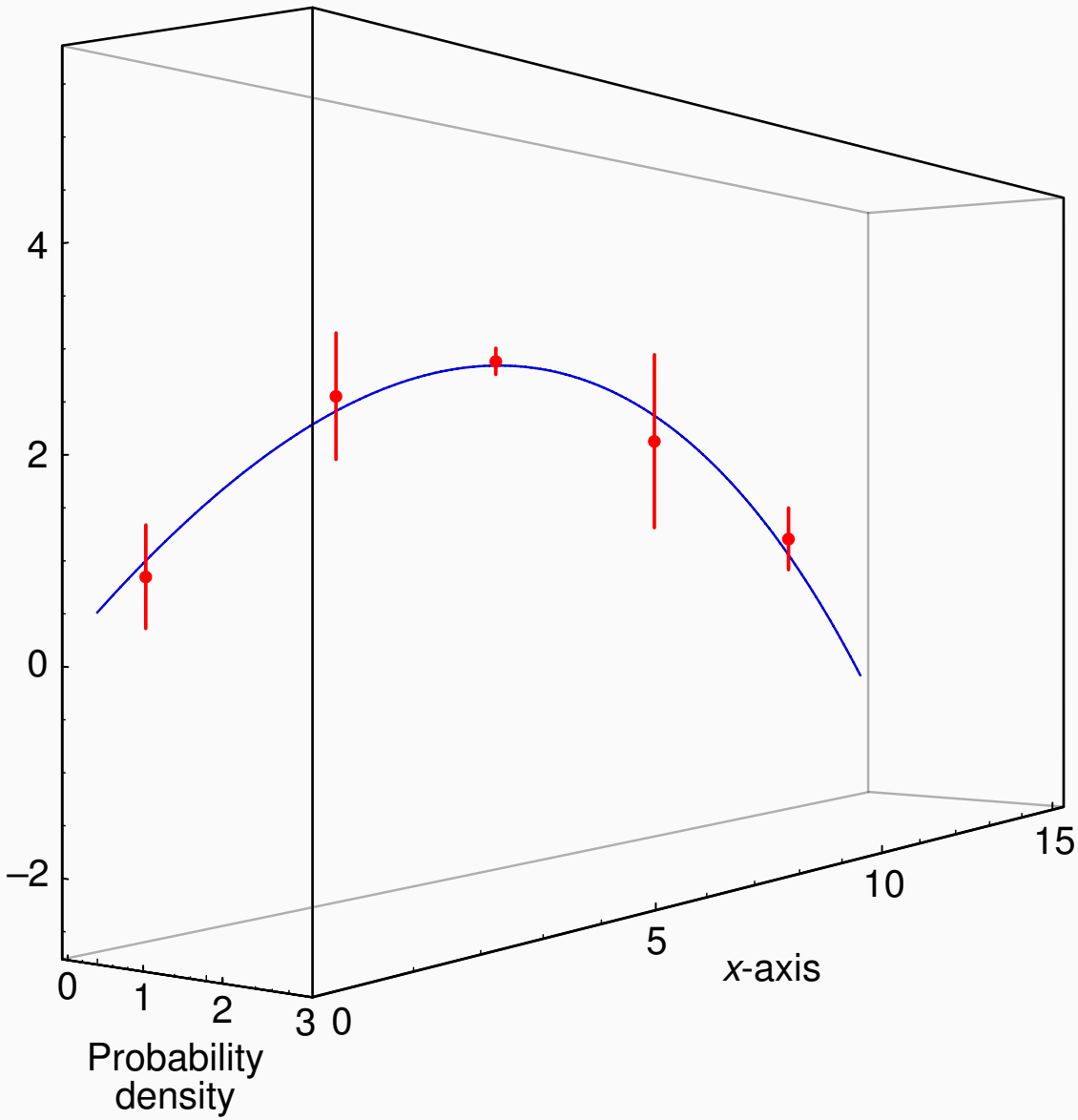
$f(x) = A_1 + A_2\,x + A_3\,x^2$

$x_1, \ldots, x_n$

$\sigma_1, \ldots, \sigma_n$

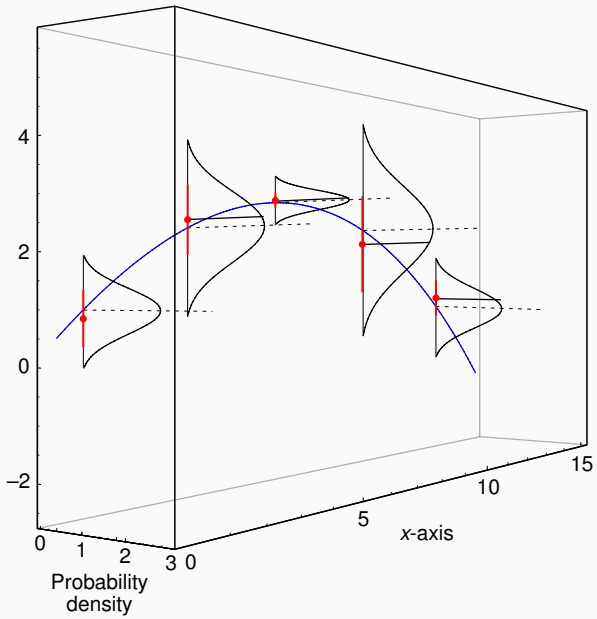
$\theta = \{A_1, A_2, A_3\}$

$p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(f(x), \sigma_i)$



Likelihood

The likelihood of all the data
is the product of the likelihood
for each data point



$$p(D|\theta, I) = \prod_{i=1}^{N} p(D_i|\theta, I)$$

$$= \prod_{i=1}^{N} \mathcal{N}_{\text{pdf}}(d_i|f(x_i), \sigma_i)$$

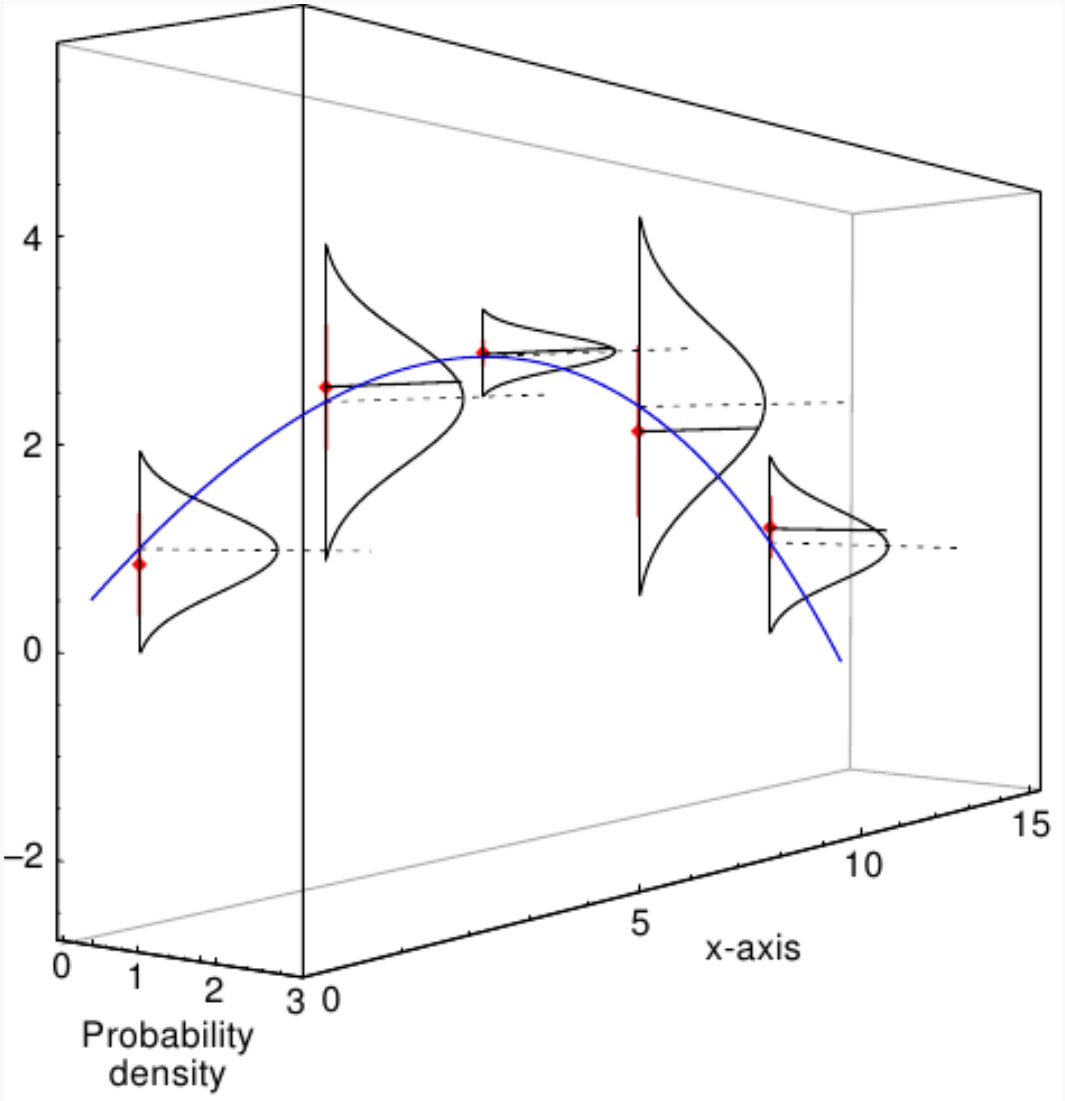
$$= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{ -\frac{[d_i - f(x_i)]^2}{2\sigma_i^2} \right\}$$

$$= (2\pi)^{-N/2} \left(\prod_{i=1}^{N} \frac{1}{\sigma_i} \right) \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \frac{[d_i - f(x_i)]^2}{2\sigma_i^2} \right\}$$

a product of many small numbers is very small: use logarithms!

Likelihood

the likelihood for a fixed dataset is a function of the parameters



fixed D,

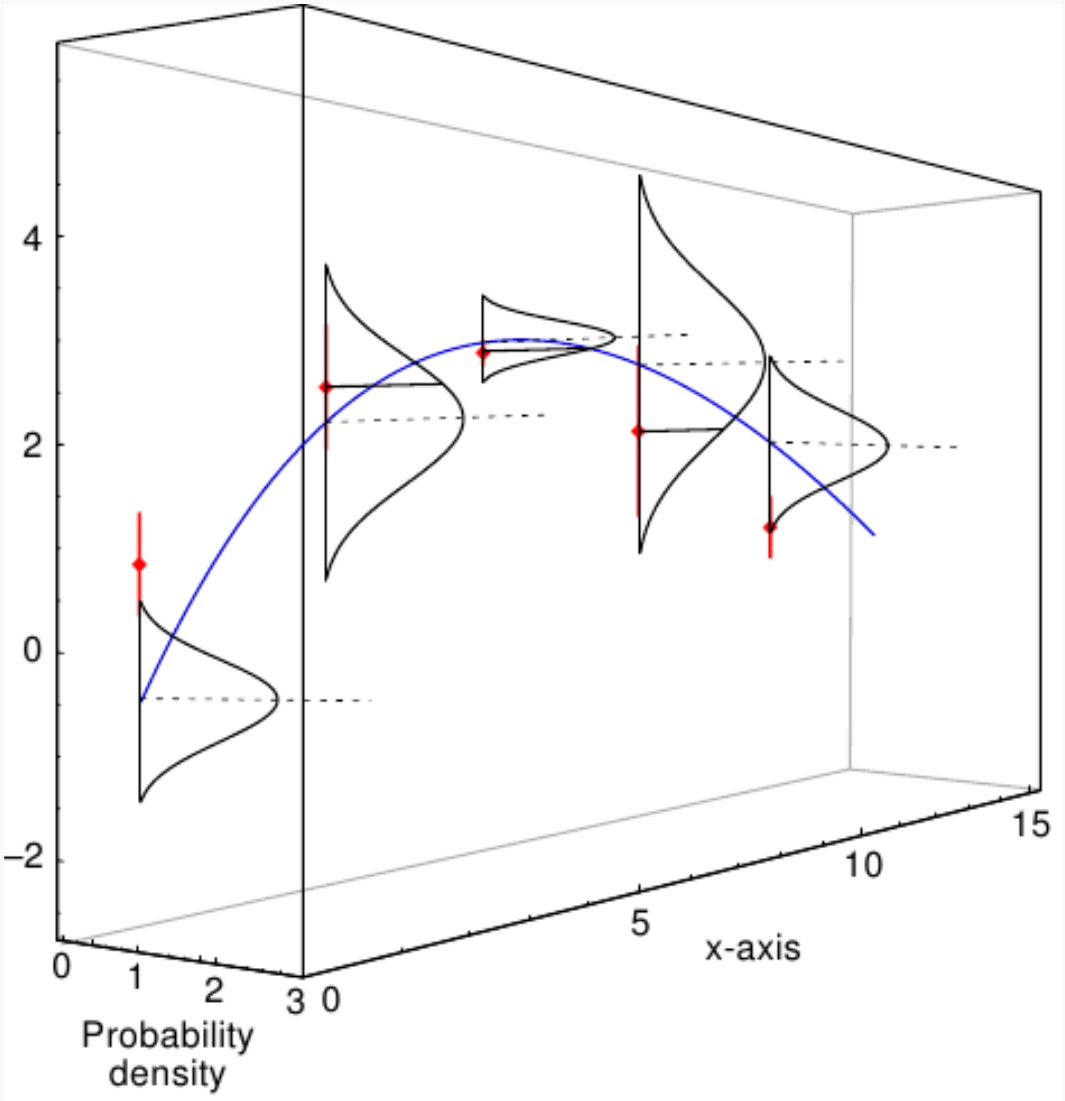
changing θ

changing $f(x)$

higher likelihood

Likelihood

the likelihood for a fixed dataset is a function of the parameters



fixed D,

changing θ

changing $f(x)$

lower likelihood

Example: exoplanets

The radial-velocity signal produced by an orbiting planet is given by a Keplerian function $rv = \mathrm{kep}(t, P, K, e, \omega, \chi)$

- assume Gaussian likelihood
  $p(D_i|\theta, \mathcal{I}) \sim \mathcal{N}(kep(\theta, t), \sigma_i)$
- assume independent data points
  $p(D|\theta, \mathcal{I}) = \prod_{i=1}^{N} p(D_i|\theta, \mathcal{I})$

```
loglike = sum([norm(loc=kep(t,P,K,e,w,X), scale=e).logpdf(rv)
                    for t,rv,e in zip(time, vrad, error)])
```

Assigning likelihood

Take-home messages:

- what you put in the likelihood is also **prior information**
- the model you use should be able to generate all the data
- is independence a good assumption?
- the Gaussian is very sensitive to outliers
  (more on this in the practical classes)

Q/A

priors and likelihood

now what?

