Introduction to Bayesian Statistics and Markov Chain Monte Carlo (MCMC)

Teeraparb Chantavat

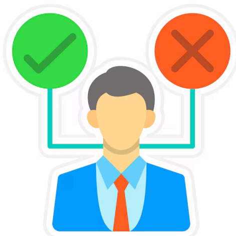
Institute for Fundamental Study
Naresuan University

COSCOM2024
15 - 21 December 2024
Kantary Bay, Rayong Thailand

Probability

- Probability as a measure of uncertainty
- Various applications in many fields e.g. physics, finance, gaming

Applications of Probability

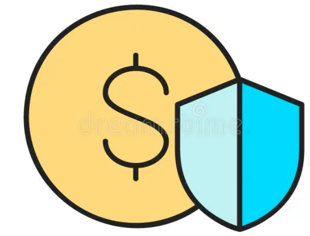


Decision Making

- Helps in making informed decisions under uncertainty.
- Enables risk assessment and management.

Scientific Research

- Used to analyze experimental data and draw conclusions.
- Aids in hypothesis testing and model building.



Insurance Industry

- Calculates premiums and assesses risks.

Applications of Probability

Weather Forecasting

- Forecasts future weather conditions based on historical data and statistical models.

Finance and Investing

- Predicts market trends and evaluates investment opportunities.

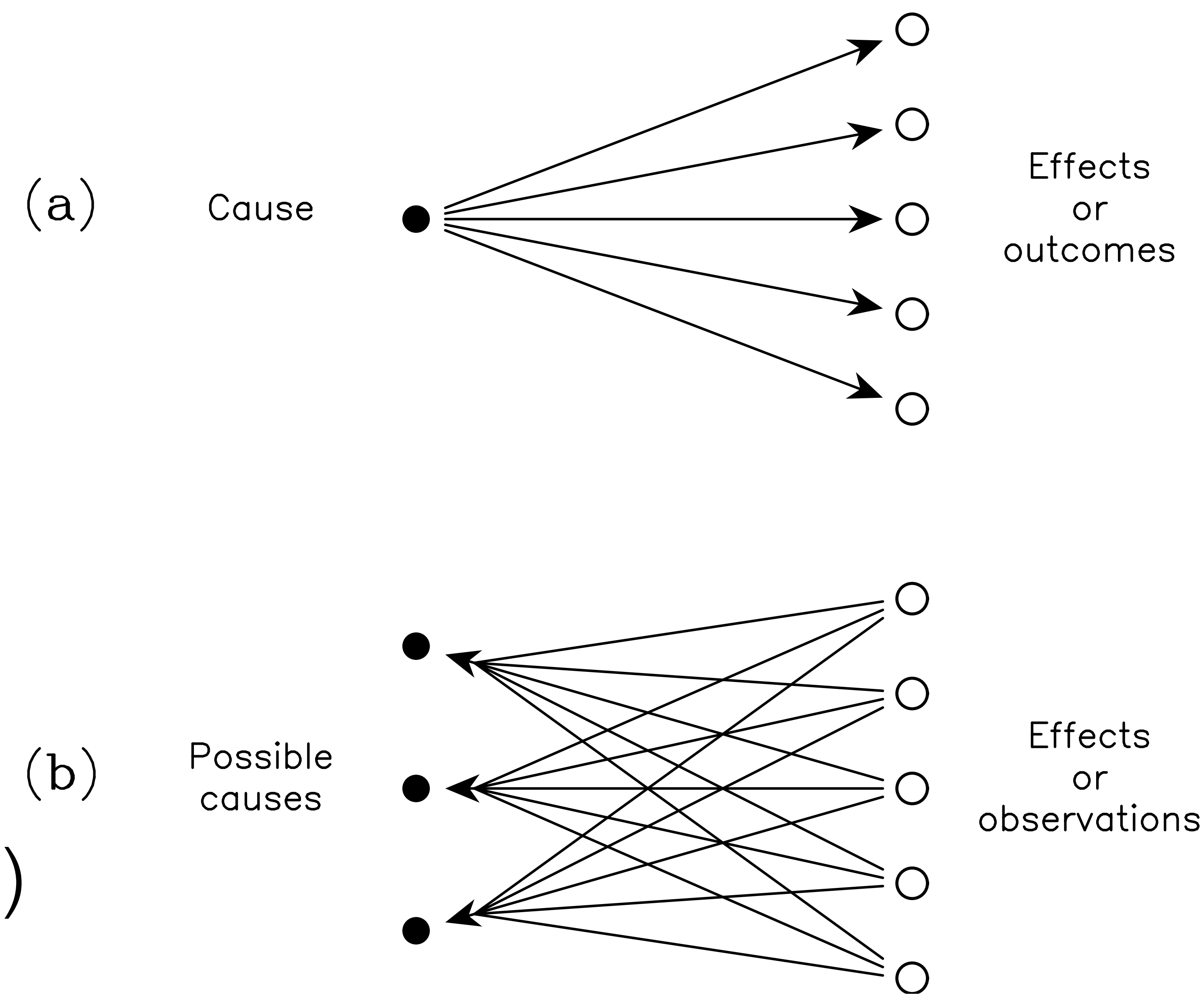
Quality Control

- Monitors product quality and identifies potential defects.

Deductive vs Inductive

Deductive Logic
(What you learn
in a science class)

Inductive Logic
(What science actually is)



Interpretation of Probability

There are different ways we can interpret probability:

- **Frequentist interpretation:**
  probability as an objective property of the world, defined as the long-run frequency of an event.

- **Bayesian interpretation:**
  probability as a degree of belief or uncertainty about a proposition. It can be updated as new evidence is obtained.

Frequentist vs Bayesian

- There are two distinct approaches to statistical inference, along with their underlying definitions of probability.

Approach	Frequentist	Bayesian
Definition of Probability	Probability is seen as the long-run relative frequency of an event occurring in repeated, independent experiments. It is based on objective, observable frequencies.	Probability is seen as a measure of belief or certainty about an event. It incorporates both prior knowledge and new evidence to update beliefs.

Frequentist vs Bayesian

Approach	Frequentist	Bayesian
Parameters	Parameters are fixed, unknown values. Inference is about estimating these fixed values based on observed data.	Parameters are considered random variables with probability distributions. Inference involves updating prior distributions with observed data to obtain posterior distributions.
Subjectivity	It is considered an objective approach, as probabilities are based on observed frequencies, and conclusions are not influenced by subjective beliefs.	Acknowledges subjectivity, as it allows the incorporation of prior beliefs. Bayesian inference is sensitive to the choice of priors.

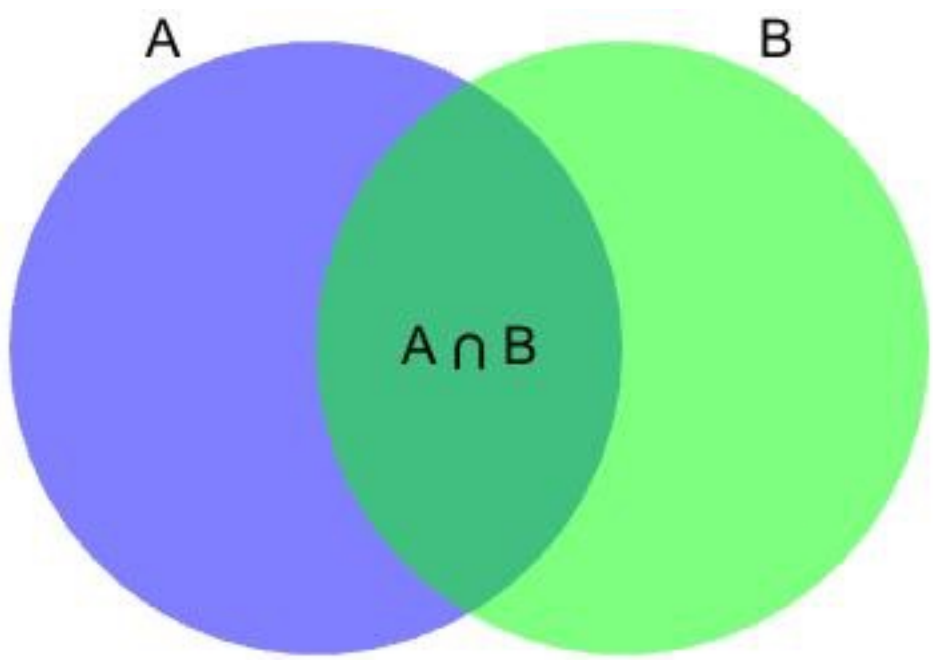
Frequentist vs Bayesian

Approach	Frequentist	Bayesian
Hypothesis Testing	Emphasizes hypothesis testing, focusing on rejecting or failing to reject null hypotheses based on the observed data.	While hypothesis testing is possible, Bayesian inference often focuses on estimating parameters and updating beliefs rather than strict hypothesis testing.
Prior Information	Typically does not incorporate prior beliefs or subjective information about parameters.	Incorporates prior information, allowing researchers to include existing knowledge or beliefs about parameters in the analysis.