We have a prior for the parameters, $p(\theta|\mathcal{I})$
We have a likelihood, $p(D|\theta,\mathcal{I})$

$$p(\theta|D,\mathcal{I}) \propto p(\theta|\mathcal{I})\, p(D|\theta,\mathcal{I})$$

can we just multiply them and be done with it?

- in simple cases, yes

  (one parameter, conjugate priors, analytical form of posterior is known)

- otherwise we can **sample** from the posterior

  no limit on number of parameters, no restrictions on priors,

  but slower because it is numerical instead of analytical

how to sample from a distribution

- inverse cdf method
  the inverse cumulative distribution function

  transforms from U(0,1) to p(y)
- rejection sampling
  samples from a simple distribution $kq(z)$

  reject if they fall in the grey area
- MCMC

  build a Markov chain with $p(\theta)$ as target distribution

how to sample from a distribution

- inverse cdf method

  the inverse cumulative distribution function

  transforms from U(0,1) to p(y)

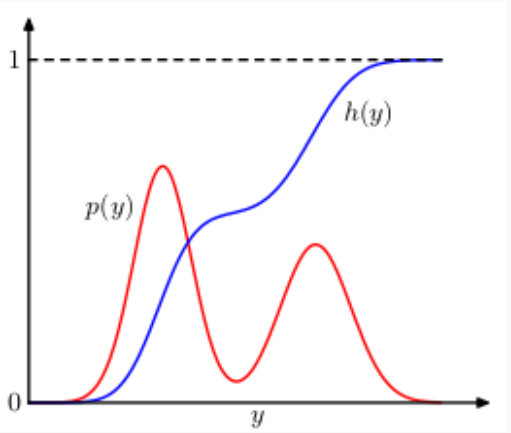
- rejection sampling

  samples from a simple distribution $kq(z)$

  reject if they fall in the grey area

- MCMC

  build a Markov chain with $p(\theta)$ as target distribution



how to sample from a distribution

- inverse cdf method

  the inverse cumulative distribution function

  transforms from U(0,1) to p(y)

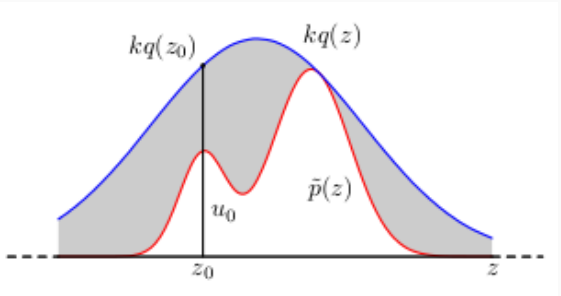
- rejection sampling

  samples from a simple distribution $kq(z)$

  reject if they fall in the grey area

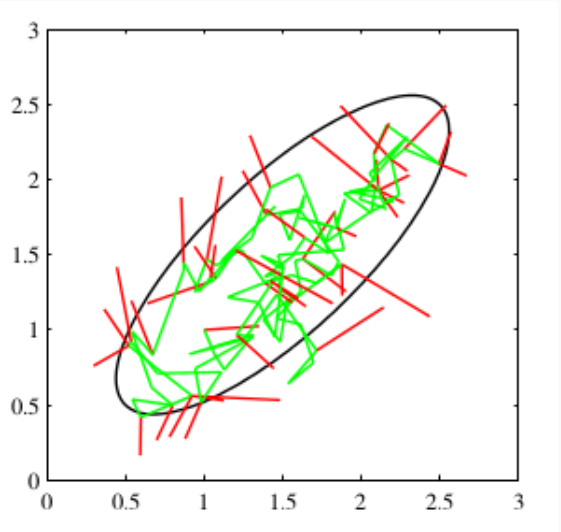
- MCMC

  build a Markov chain with $p(\theta)$ as target distribution



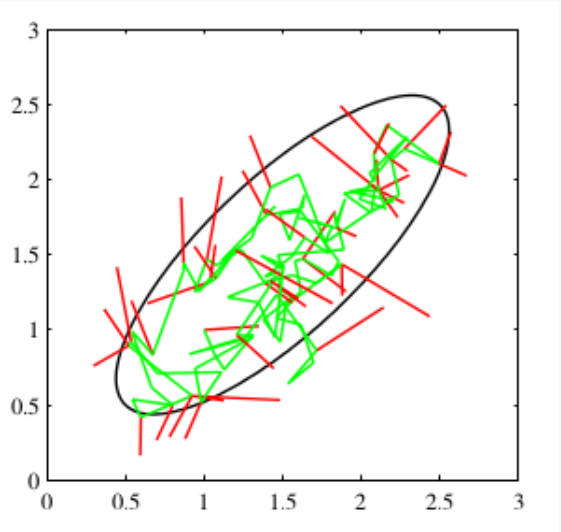
how to sample from a distribution

- inverse cdf method
  the inverse cumulative distribution function
  transforms from U(0,1) to p(y)
- rejection sampling
  samples from a simple distribution $kq(z)$
  reject if they fall in the grey area
- MCMC
  build a Markov chain with $p(\theta)$ as target distribution



how to sample from a distribution

- inverse cdf method
  the inverse cumulative distribution function
  transforms from U(0,1) to p(y)
- rejection sampling
  samples from a simple distribution $kq(z)$
  reject if they fall in the grey area
- MCMC
  build a Markov chain with $p(\theta)$ as target distribution
  still works for many parameters
  only need to be able to evaluate $p(\theta)$



MCMC: how to

Metropolis-Hastings is the most common MCMC algorithm (but there are many, many more!)

It goes like this

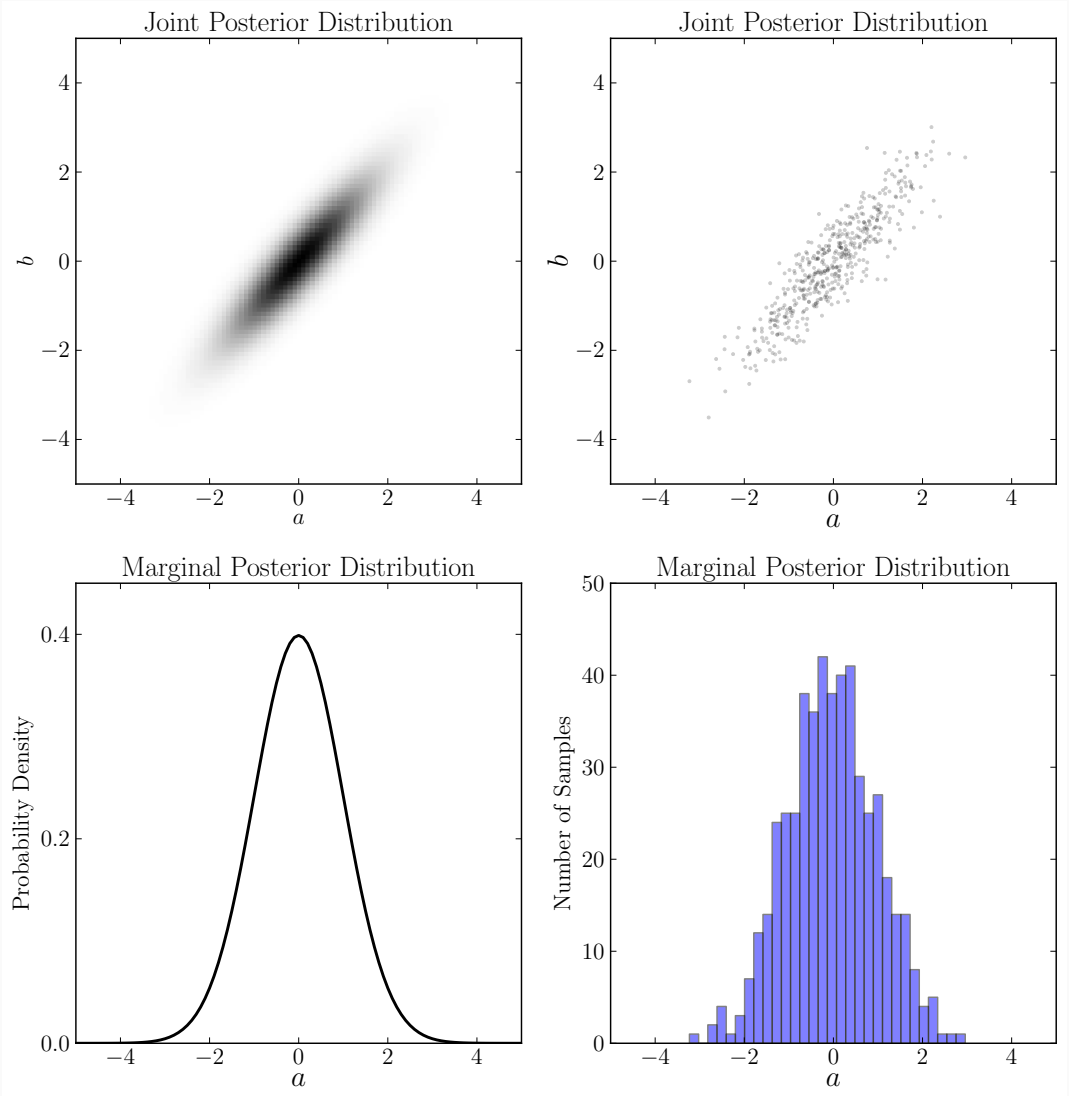
1. Choose an arbitrary point x_0 to be the first sample
2. **For each iteration t:**
   - 2.1 Generate a candidate x' for the next sample by picking from the distribution $Q(x'|x_t)$.
   - 2.2 Calculate the acceptance ratio

$$r = p(x')/p(x_t)$$

   - 2.3 If $r \geq 1$, then the candidate is more likely than x_t:
     - automatically accept the candidate by setting $x_{t+1} = x'$.

     Otherwise, accept the candidate with probability r
   - 2.4 If the candidate is rejected, set $x_{t+1} = x_t$

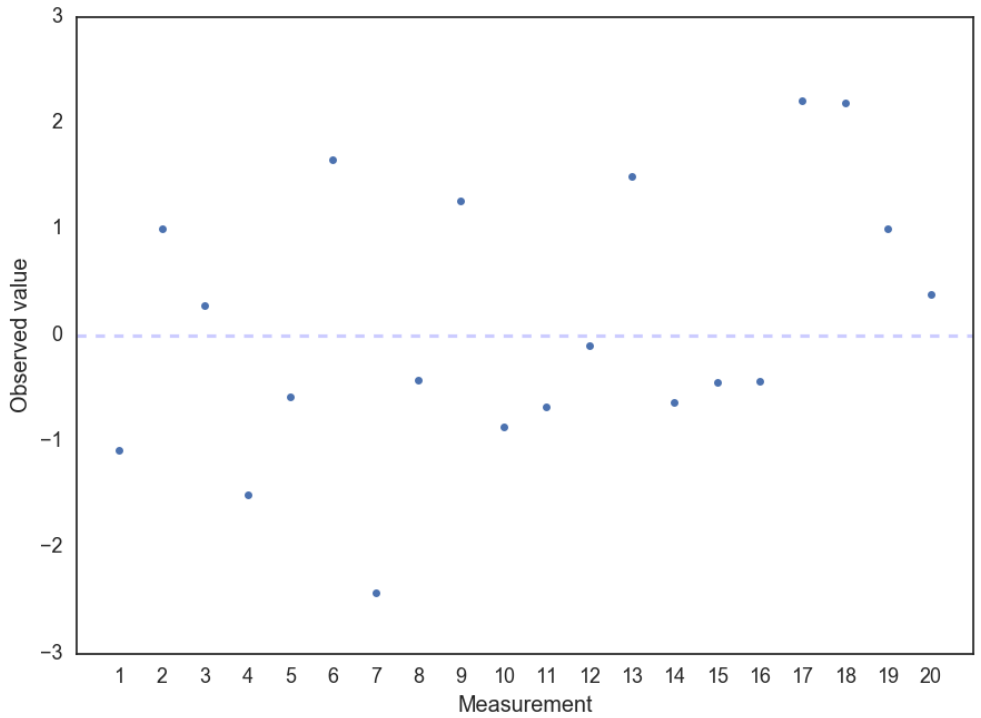
MCMC is awesome!