Conditional Probability

Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



$P(A)$	Observing the data.
$P(B)$	The theory is true.
$P(A \mid B)$	The data is observed given that the theory is true

Bayesian Rule

Symmetry Rule

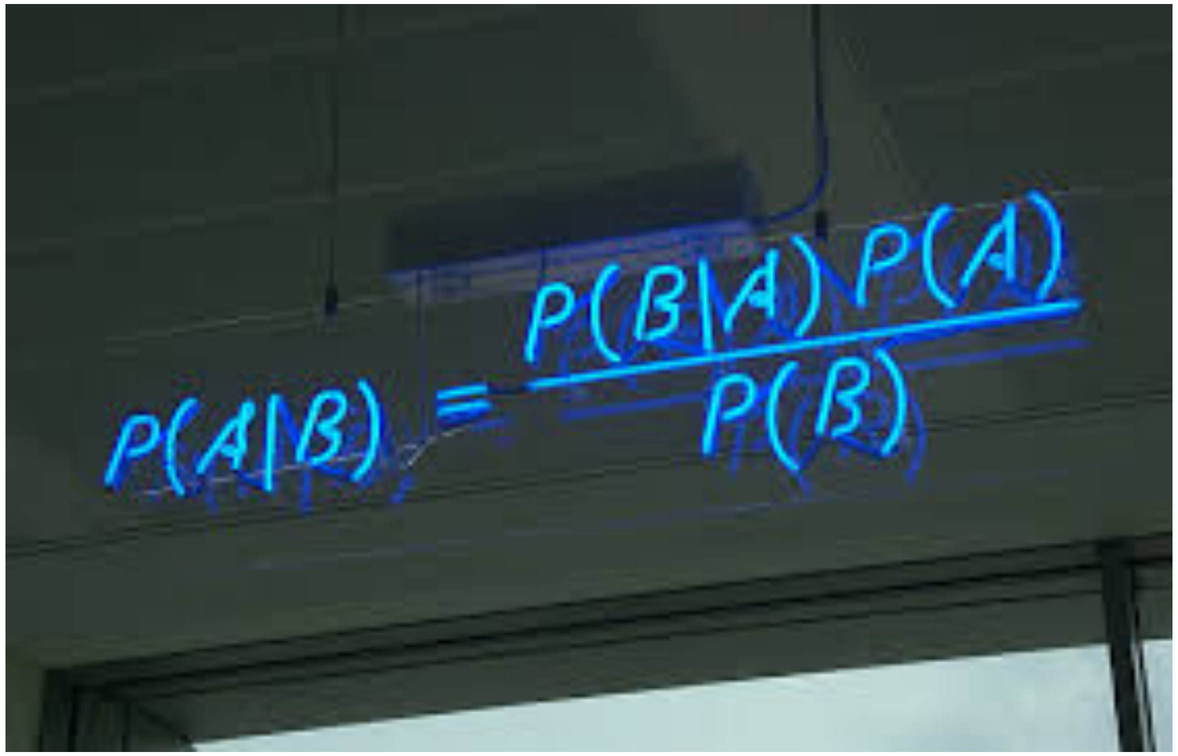
$$P(B \cap A) = P(A \cap B)$$

$$P(B \mid A)P(A) = P(A \mid B)P(B)$$

Bayesian Rule

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

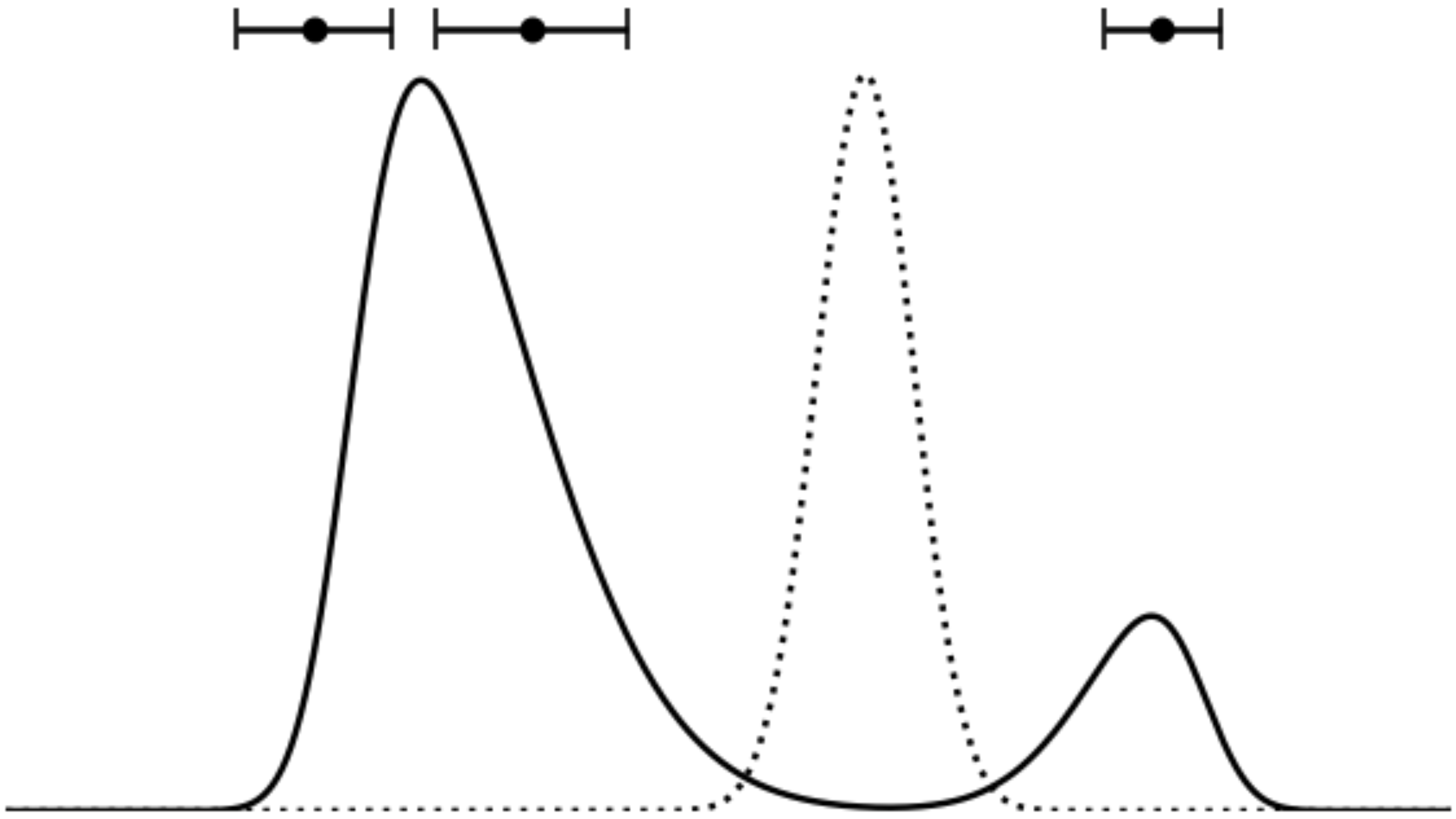
Bayesian Rule



"There are no problems left in statistics except the assessment of probability"

Lindley (2000)

Bayesian Statistics



Glossary

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$$

$P(H)$	Probability that the hypothesis is true.	**Prior**
$P(D \mid H)$	Probability that the data is observed given that the hypothesis is true.	**Likelihood**
$P(D)$	Probability that the collections of data is liable.	**Evidence**
$P(H \mid D)$	Probability that the hypothesis is true given that the data is true.	**Posterior**

Hypothesis Space

- A theory usually have many parameters, for example a two-parameter model

$$\boldsymbol{\Theta} = \{\Theta_1, \Theta_2\}$$

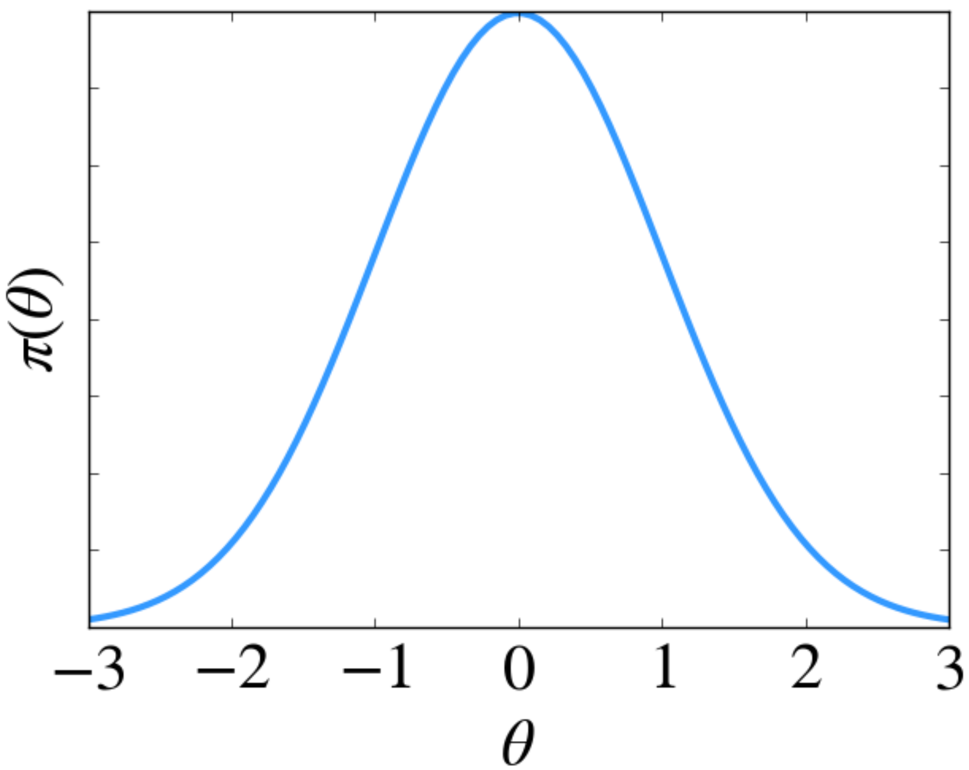
- The **hypothesis** is the assumption that the parameter

$$\text{Hypothesis 1} \quad (H_1) \quad : \quad \theta_1 = 1.0, \quad \theta_2 = 1.2$$

$$H_1 \quad \equiv \quad \boldsymbol{\theta}_1 = \{\theta_1, \theta_2\}$$

Prior Probability

- The **prior probability** is the distribution of the parameters we know before the experiment **(degree of belief)**.
- We can have a uniform distribution for total ignorance or a normal distribution if mean and standard deviation are given.



Likelihood

- In most cases, we are working with the logarithm of the likelihood function called **log-likelihood.**

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})$$

- Expanding around the maximum of the log-likelihood at $\boldsymbol{\theta}_0$ i.e.

$$\left.\frac{\partial L}{\partial \theta_\alpha}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0$$

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = L(\boldsymbol{x}|\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{\alpha,\beta} \left.\frac{\partial^2 L}{\partial \theta_\alpha \partial \theta_\beta}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \left(\theta_\alpha - \theta_{\alpha 0}\right)\left(\theta_\beta - \theta_{\beta 0}\right).$$

Likelihood

- We define the **precision matrix** $\boldsymbol{P}$ as

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = L(\boldsymbol{x}|\boldsymbol{\theta}) - \frac{1}{2}\big(\boldsymbol{\theta} - \boldsymbol{\theta}_0\big)^T \cdot \boldsymbol{P} \cdot \big(\boldsymbol{\theta} - \boldsymbol{\theta}_0\big),$$

where

$$P_{\alpha\beta} \equiv -\frac{\partial^2 L}{\partial\theta_\alpha \partial\theta_\beta}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

- The **likelihood** is then given by

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\big(\boldsymbol{\theta} - \boldsymbol{\theta}_0\big)^T \cdot \boldsymbol{P} \cdot \big(\boldsymbol{\theta} - \boldsymbol{\theta}_0\big)\right).$$

Likelihood

- The inverse of the **precision matrix** is called covariance matrix

$$\boldsymbol{C} \equiv \boldsymbol{P}^{-1}$$

then

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \cdot \boldsymbol{C}^{-1} \cdot (\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right).$$

- The variance of the parameter can be estimated as

$$\mathrm{Var}(\theta_\alpha) = C_{\alpha\alpha} = \sigma^2_{\theta_\alpha}.$$

Marginalization

Suppose that we have a proposition B with its negative counterpart $\overline{B}$. From the sum rule

$$P(A, B|I) + P(A, \overline{B}|I) = P(A|I).$$

This is called **marginalisation**.

$$P(A, B_1|I) + P(A, B_2|I) + \ldots + P(A, B_N|I) = 1,$$

or

$$\int \mathrm{d}B P(A, B|I) = P(A|I).$$

Evidence

- The evidence is usually considered as a normalization constants — nothing to do with **parameter estimations**.

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\boldsymbol{\pi}(\boldsymbol{\theta})$$

- However, the evidence is important for **model comparison** .

$$\mathcal{P}(\boldsymbol{\theta}_1|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}_1)\boldsymbol{\pi}(\boldsymbol{\theta}_1)}{D(\boldsymbol{x}|\mathcal{M}_1)}, \quad \mathcal{P}(\boldsymbol{\theta}_2|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}_2)\boldsymbol{\pi}(\boldsymbol{\theta}_2)}{D(\boldsymbol{x}|\mathcal{M}_2)}$$

Evidence

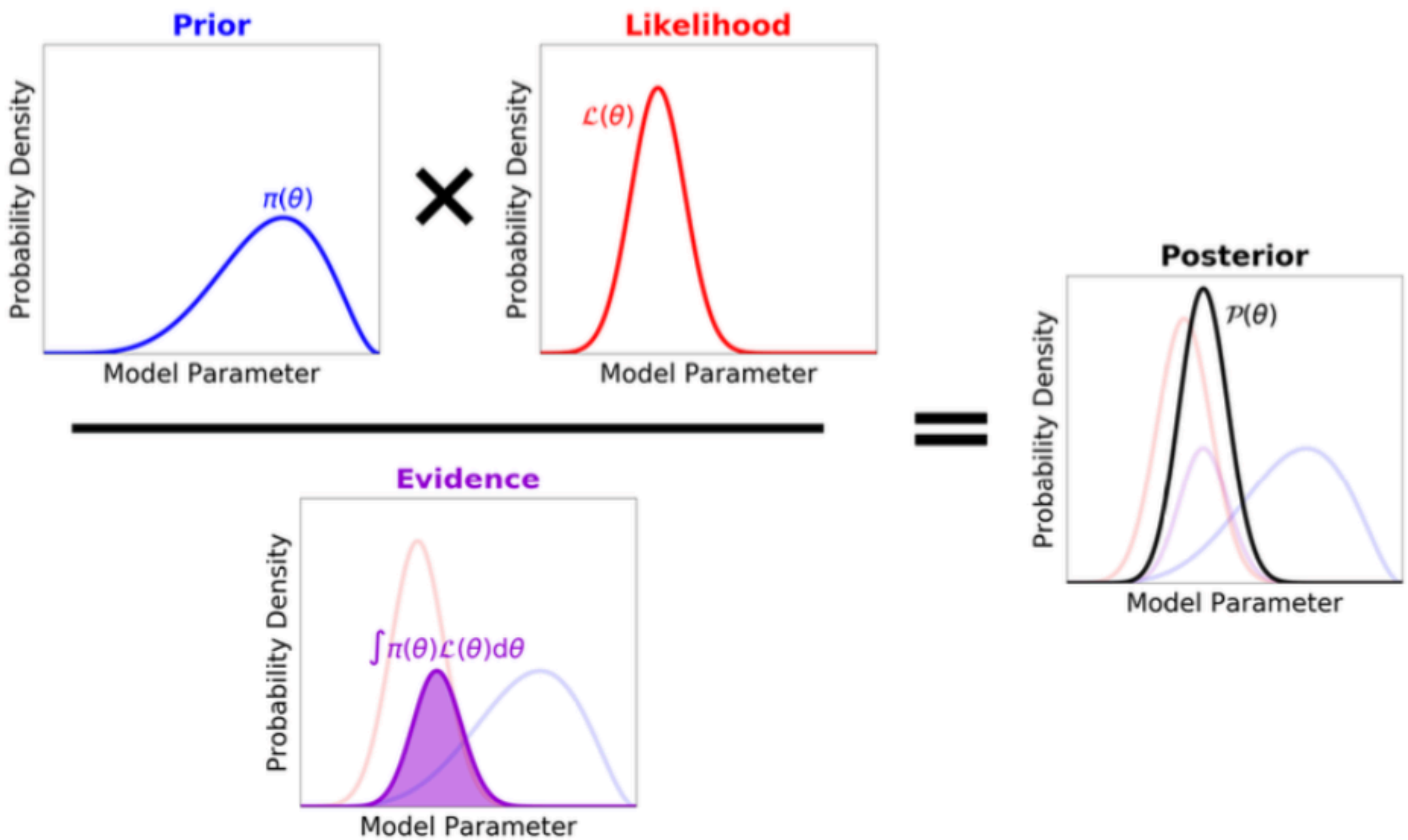
The evidence could be computed by marginalize over the hypothesis space.

$$\mathcal{Z} = \int \mathcal{L}(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} \equiv \int \tilde{\mathcal{P}}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta}$$

where $\tilde{\mathcal{P}}(\mathbf{\Theta}) \equiv \mathcal{L}(\mathbf{\Theta})\pi(\mathbf{\Theta})$ is the unnormalized posterior.

Posterior Probability

- This is the revised probability of the event or hypothesis after considering the new data.



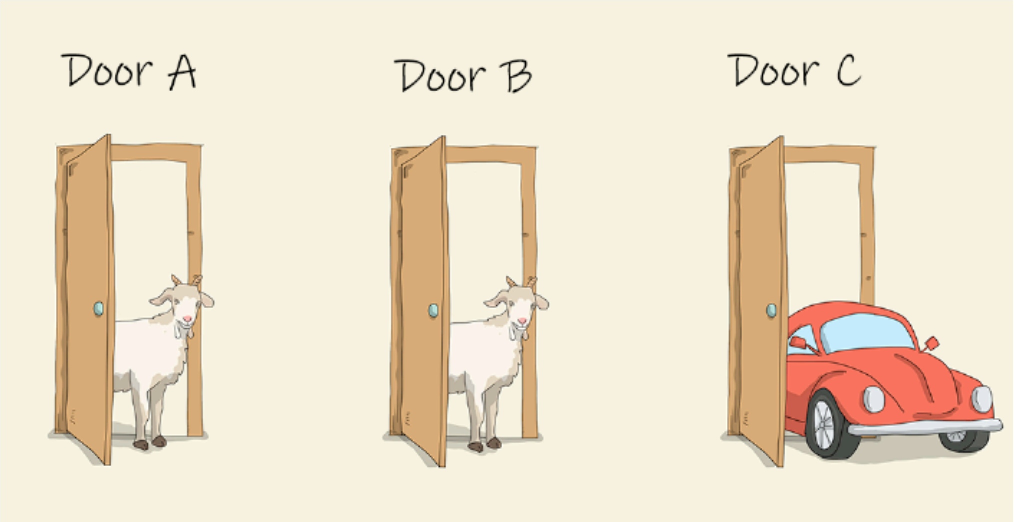
Bayesian Rule

$$P(\text{Hypothesis}|\text{Data, Prior Information}) \propto P(\text{Data}|\text{Hypothesis, Prior Information}) \times P(\text{Hypothesis}|\text{Prior Information})$$