Credit: Brendon Brewer

MCMC step-by-step

Let's use a very simple example to see what MCMC does.



We have 20 measurements of μ.

Use a Gaussian likelihood

$$p(D|\mu, I)$$

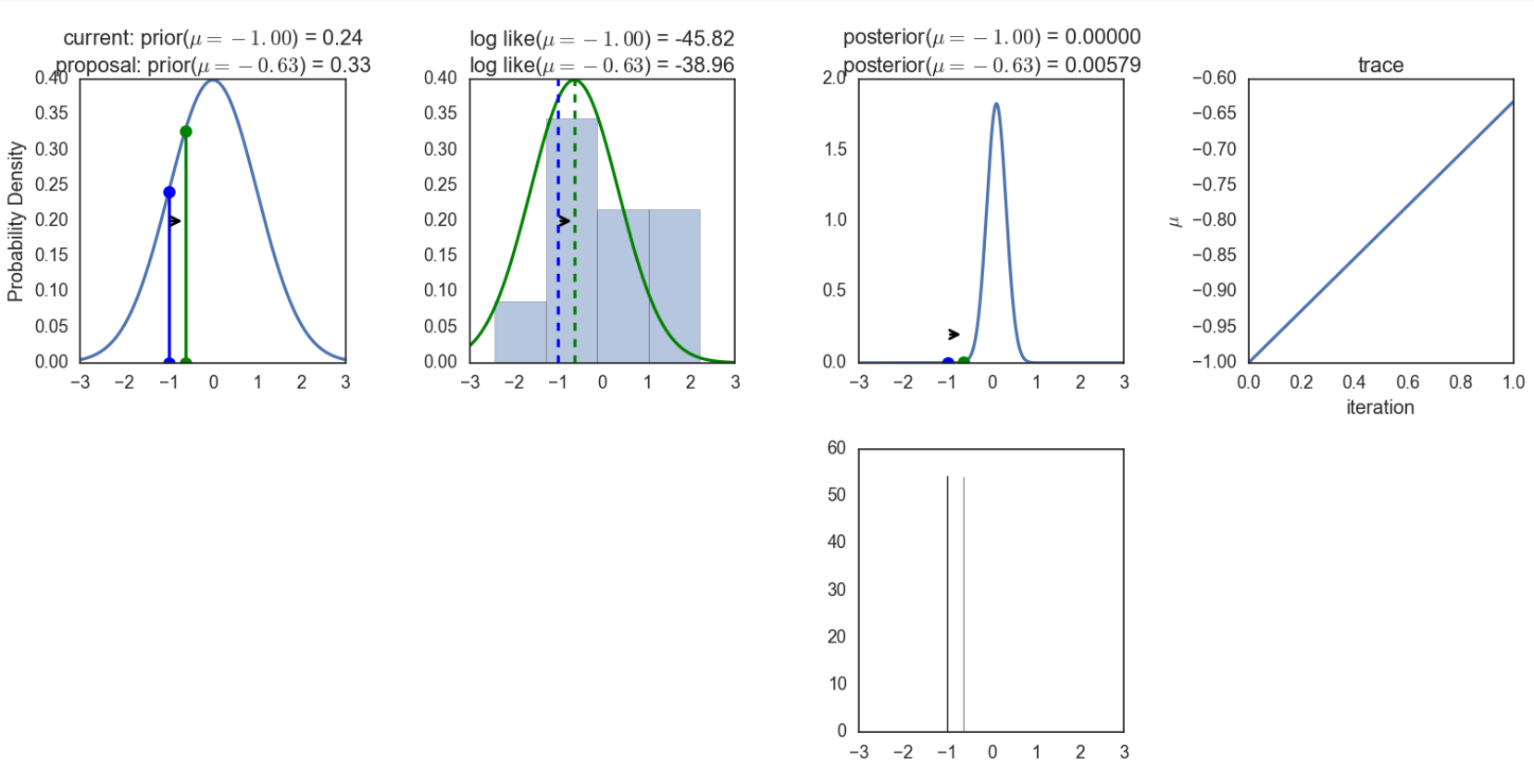
and a Gaussian prior

$$p(\mu|I)$$

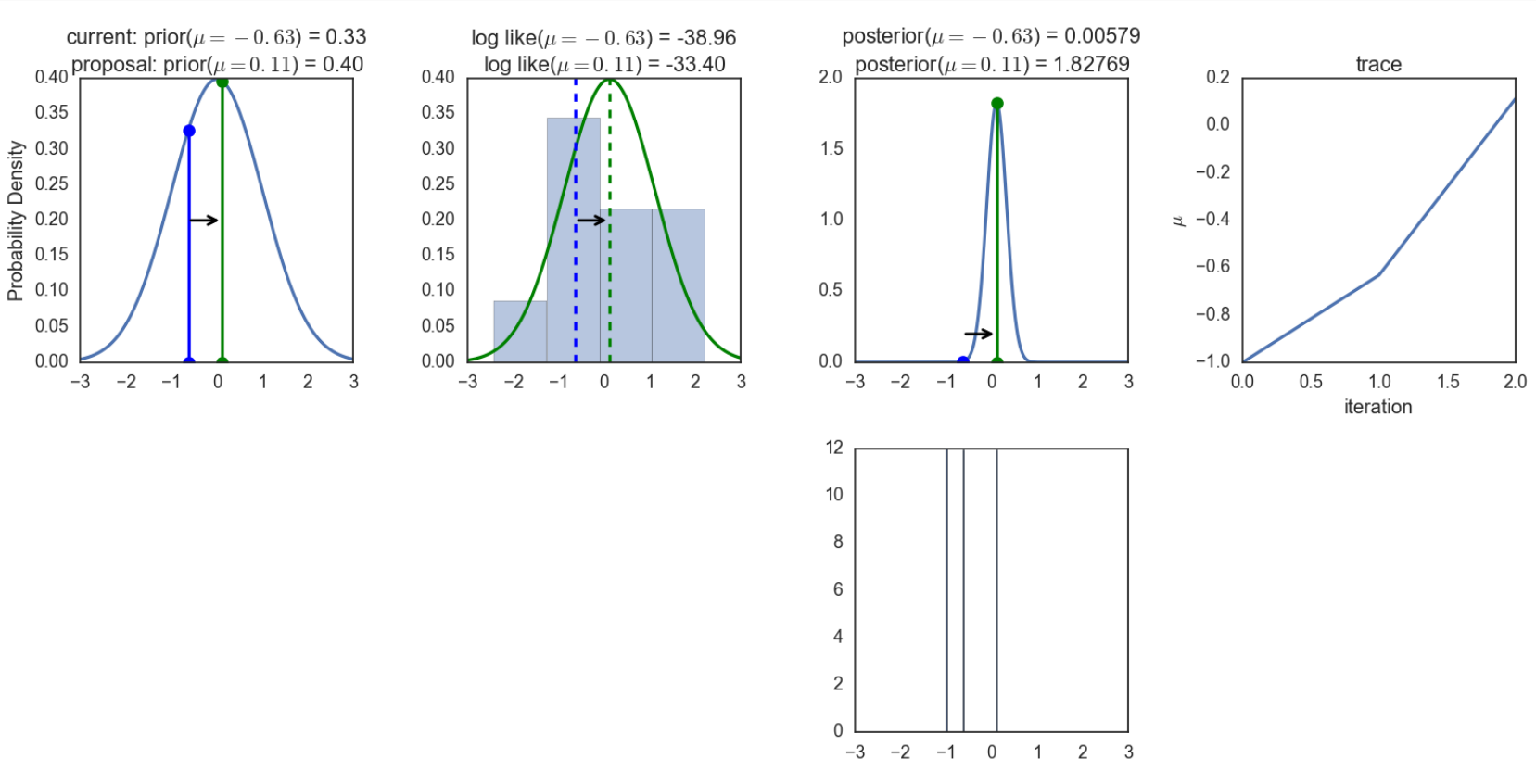
Run MCMC
to get samples from posterior

$$p(\mu|D, I)$$

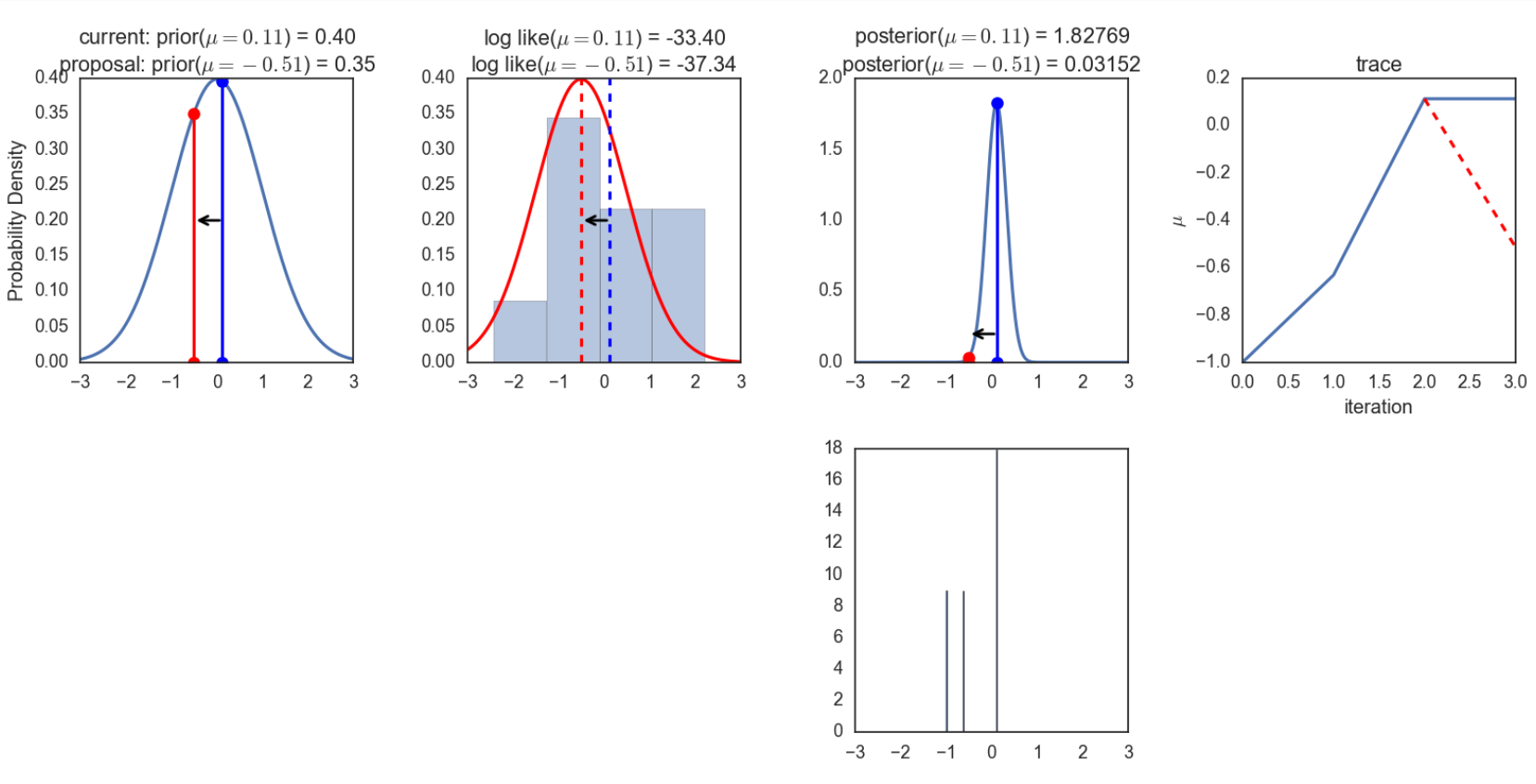
MCMC step-by-step



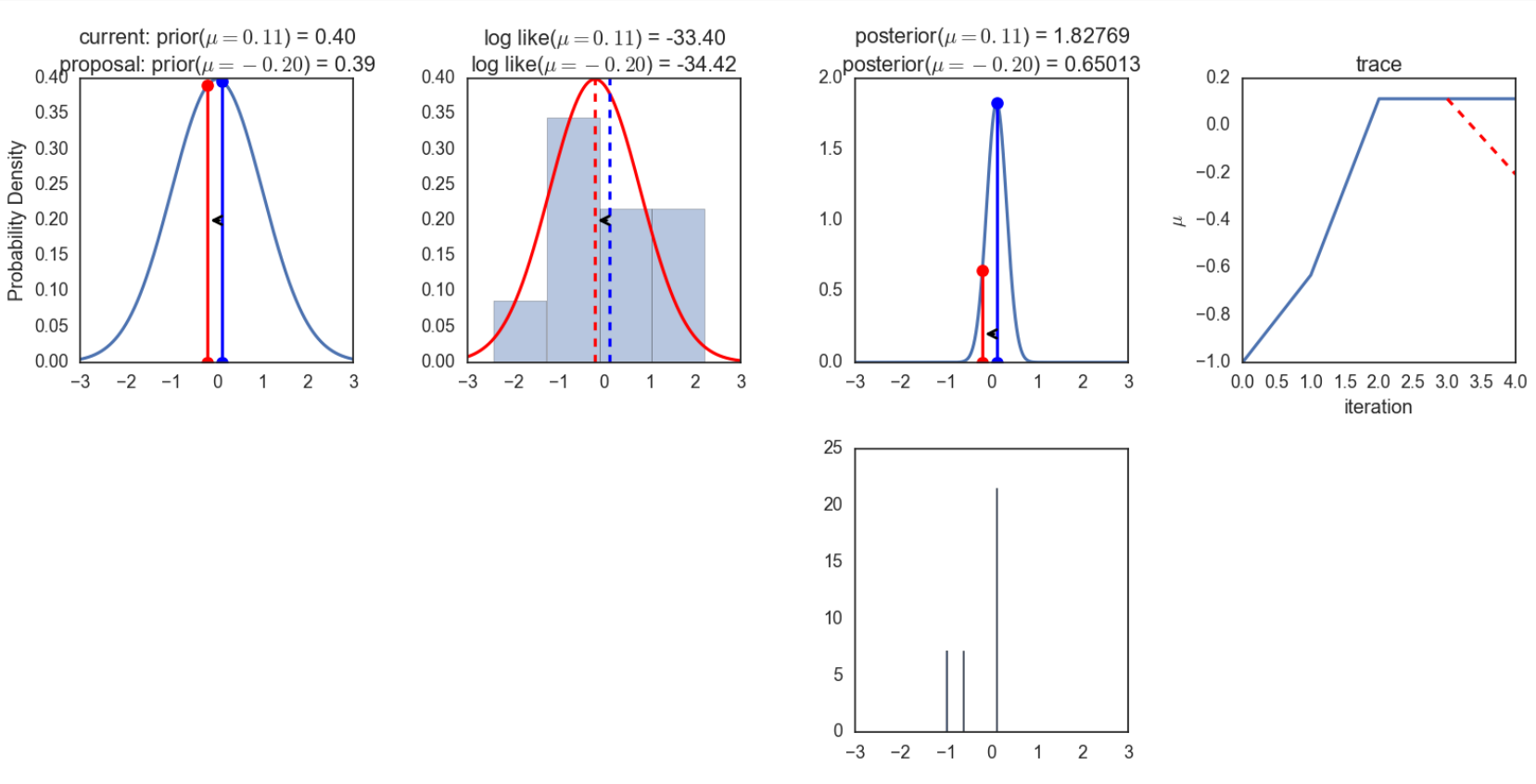