OR

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Example: Monte Hall Problem

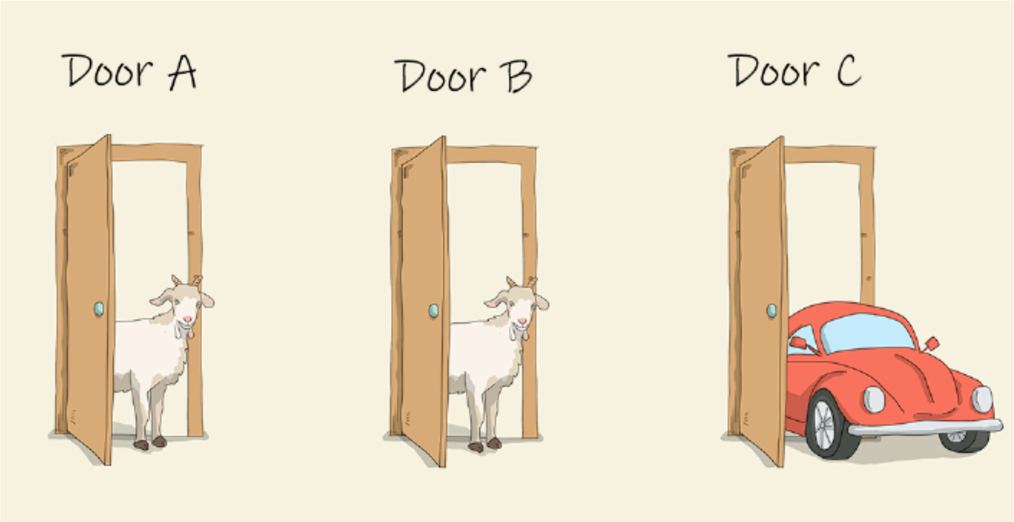


- Monty shows you three closed doors and tells you that there is a prize behind each door: one prize is a car the other two are less valuable prizes like goats. The prizes are arranged at random.

- The object of the game is to guess which door has the car. If you guess right, you get to keep the car.

- You pick a door, which we will call Door A. We'll call the other doors B and C.

Example: Monte Hall Problem

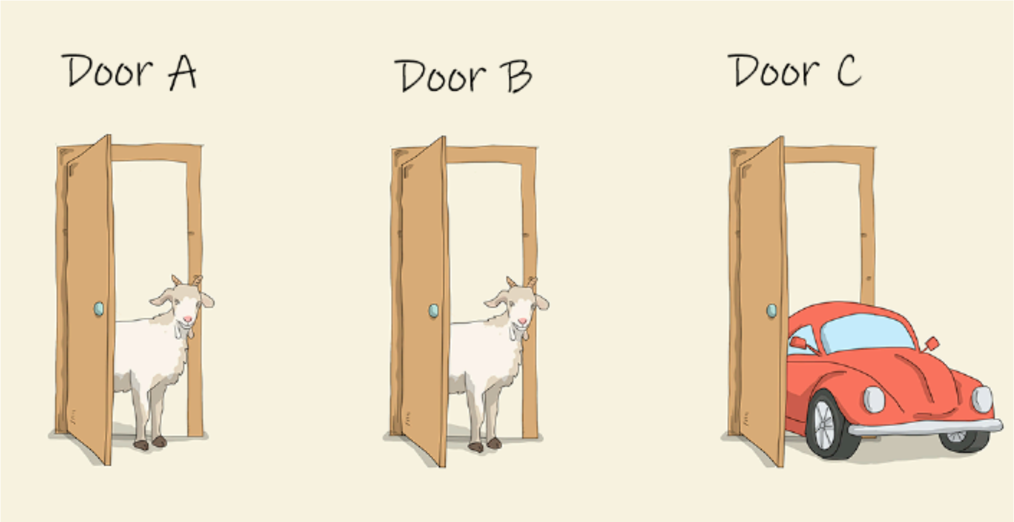


- Before opening the door you chose, Monty increases the suspense by opening either Door B or C, whichever does not have the car. (If the car is actually behind Door A, Monty can safely open B or C, so he chooses one at random.)

- Then Monty offers you the option to stick with your original choice or switch to the one remaining unopened door.

The question is, should you **stick** or **switch** or does it make no difference?

Example: Monte Hall Problem



Choice	Prior $\pi(\Theta)$	Likelihood $\mathcal{L}(\Theta)$	$\pi(\Theta) \cdot \mathcal{L}(\Theta)$	Posterior $\mathcal{P}(\Theta \| D)$
A	$1/3$	$1/2$	$1/6$	$1/3$
B	$1/3$	0	0	0
C	$1/3$	1	$1/3$	$2/3$

What are Posteriors Good for?

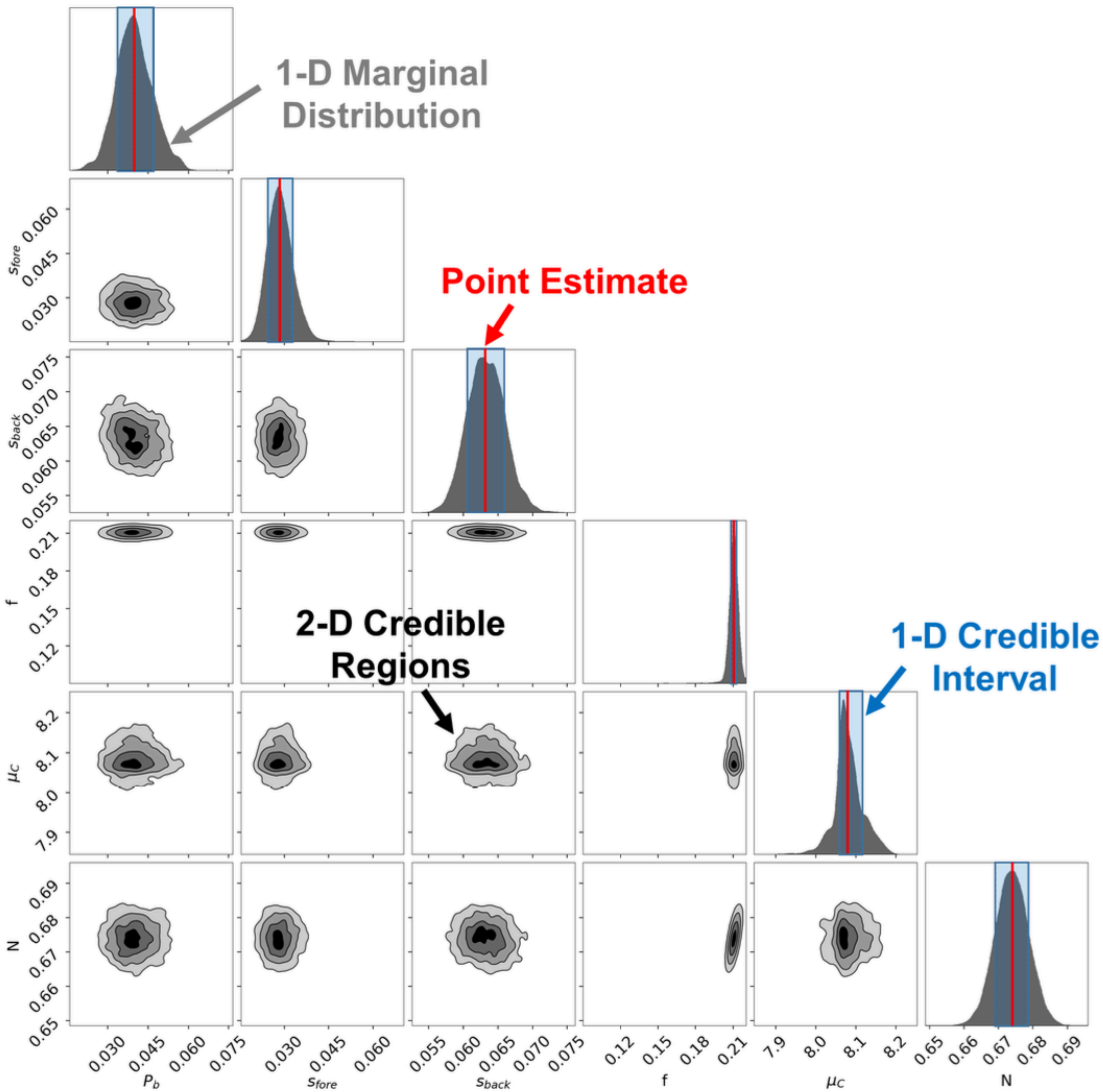
- **Making educated guesses:**
  This is the revised probability of the event or hypothesis after considering the new data.

- **Quantifying uncertainty:**
  Provide constraints on the range of possible model parameter values.

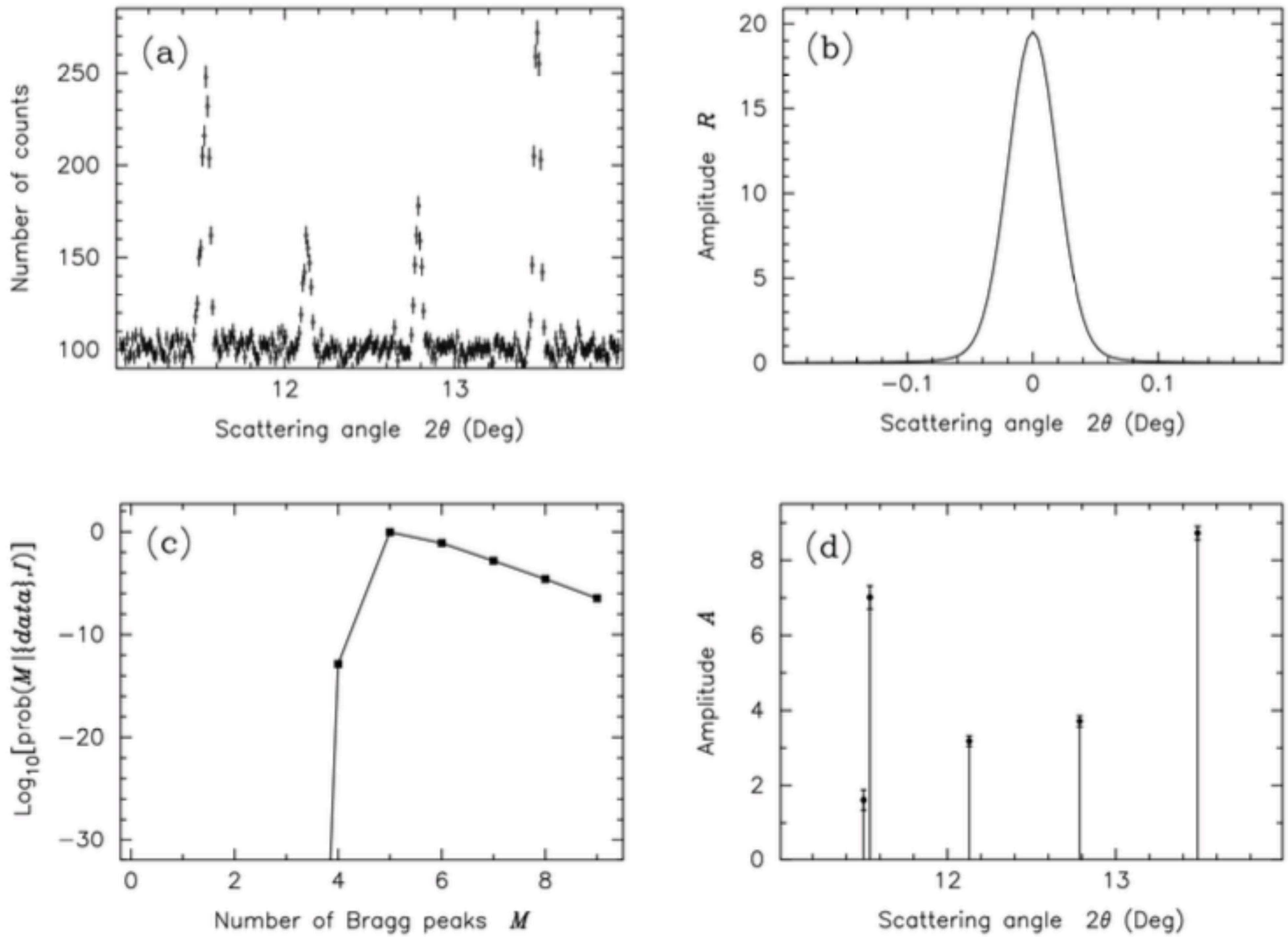
- **Generating predictions:**
  Predict observables or other variables that depend on the model parameters.

- **Comparing models:**
  Use the evidences from different models to determine which models are more favorable.

2D Marginalized Posterior PDF



Model Comparison



Variance and Covariance

Variance is the average of the square deviation from the mean of a parameter,

$$\mathrm{Var}\left(\theta\right) = \mathrm{E}\Big(\left(\theta - \mathrm{E}\left(\theta\right)\right)^2\Big).$$

Covariance is the average of the joint deviation from the mean of two parameters,

$$\mathrm{Cov}\left(\theta_i, \theta_j\right) = \mathrm{E}\Big(\left(\theta_i - \mathrm{E}\left(\theta_i\right)\right)\left(\theta_j - \mathrm{E}\left(\theta_j\right)\right)\Big).$$

Covariance Matrix

- The **covariance matrix** is related to the correlation matrix

$$\boldsymbol{C} = \begin{pmatrix} \mathrm{Var}(\theta_1) & \dots & \mathrm{Cov}(\theta_1, \theta_n) \\ \vdots & \ddots & \\ \mathrm{Cov}(\theta_n, \theta_1) & \dots & \mathrm{Var}(\theta_n) \end{pmatrix}$$

$$\boldsymbol{C} = \begin{pmatrix} \sigma_{\theta_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{\theta_n} \end{pmatrix} \cdot \boldsymbol{R} \cdot \begin{pmatrix} \sigma_{\theta_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{\theta_n} \end{pmatrix}$$

where $\boldsymbol{R}$ is the **correlation matrix**.

Correlation Matrix

- The **correlation matrix**

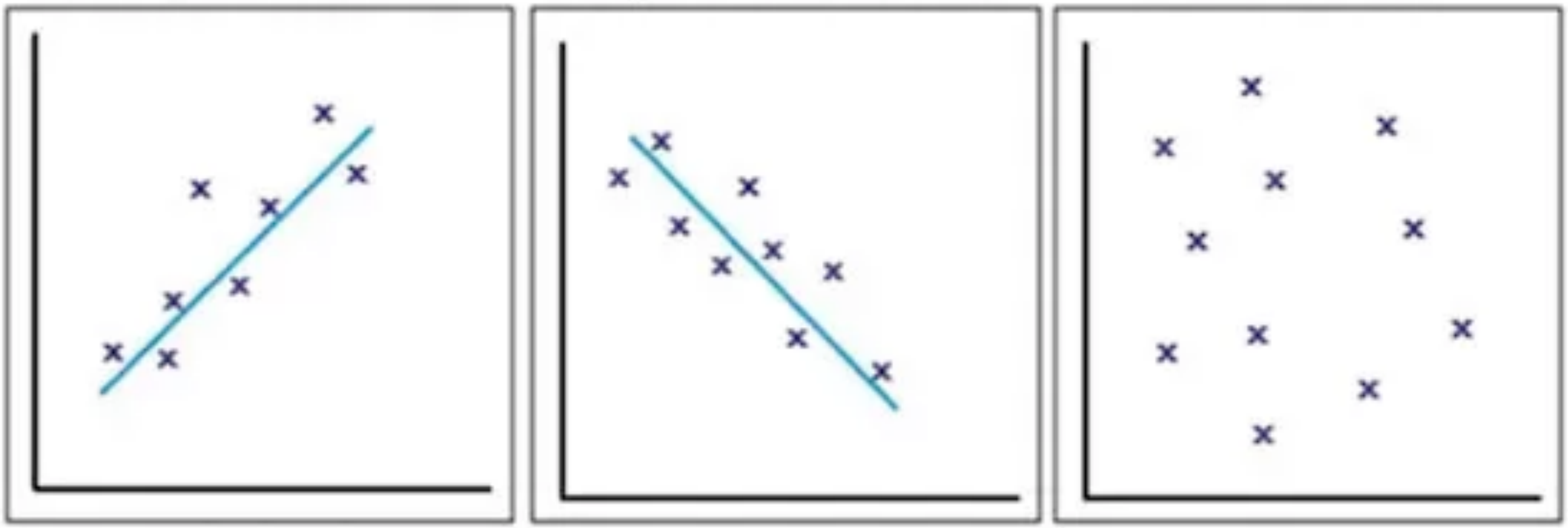
$$\boldsymbol{R} = \begin{pmatrix} 1 & \dots & \mathrm{Corr}(\theta_1, \theta_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Corr}(\theta_n, \theta_1) & \dots & 1 \end{pmatrix}$$

- where

$$\mathrm{Corr}(\theta_\alpha, \theta_\beta) = \frac{\mathrm{Cov}(\theta_\alpha, \theta_\beta)}{\sigma_{\theta_\alpha}\sigma_{\theta_\beta}}$$

Correlations

Positive correlation	Negative correlation	No correlation
The points lie close to a straight line, which has a positive gradient.	The points lie close to a straight line, which has a negative gradient.	There is no pattern to the points.
This shows that as one variable **increases** the other **increases**.	This shows that as one variable **increases**, the other **decreases**.	This shows that there is **no connection** between the two variables.