MCMC step-by-step



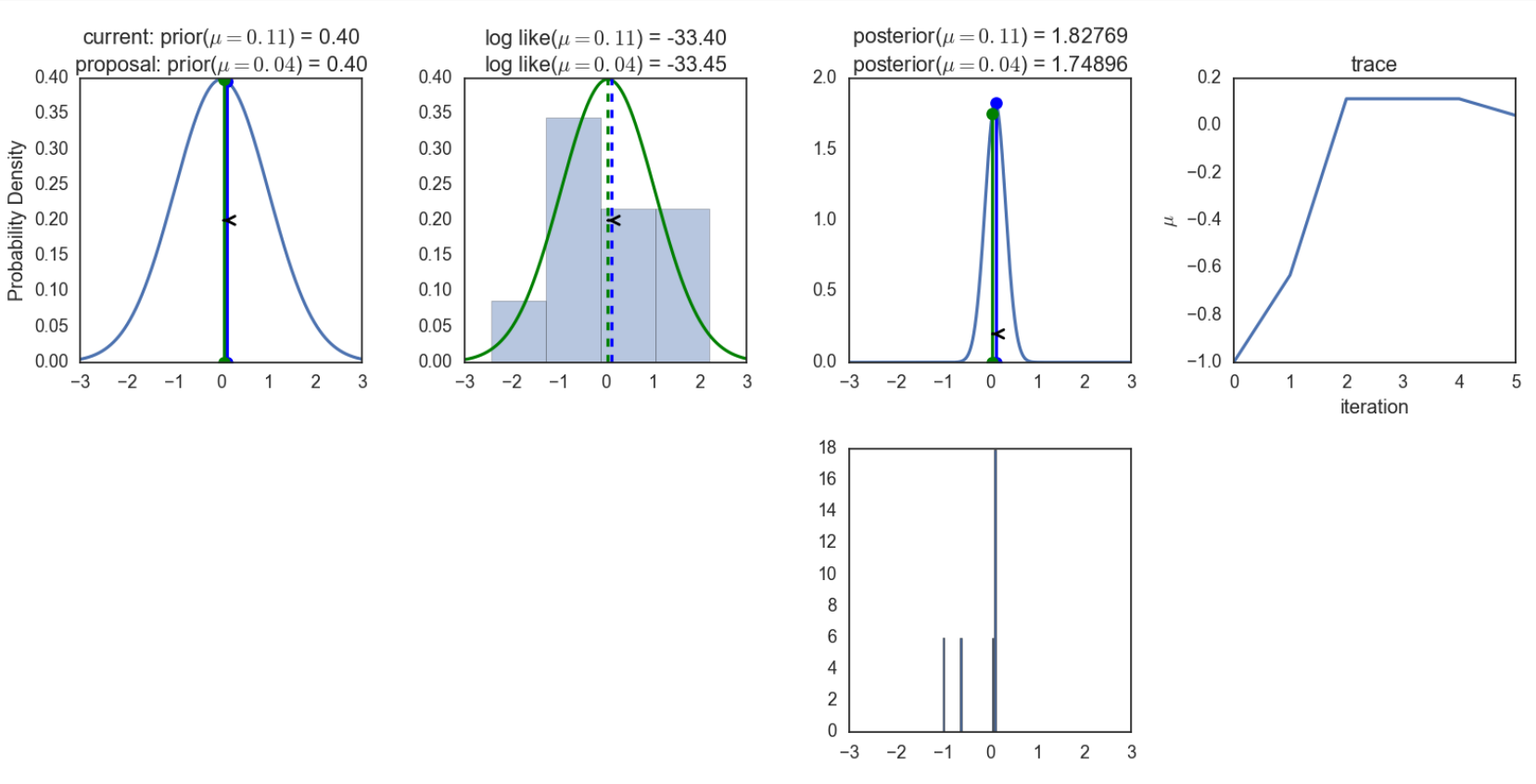
MCMC step-by-step



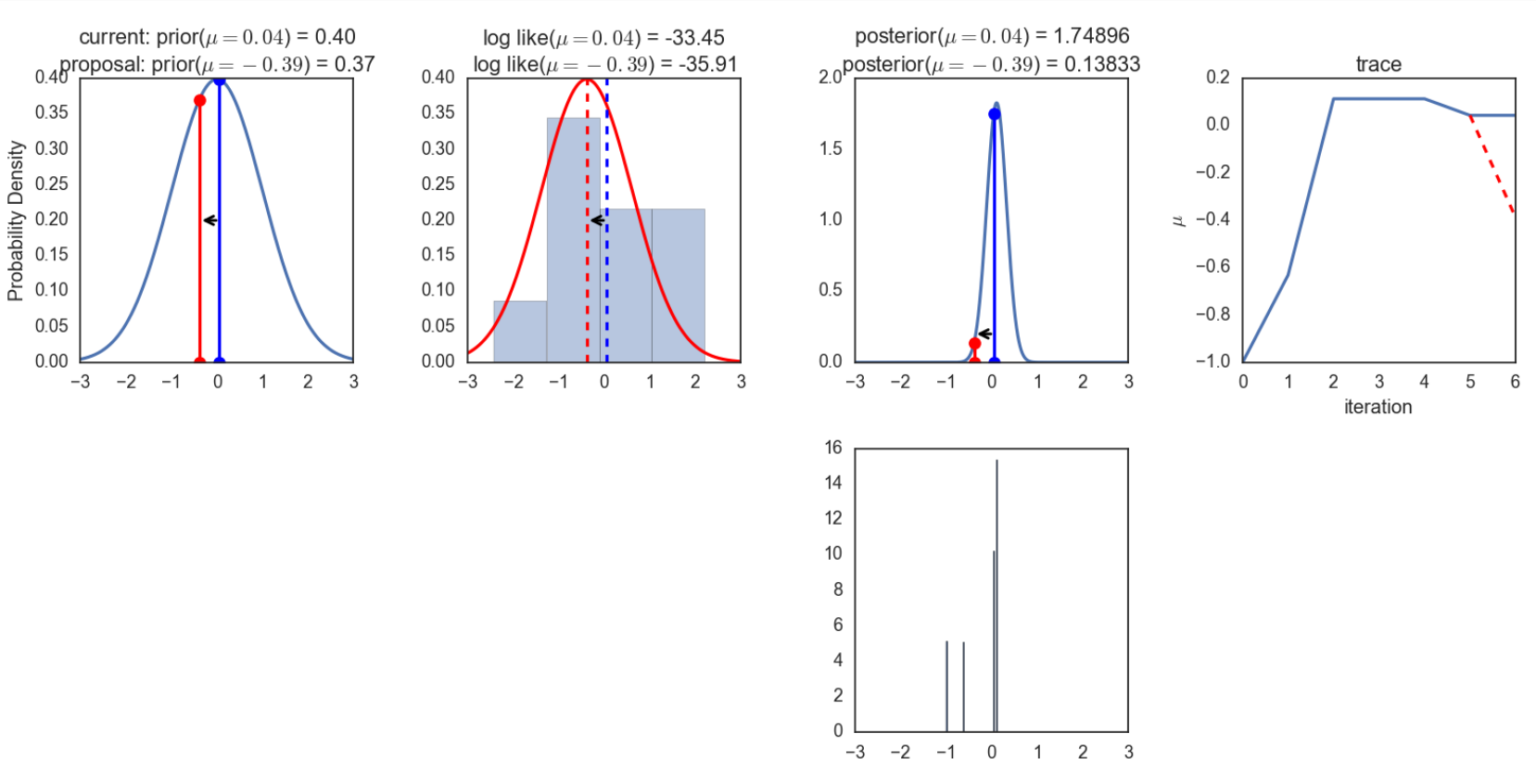
MCMC step-by-step



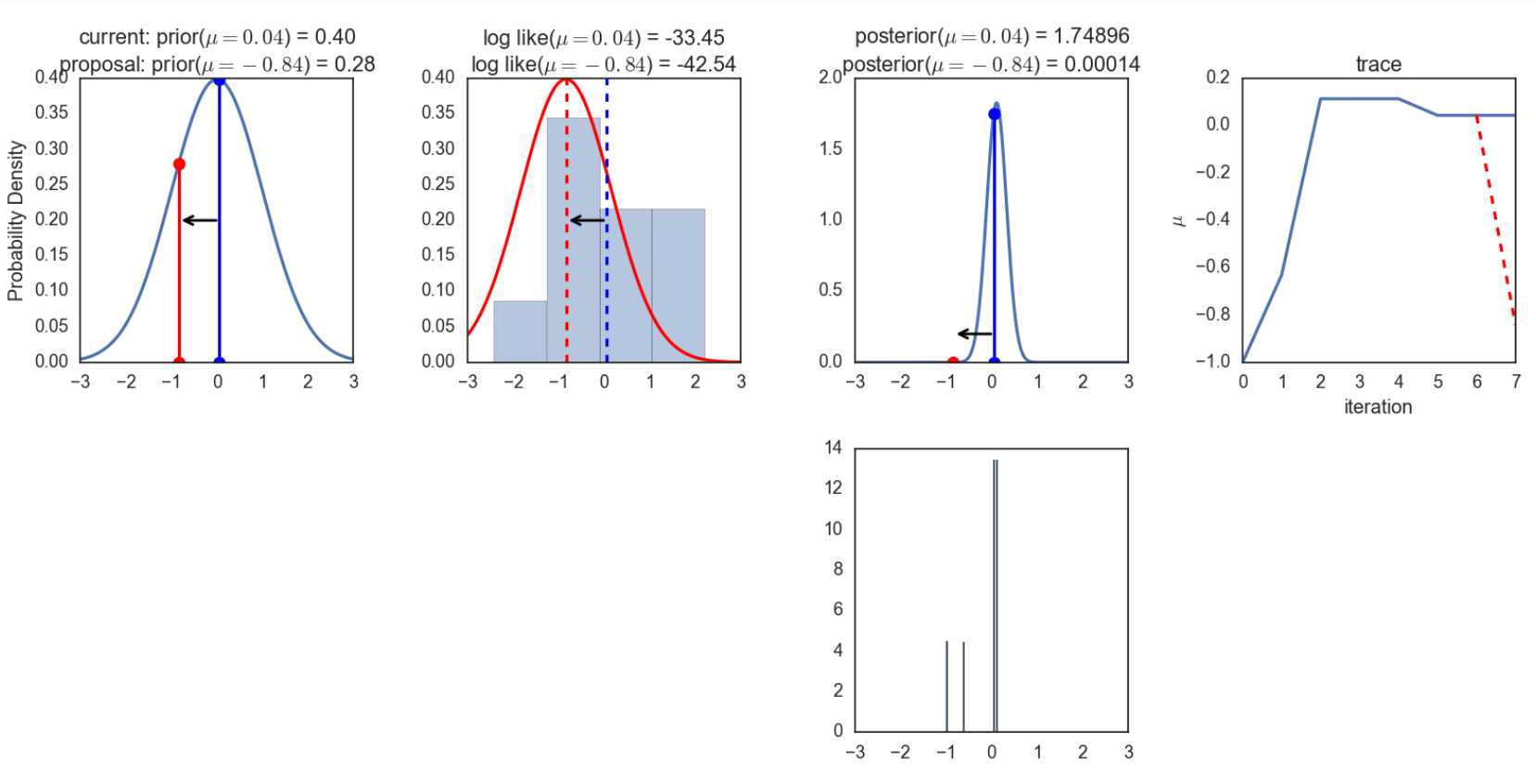
MCMC step-by-step



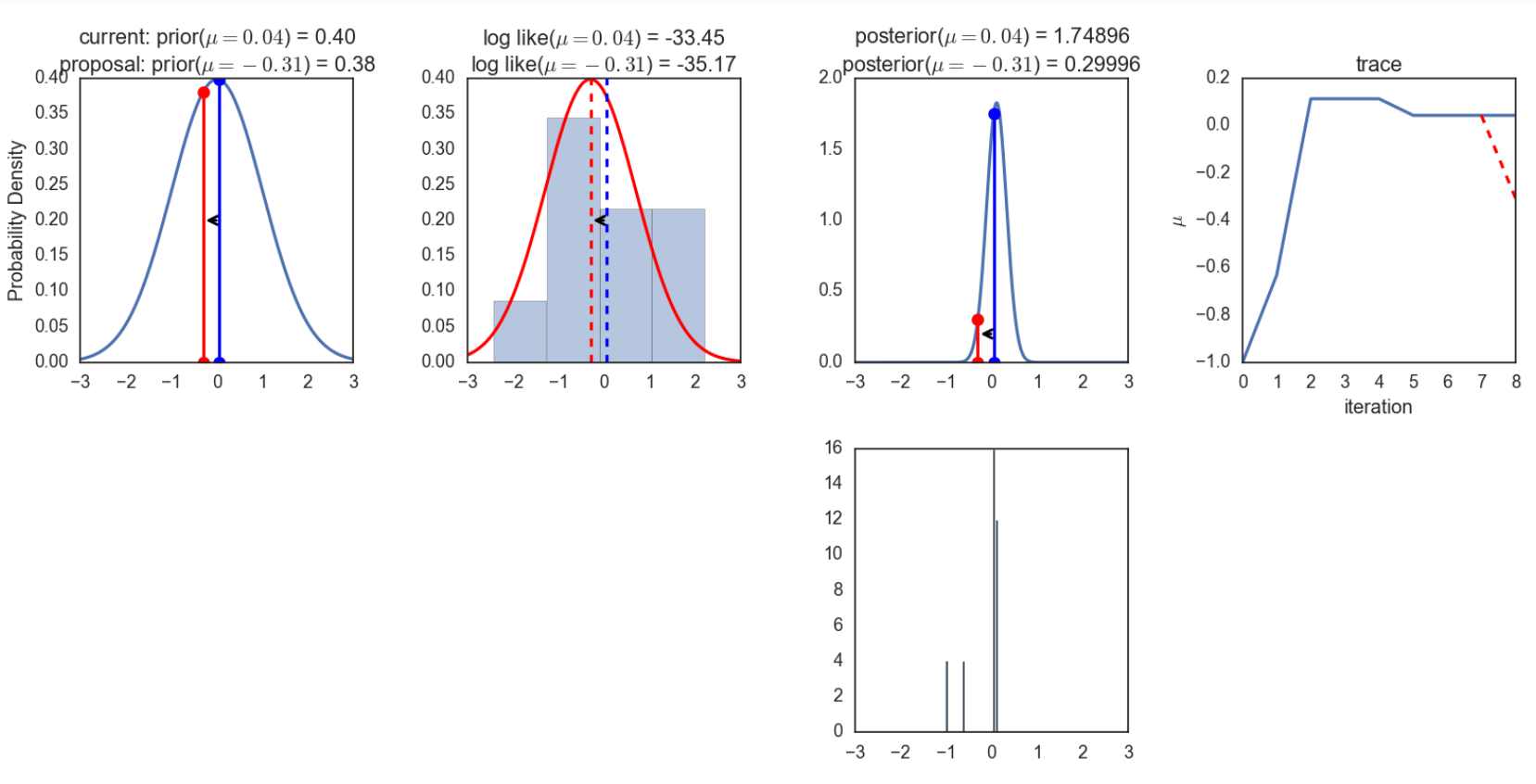
MCMC step-by-step



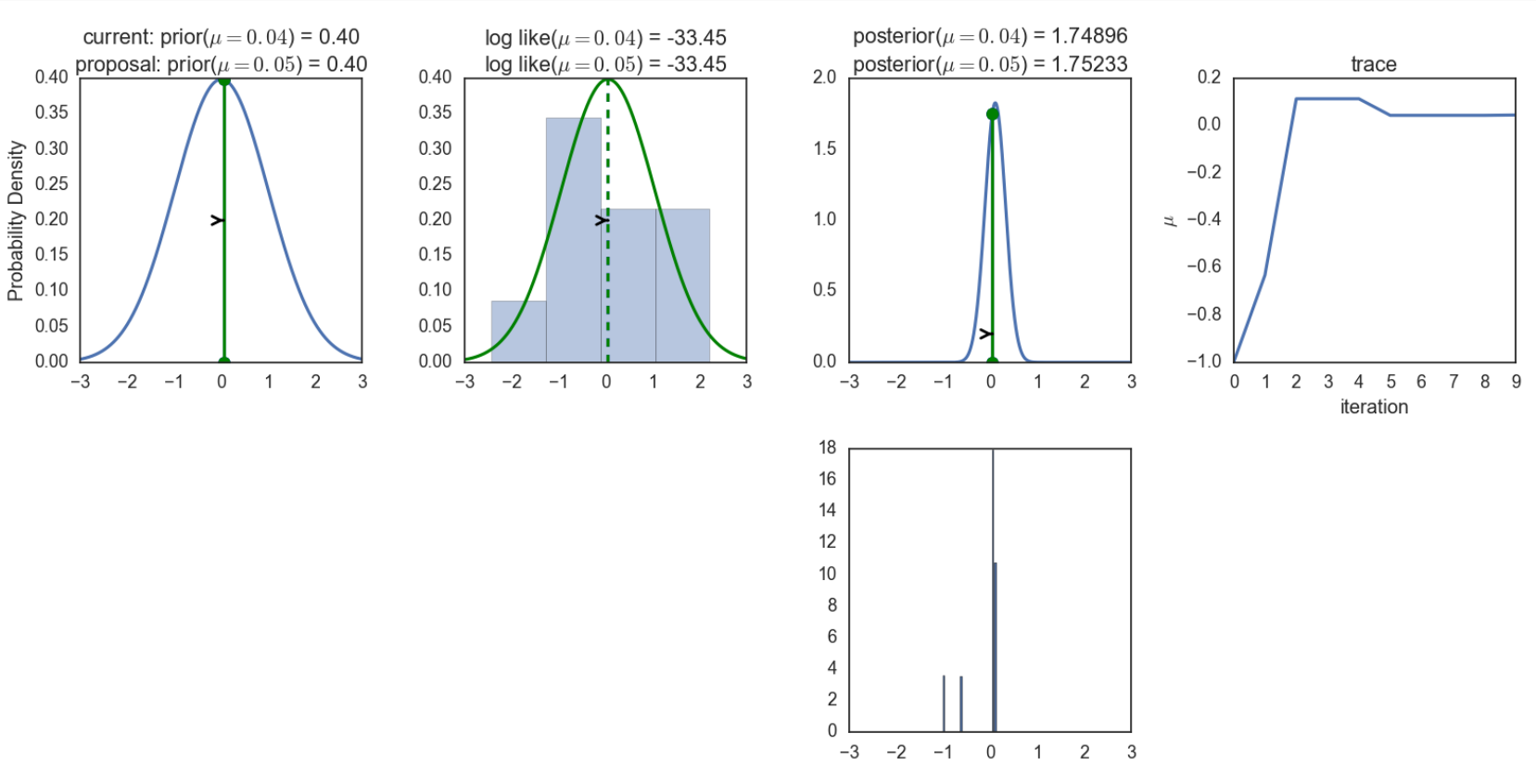
MCMC step-by-step



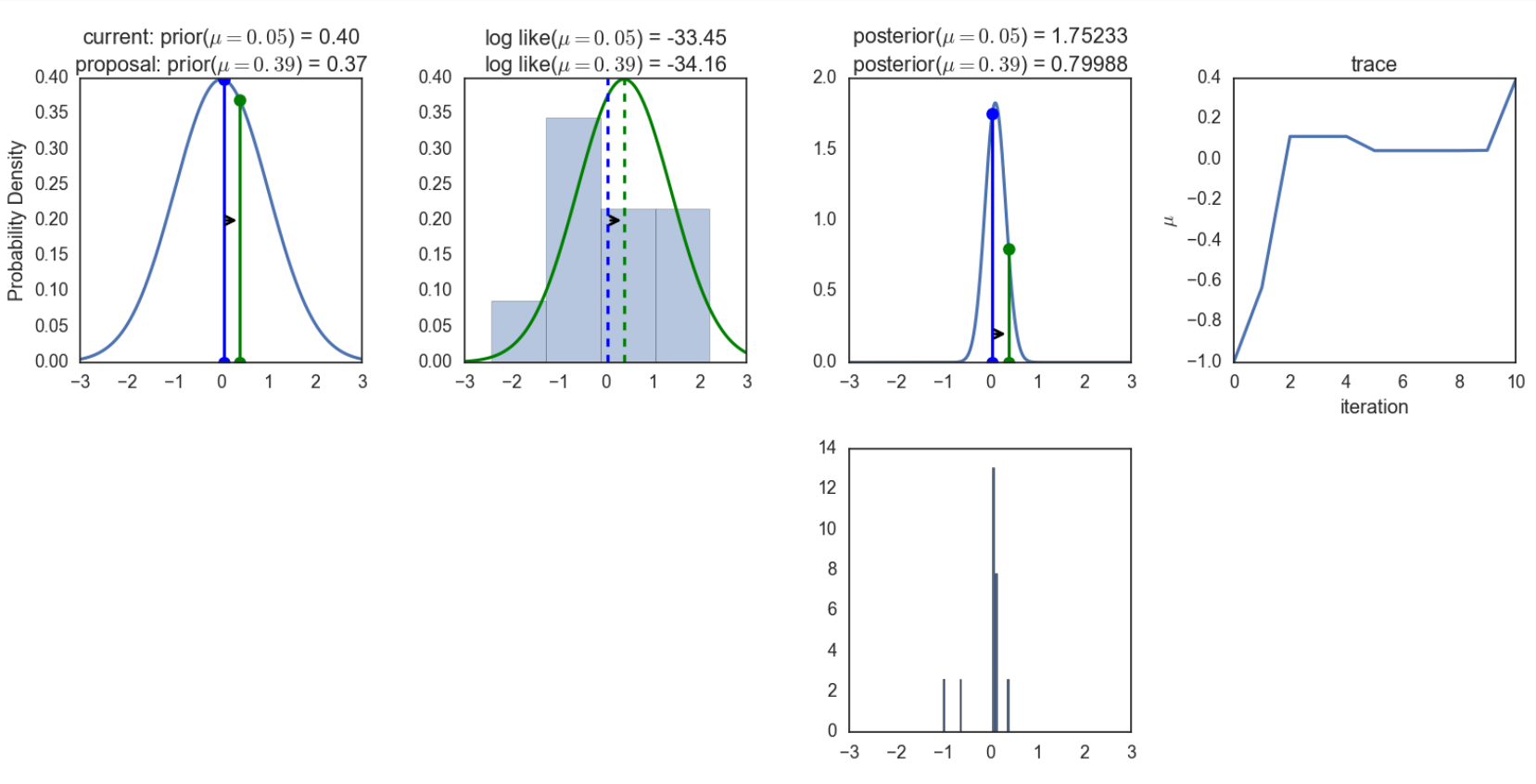
MCMC step-by-step



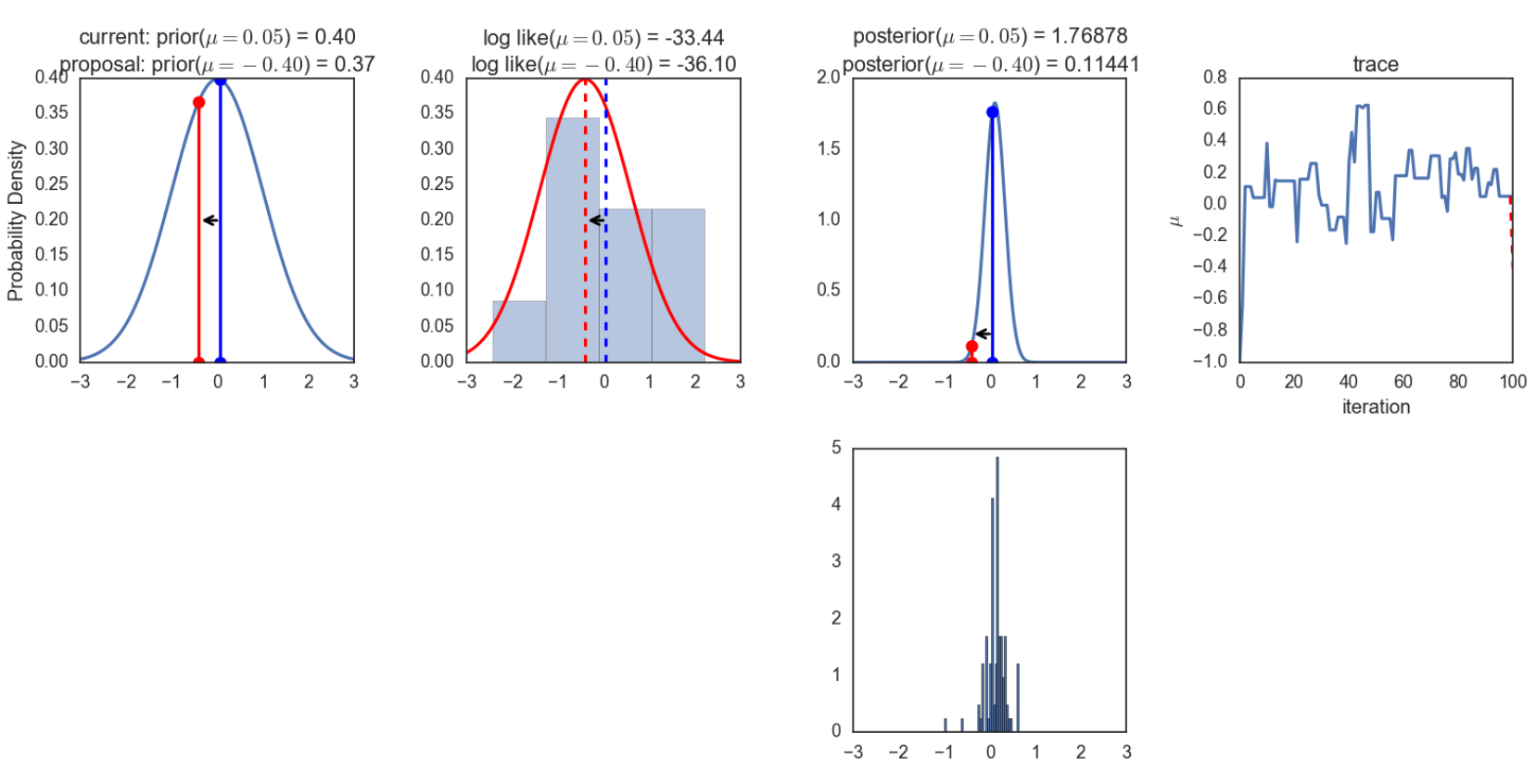