Parameter Estimation

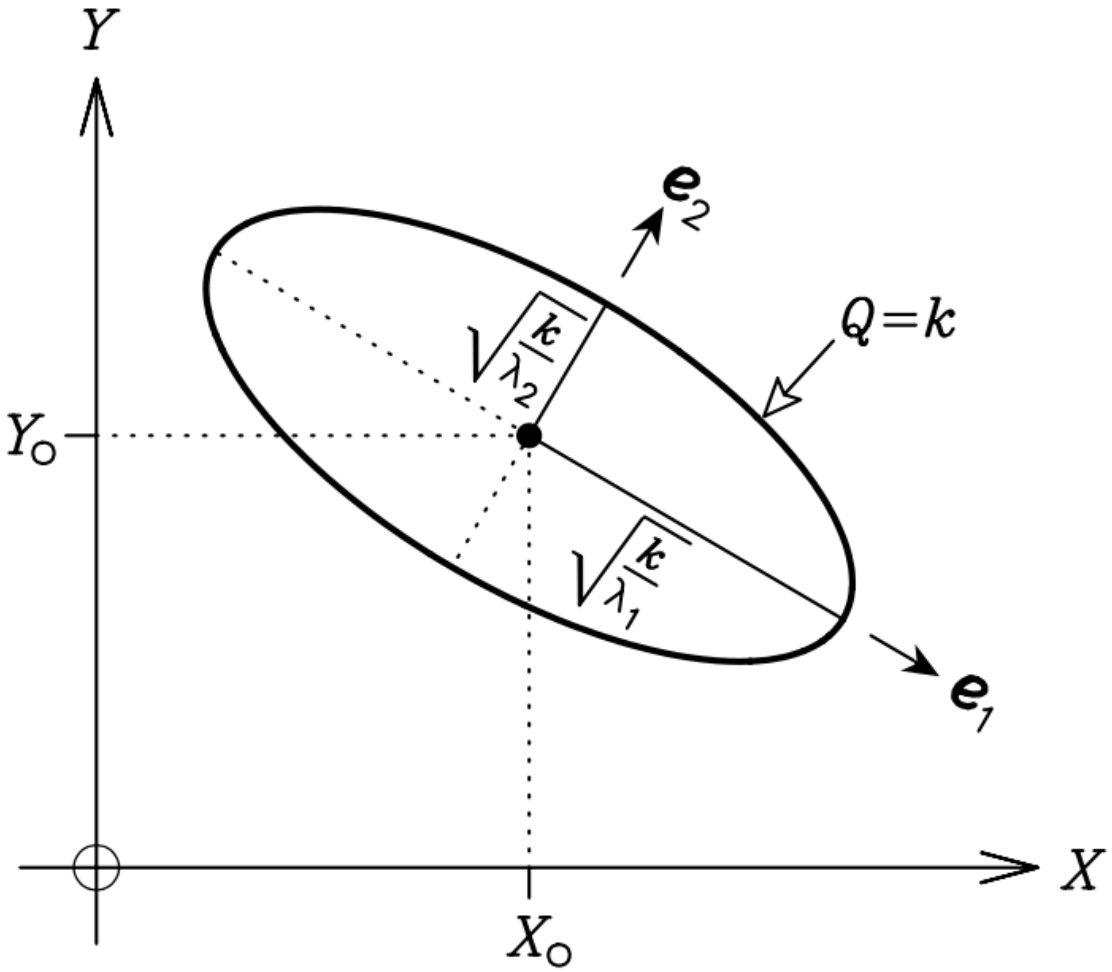
- The posterior encodes our inference about the parameter in the model, given the data and the relevant background information.

- We wish to summarize this with just two numbers: the best estimate (**mean**) and a measure of its reliability (**deviation**).

- With posterior we could either calculate the average value or the maximum likelihood value;

$$\left.\frac{\mathrm{d}\mathcal{P}}{\mathrm{d}\theta}\right|_{\theta=\theta_0} = 0 \quad \text{or} \quad \nabla_{\boldsymbol{\Theta}}\mathcal{P} = 0.$$

Parameter Estimation



- The approximation 2D marginalized probability density will an ellipse.

- For a 1sigma confidence region (68%)

$$\chi^2_{1\sigma} = 2.30$$

- Other confidence regions

$$\chi^2_{2\sigma} = 6.18$$

$$\chi^2_{3\sigma} = 11.83$$

Approximating Posterior with Grids

- The posterior pdf are usually has no analytic form, which we will have to use numerical method to approximate the posterior.
- In 1D, we can approximate it using standard numerical techniques such as a **Riemann sum** over a **discrete grid** of points:

$$\mathbb{E}_{\mathcal{P}}\left(f(\mathbf{\Theta})\right) = \int f(\mathbf{\Theta})\mathcal{P}(\mathbf{\Theta})\mathrm{d}\mathbf{\Theta} \approx \sum_{i=1}^{n} f(\mathbf{\Theta}_i)\mathcal{P}(\mathbf{\Theta}_i)\Delta\mathbf{\Theta}_i$$

where

$$\Delta\mathbf{\Theta}_i = \mathbf{\Theta}_{j+1} - \mathbf{\Theta}_j$$

Approximating Posterior with Grids

- We could take the mid points as the sampling points:

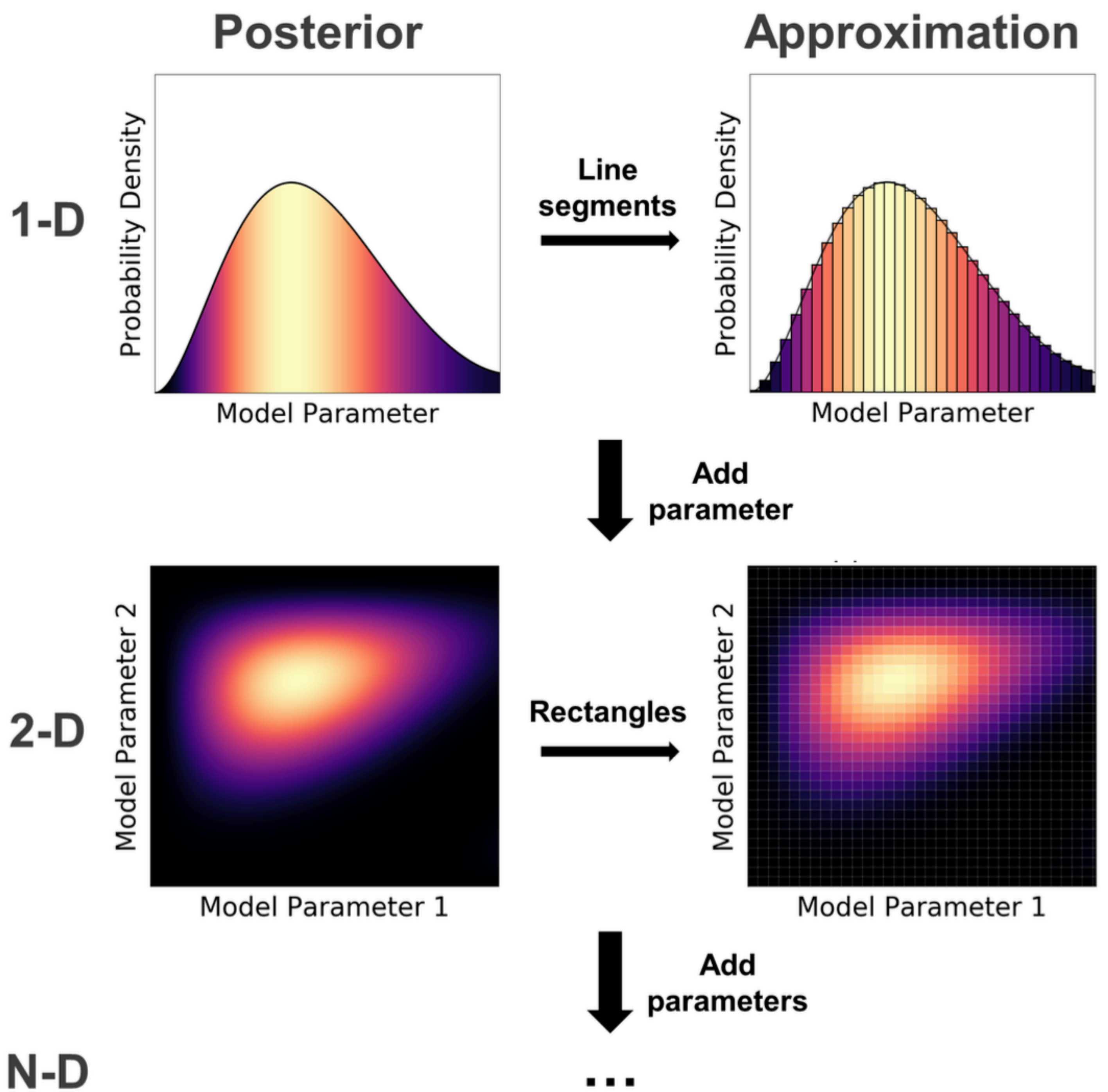
$$\boldsymbol{\Theta}_i = \frac{\boldsymbol{\Theta}_{j+1} + \boldsymbol{\Theta}_j}{2}$$

- We could generalized to higher dimension in a similar way,

$$\Delta\boldsymbol{\Theta}_i = \prod_{j=1}^{d} \Delta\Theta_{i,j}$$

- However the number of sampling points will increase exponentially - this is the **curse of dimensionality**.

Approximating Posterior with Grids



Effective Sampling Size

- Uniform sampling method has a drawback of spending a lot of computational time on the region with low probability i.e. $\tilde{\mathcal{P}}(\boldsymbol{\Theta})$ is small.
- For high dimensional space, most of the volume will have low probability.
- We will take the posterior into account as the weight of the grid point;

$$w_i \equiv \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i$$

Effective Sampling Size

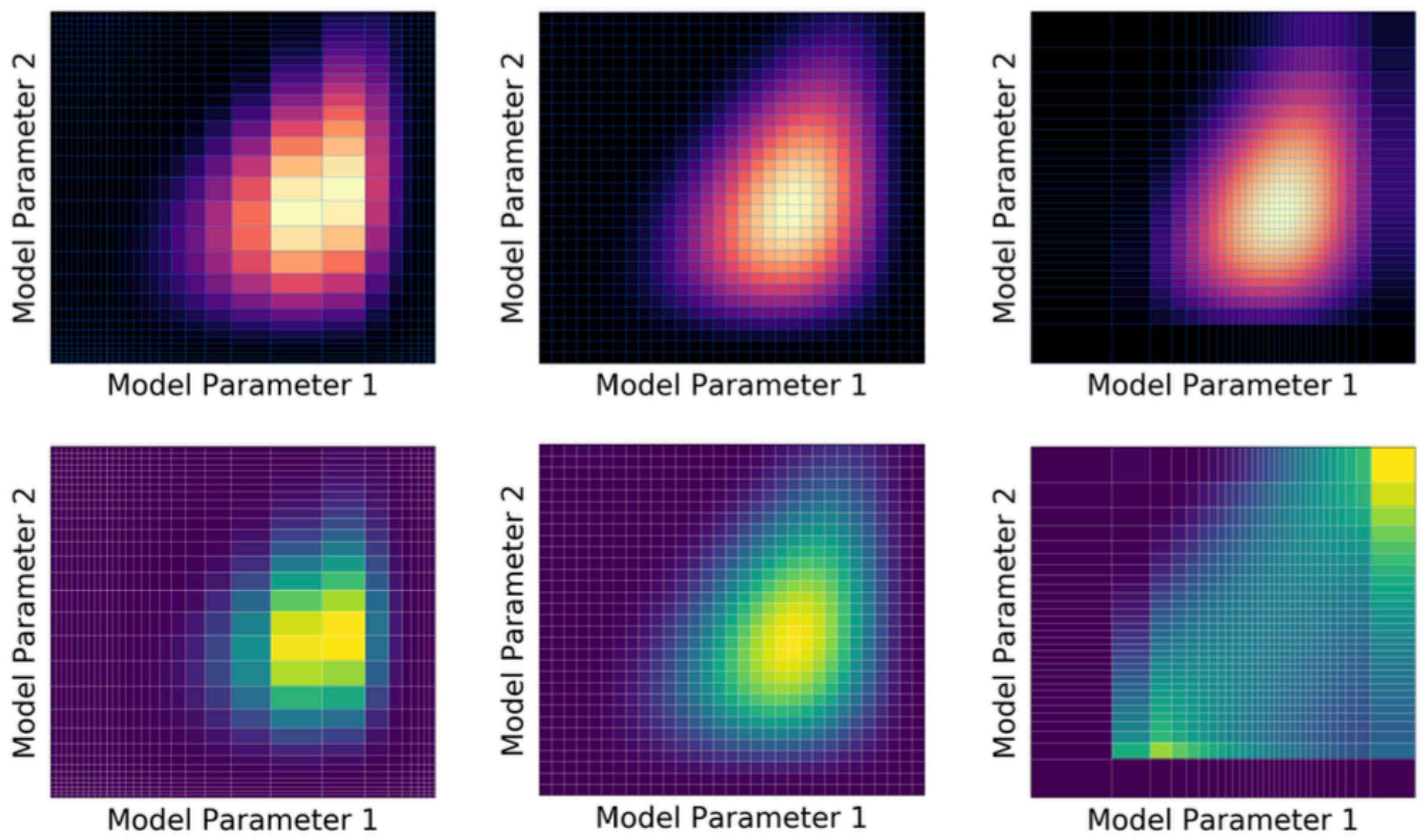
Poor Spacing
ESS ~ 60

Uniform Spacing
ESS ~ 370

Optimal Spacing
ESS ~ 830

Posterior Approximation (30 x 30 grid)

Estimated Weights



Convergence and Consistency

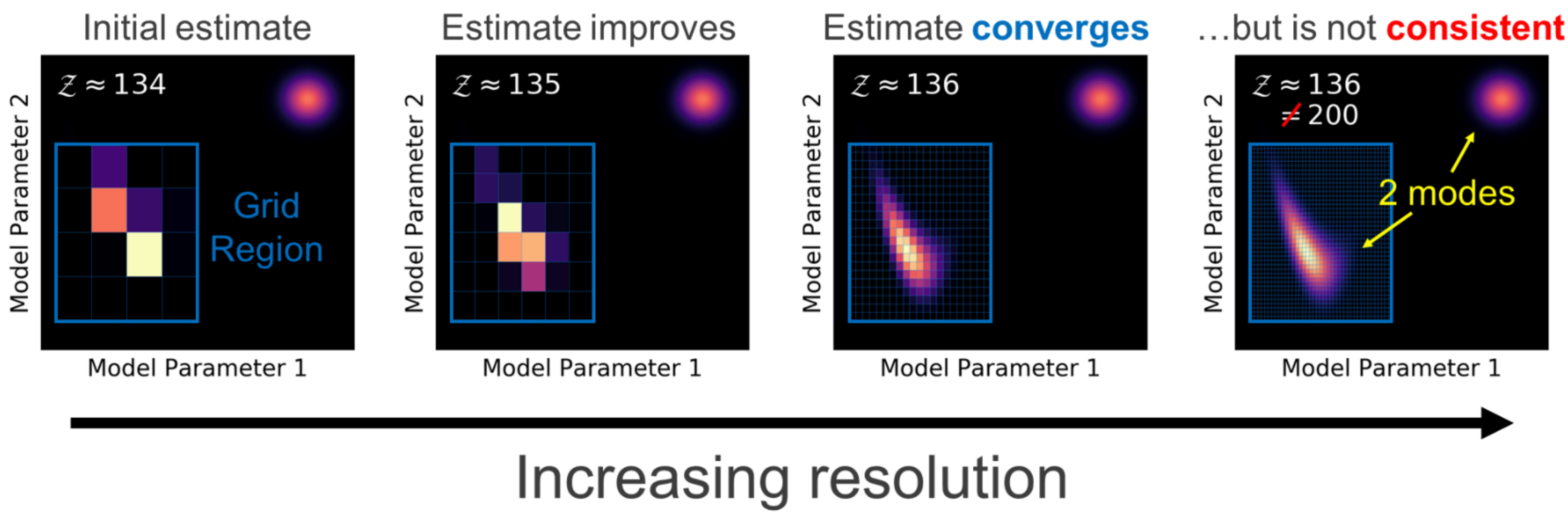
- **Convergence** is the idea that, while our estimates using n samples (grid points) might be noisy, it approaches some fiducial value as $n \to \infty$:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} = C$$

- **Consistency** is subsequently the idea that the value we converge to is the true value we are interested in estimating:

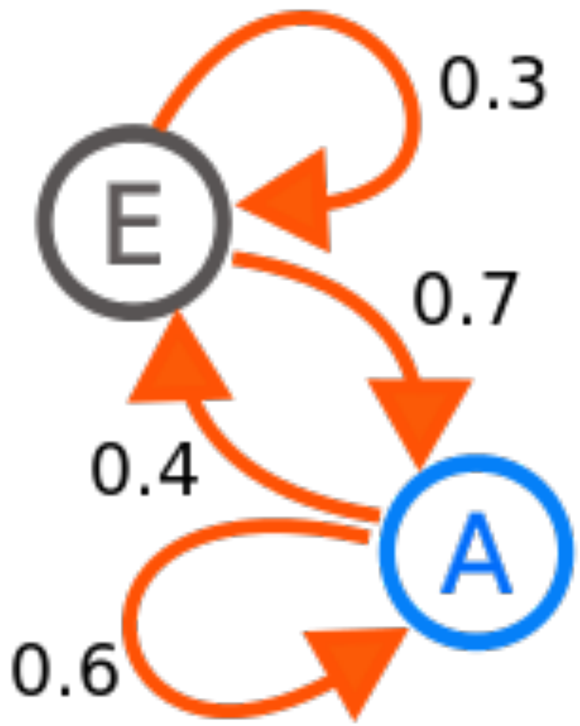
$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(\boldsymbol{\Theta}_i)\tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i}{\sum_{i=1}^{n} \tilde{\mathcal{P}}(\boldsymbol{\Theta}_i)\Delta\boldsymbol{\Theta}_i} = \mathbb{E}_{\mathcal{P}}\left(f(\boldsymbol{\Theta})\right)$$

Convergence and Consistency



Markov Chain Monte Carlo (MCMC)

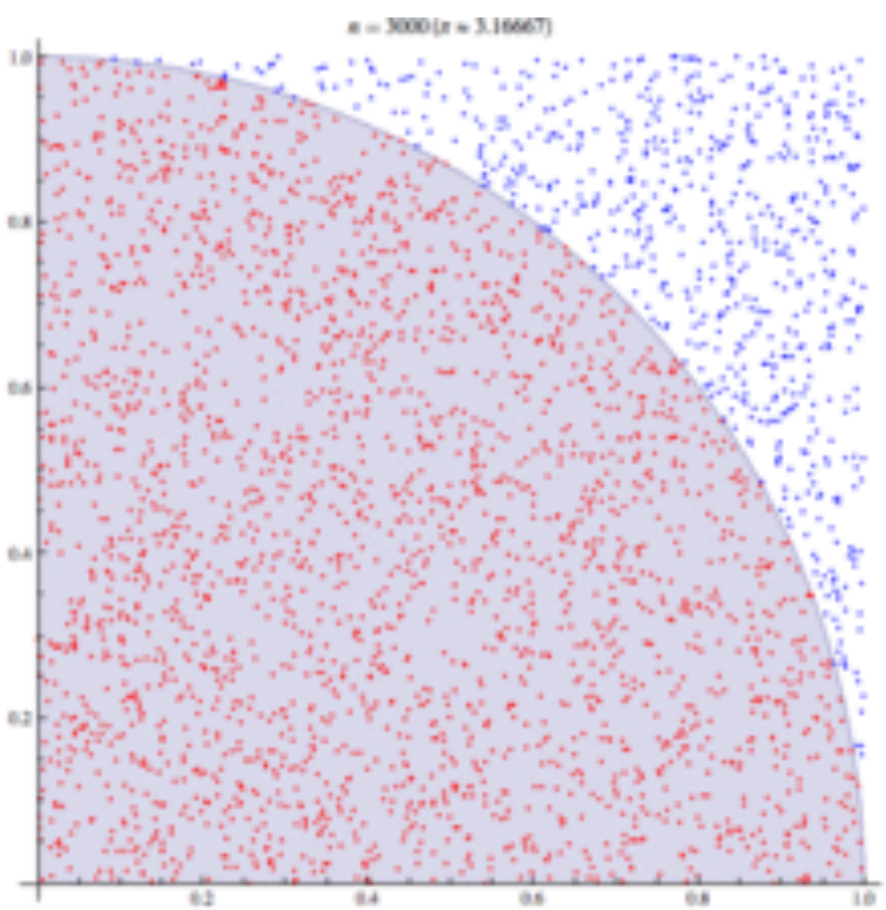
- A Markov chain is a chain of states in a parameter space that is “memoryless” (Markov property).