MCMC step-by-step



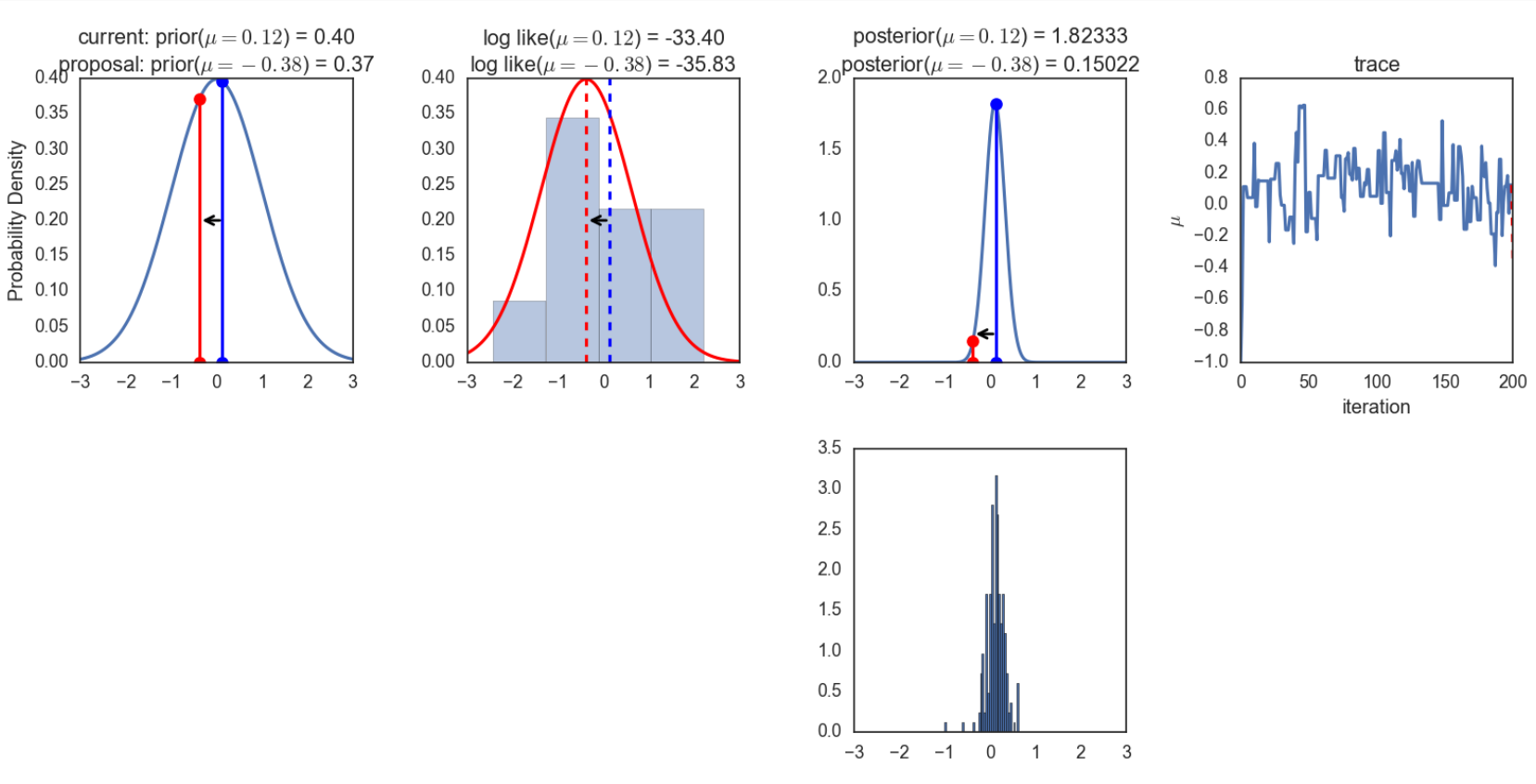
MCMC step-by-step



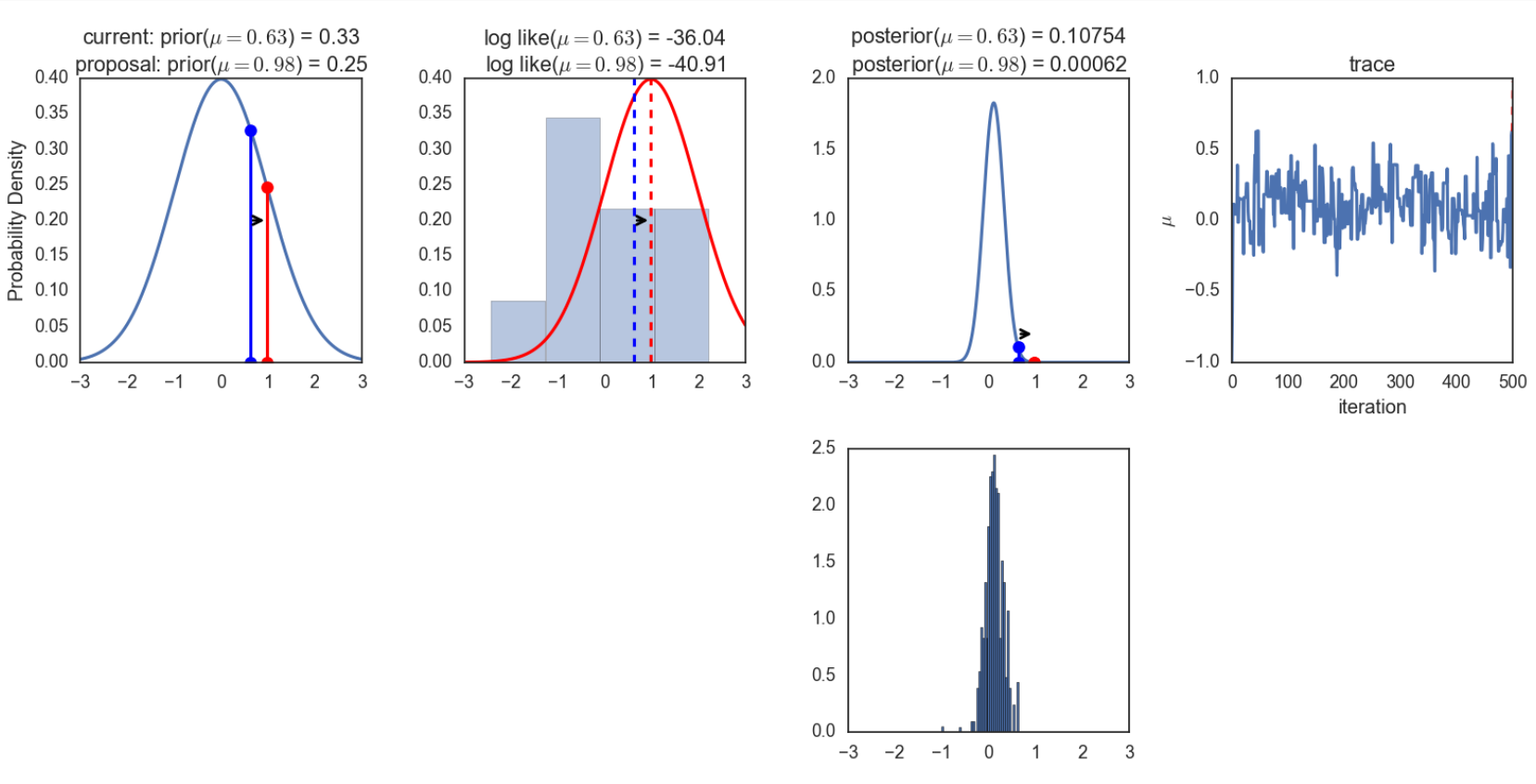
MCMC step-by-step



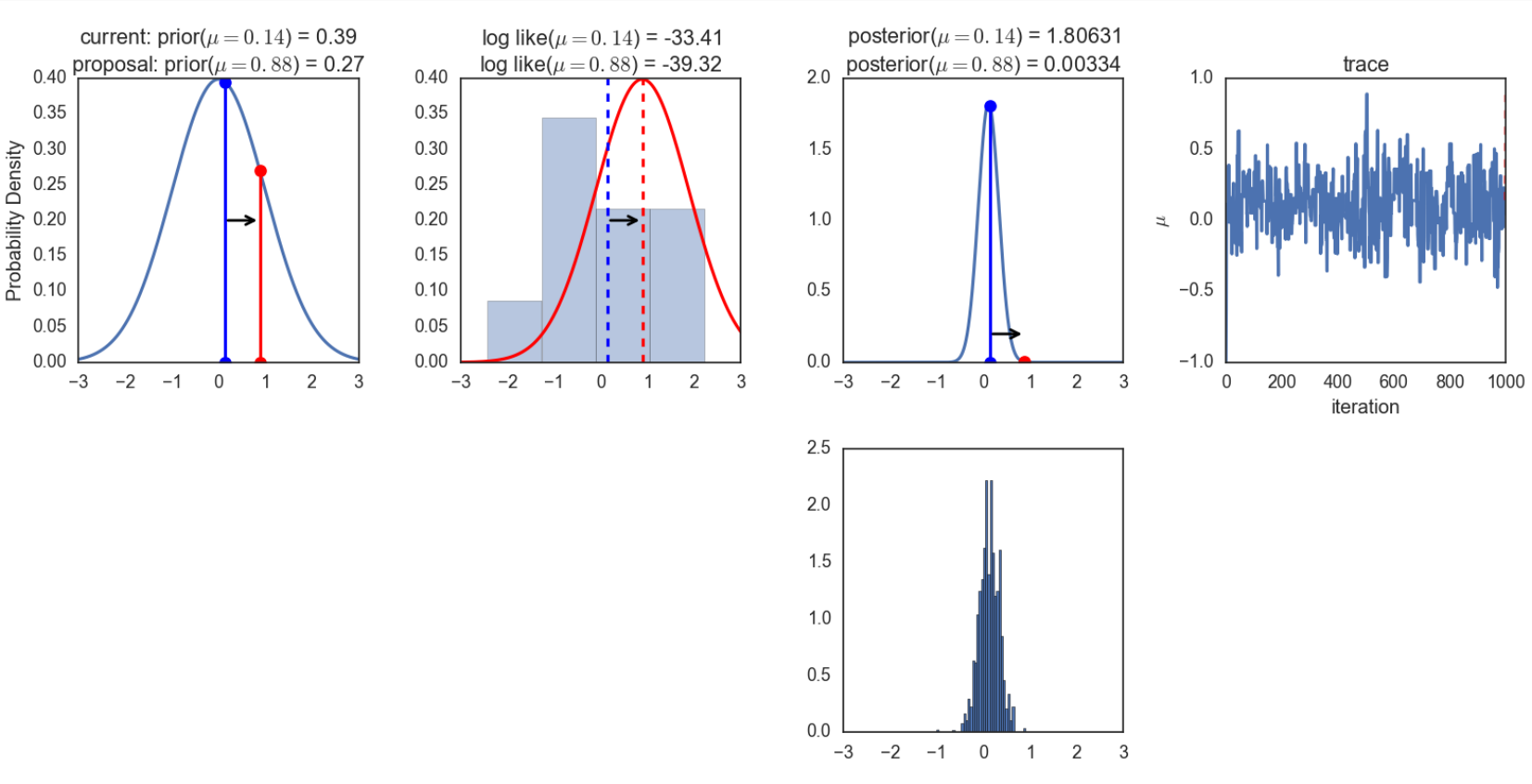
MCMC step-by-step



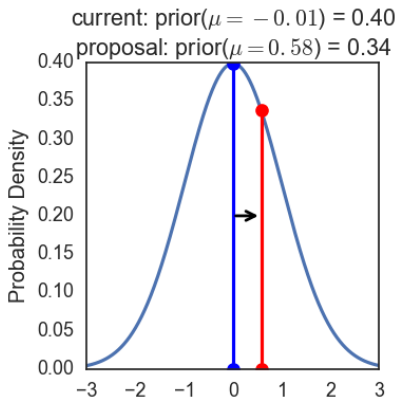
MCMC step-by-step

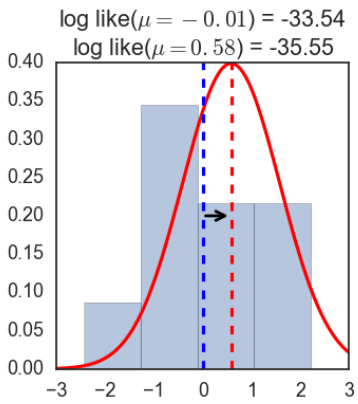


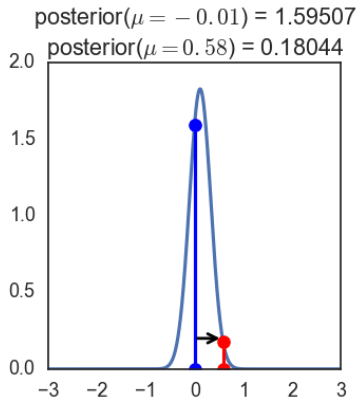
MCMC step-by-step

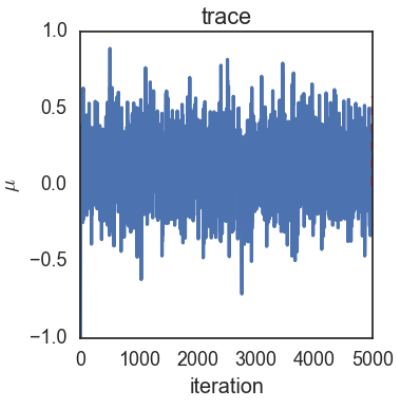


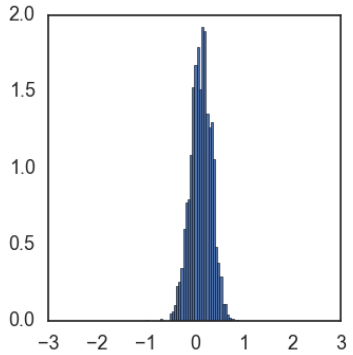
MCMC step-by-step











Q/A

MCMC

Finally a real example

The planets and activity of CoRoT-7

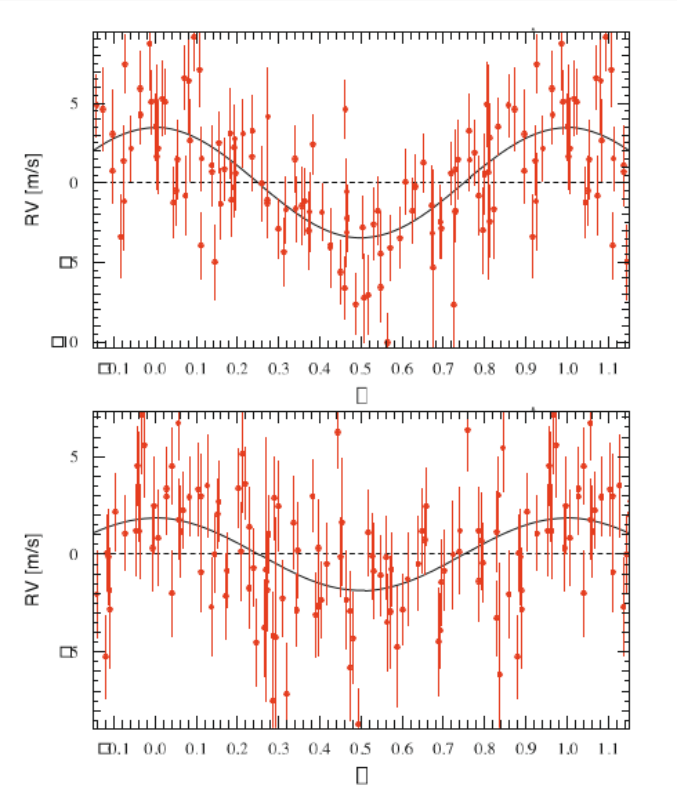
- CoRoT-7b was the first super-Earth with measured radius (Léger et al. 2009)

credit: Raphaëlle Haywood

Finally a real example

The planets and activity of CoRoT-7

- A follow-up RV campaign revealed CoRoT-7c (maybe d)
  (Queloz et al. 2009)



Finally a real example

The planets and activity of CoRoT-7

- Because the star is active (2% variation in photometry) many authors considered different activity corrections and got different planet masses, number of planets, etc.

  Hatzes et al. (2010, 2011), Lanza et al. (2010), Boisse et al. (2011)

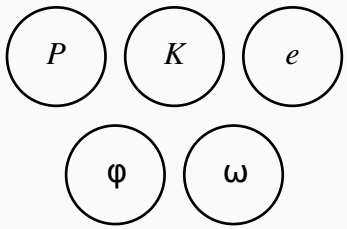
  Ferraz-Mello et al. (2011), Pont et al. (2011)

- Simultaneous CoRoT photometry and HARPS RV were obtained in 2012

  Haywood et al. (2014)

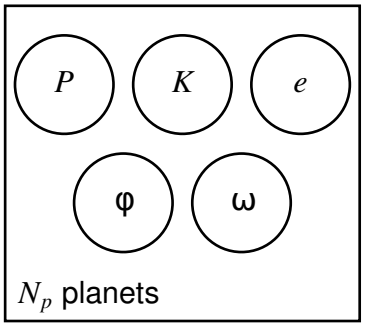
Finally a real example

The planets and activity of CoRoT-7



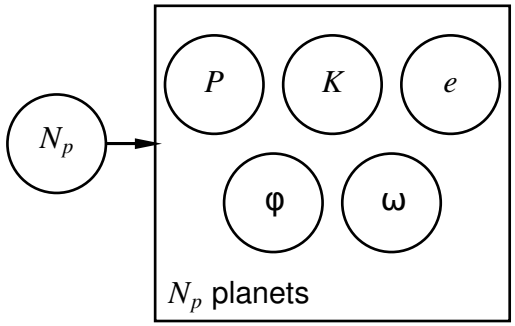
Finally a real example

The planets and activity of CoRoT-7



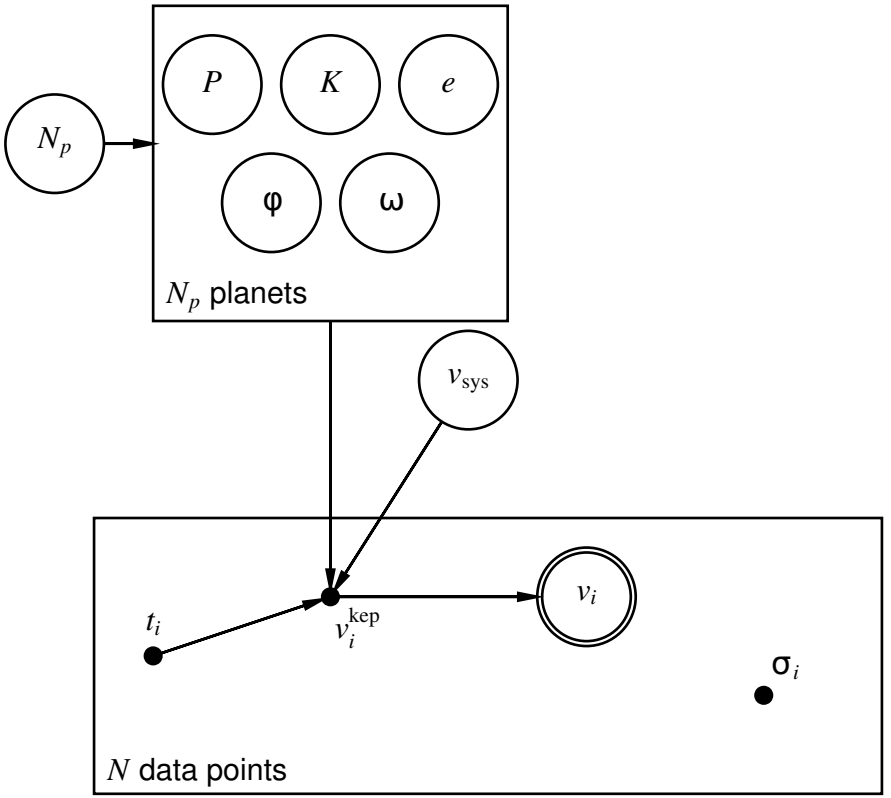
Finally a real example

The planets and activity of CoRoT-7



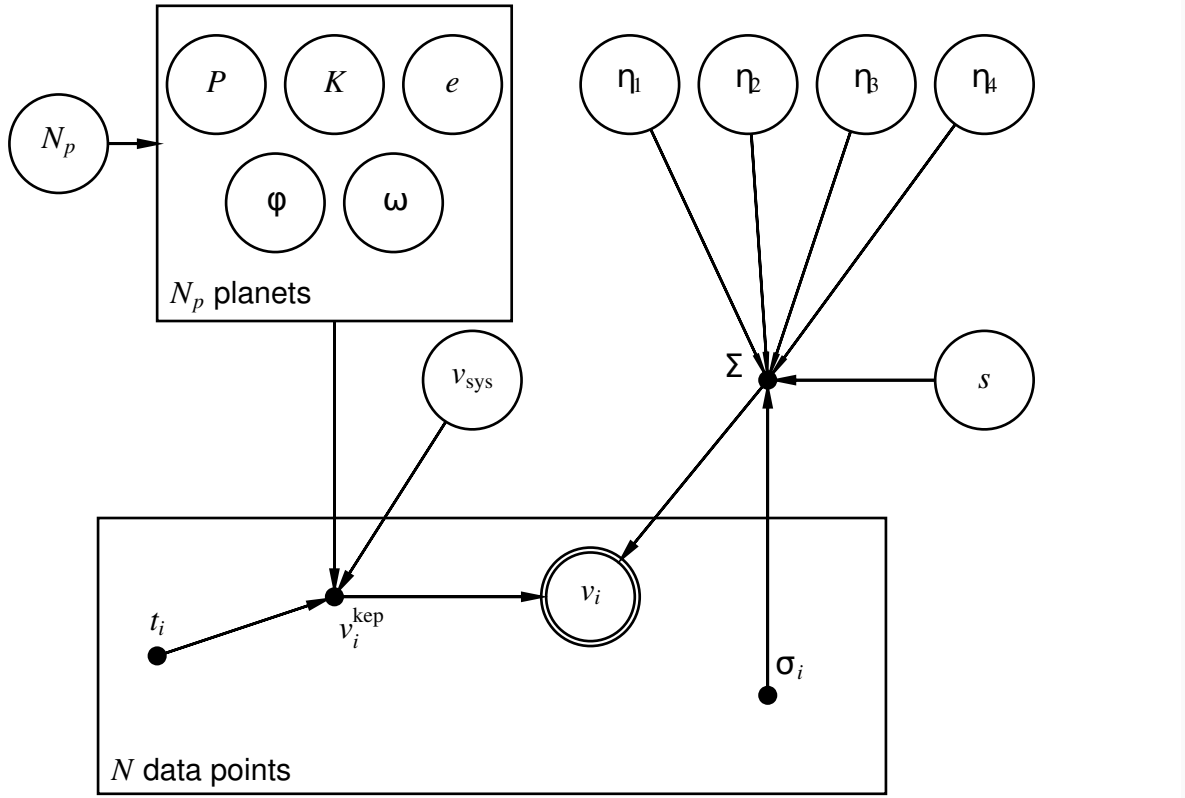
Finally a real example

The planets and activity of CoRoT-7



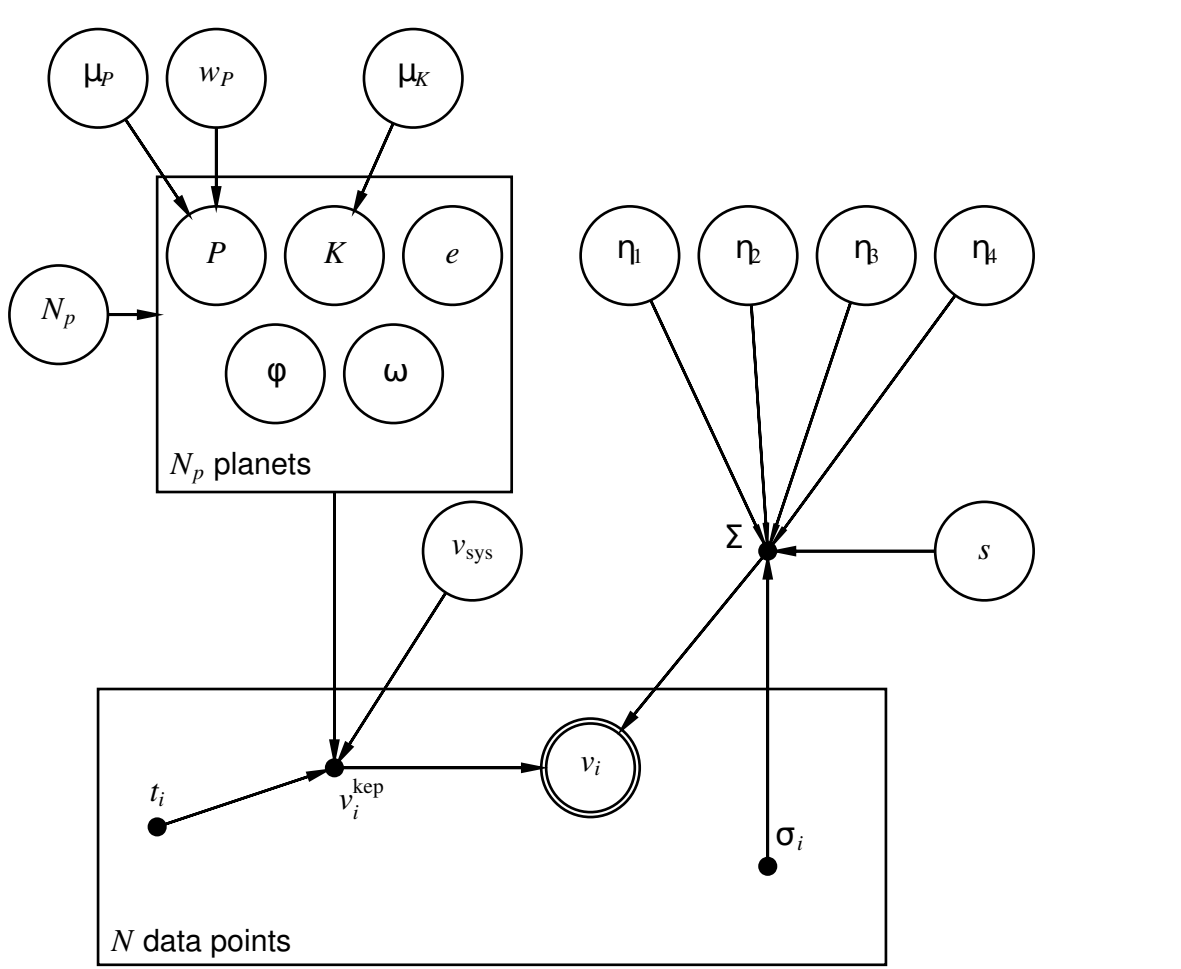
Finally a real example

The planets and activity of CoRoT-7



Finally a real example

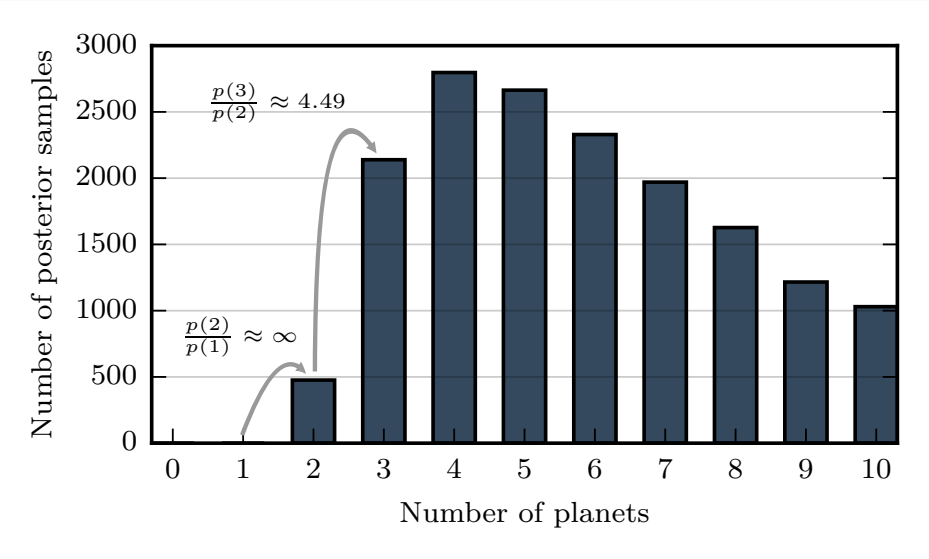
The planets and activity of CoRoT-7



Finally a real example

The planets and activity of CoRoT-7

- Our model uses all RV observations (2008 and 2012)
  And **only** RV observations, no photometry
- It fits the number of planets *together* with their parameters
- The Gaussian process (our model for the activity) provides an estimate of the rotation period.



Faria et al. (2016)

Q/A

CoRoT-7 example

Conclusion

Take-home messages

- The rules of probability tell you what you can and can't do
- Bayes' theorem "inverts" the probabilities

$$p(D|\theta, \mathcal{I}) \rightarrow p(\theta|D, \mathcal{I})$$

  but requires a prior $p(\theta|\mathcal{I})$
- Priors are not that hard to set, most of the times
- The likelihood connects θ and D; it is also prior information
- MCMC is almost generally applicable and easy to use and understand (in `Python` see `emcee`, `PyMC`, `PyStan`)

Things I didn't mention but are important

- I mostly forgot about the evidence $p(D|\mathcal{I})$; you shouldn't. This term is important when comparing two or more models but it's still a bit hard to calculate in some cases

Things I didn't mention but are important

- I mostly forgot about the evidence $p(D|\mathcal{I})$; you shouldn't.
  This term is important when comparing two or more models
  but it's still a bit hard to calculate in some cases

- "*You can't fit models with more parameters than data points*"
  That is a myth, sure you can!
  We will do it in the practical classes

Things I didn't mention but are important

- I mostly forgot about the evidence $p(D|\mathcal{I})$; you shouldn't.
  This term is important when comparing two or more models
  but it's still a bit hard to calculate in some cases

- "*You can't fit models with more parameters than data points*"
  That is a myth, sure you can!
  We will do it in the practical classes

- It's ok to have the prior depend on free parameters.

Things I didn't mention but are important

- I mostly forgot about the evidence $p(D|\mathcal{I})$; you shouldn't.
  This term is important when comparing two or more models
  but it's still a bit hard to calculate in some cases

- "*You can't fit models with more parameters than data points*"
  That is a myth, sure you can!
  We will do it in the practical classes

- It's ok to have the prior depend on free parameters.

- Bayesian methods, as any other methods, are not immune to absurdities: you still need to think...

Online Material

The slides, lecture notes and other material are in

`github.com/j-faria/Bayes-IA`

Q/A

everything

References I


A. Gelman.
Bayesian data analysis.
Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 3rd edition, 2014.


P. C. Gregory.
Bayesian logical data analysis for the physical sciences.
Cambridge University Press, 2010.


D. W. Hogg.
Data analysis recipes: Probability calculus for inference.
ArXiv e-prints, 1205:4446, May 2012.

References II


D. W. Hogg, J. Bovy, and D. Lang.
Data analysis recipes: Fitting a model to data.
ArXiv e-prints, 1008:4686, Aug. 2010.


H. Jeffreys.
Theory of probability.
Oxford University Press, 3rd ed edition, 1998.


R. E. Kass and A. E. Raftery.
Bayes Factors.
J. Am. Stat. Assoc., 90(430):773, June 1995.


J. K. Kruschke.
Bayesian estimation supersedes the t test.
J. Exp. Psychol. Gen., 142(2):573–603, 2013.

References III


T. J. Loredo.
The Return of the Prodigal: Bayesian Inference in Astrophysics.
Valencia, 1994.


R. McElreath.
Statistical rethinking: a Bayesian course with examples in R and Stan.
Number 122 in Chapman & Hall/CRC texts in statistical science series. Taylor & Francis, Boca Raton, 2016.


D. S. Sivia and J. Skilling.
Data Analysis: A Bayesian Tutorial.
Oxford University Press, Oxford, 2006.

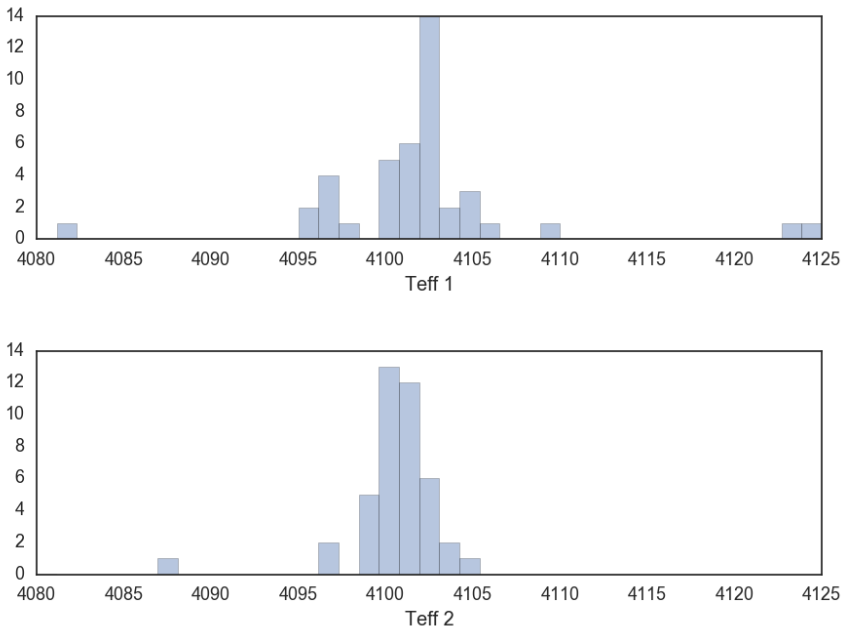
References IV


J. Vanderplas.
The Model Complexity Myth.
https://jakevdp.github.io/blog/2015/07/06/model-complexity-myth/, July 2015.

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail*
but at least they are objective and don't depend on priors

Consider two different effective temperature estimates
for the same 42 stars.



Are the means different?

t-test: t_{obs}=1.53, p=0.13

(`scipy.stats.ttest_ind`)

$p > 0.05 \rightarrow$ not significant

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail but at least they are objective and don't depend on priors*

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail*
but at least they are objective and don't depend on priors

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42
- p changes if we intended to sample from multiple groups

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail but at least they are objective and don't depend on priors*

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42
- p changes if we intended to sample from multiple groups
- p changes if we intended to sample until specific duration

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail but at least they are objective and don't depend on priors*

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42
- p changes if we intended to sample from multiple groups
- p changes if we intended to sample until specific duration
- p changes if sampling was interrupted

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail but at least they are objective and don't depend on priors*

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42
- p changes if we intended to sample from multiple groups
- p changes if we intended to sample until specific duration
- p changes if sampling was interrupted
- p changes if we got more data than we expected

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail but at least they are objective and don't depend on priors*

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42
- p changes if we intended to sample from multiple groups
- p changes if we intended to sample until specific duration
- p changes if sampling was interrupted
- p changes if we got more data than we expected
- p changes if we intended to ample until threshold t_{obs}

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail but at least they are objective and don't depend on priors*

$t_{obs} = 1.53$ is fixed given the data, but p is not!

What is hidden in the previous calculation?

- $p = 0.13$ if we intended to sample until N=42
- p changes if we intended to sample from multiple groups
- p changes if we intended to sample until specific duration
- p changes if sampling was interrupted
- p changes if we got more data than we expected
- p changes if we intended to ample until threshold t_{obs}

p depends on everything **except the actual data we have**!!

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail*
but at least they are objective and don't depend on priors

The problems with the Null-Hypothesis

ASTRONOMER: *the frequentist tests may be wrong in detail*
but at least they are objective and don't depend on priors

FUTURE YOU: Yeah, right...