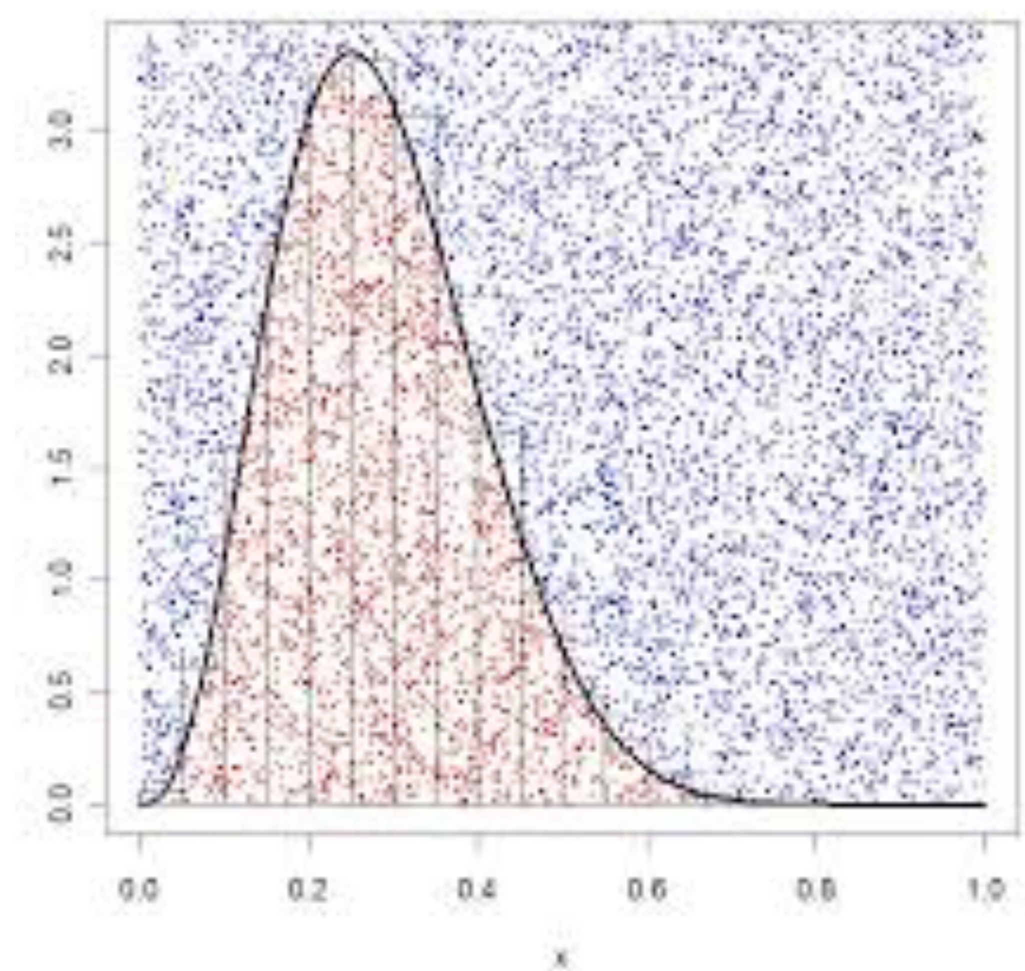


How the stage change depends only on the **current state**.

Markov Chain Monte Carlo (MCMC)

- A Monte Carlo is a method using random walk to generate the output. (rejection sampling method)





Metropolis-Hasting Algorithm

- The **Metropolis-Hastings algorithm** is an algorithm for random walks that will eventually converge to a true distribution of the parameter space.

$$P(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1)\, q(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2)$$

transitional probability ↑

prior probability ↑

proposal distribution ↑

Metropolis-Hasting Algorithm

The change of state from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$ is governed by the **acceptance rate**

$$\alpha(\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_2)\, q(\boldsymbol{\theta}_2 \to \boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_1)\, q(\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2)}\right\}$$

We are assumed an equilibrium state; hence,

$$q(\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2) = q(\boldsymbol{\theta}_2 \to \boldsymbol{\theta}_1)$$

Therefore,

$$\alpha(\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_2)}{\pi(\boldsymbol{\theta}_1)}\right\}$$

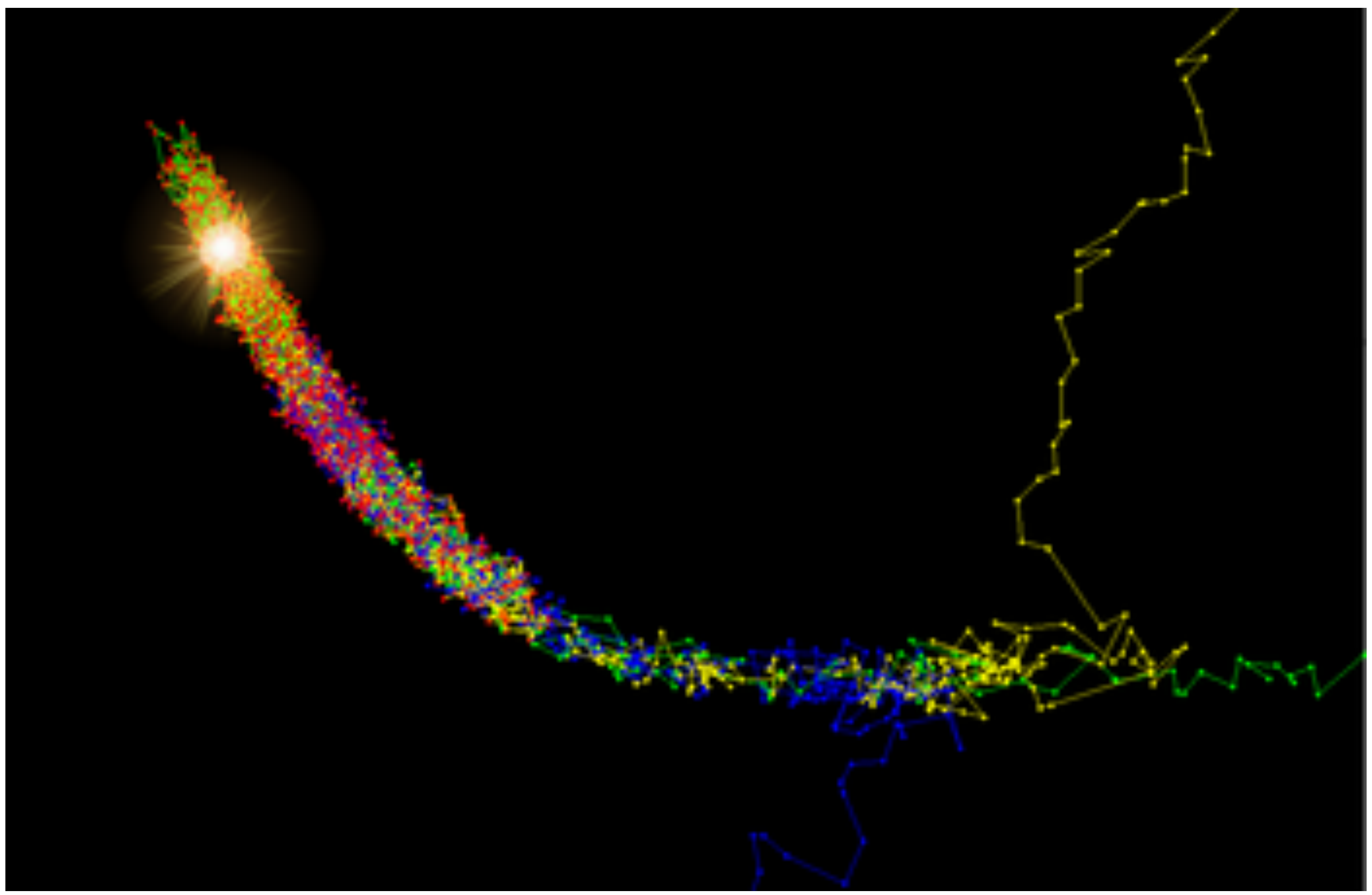
Metropolis-Hasting Algorithm

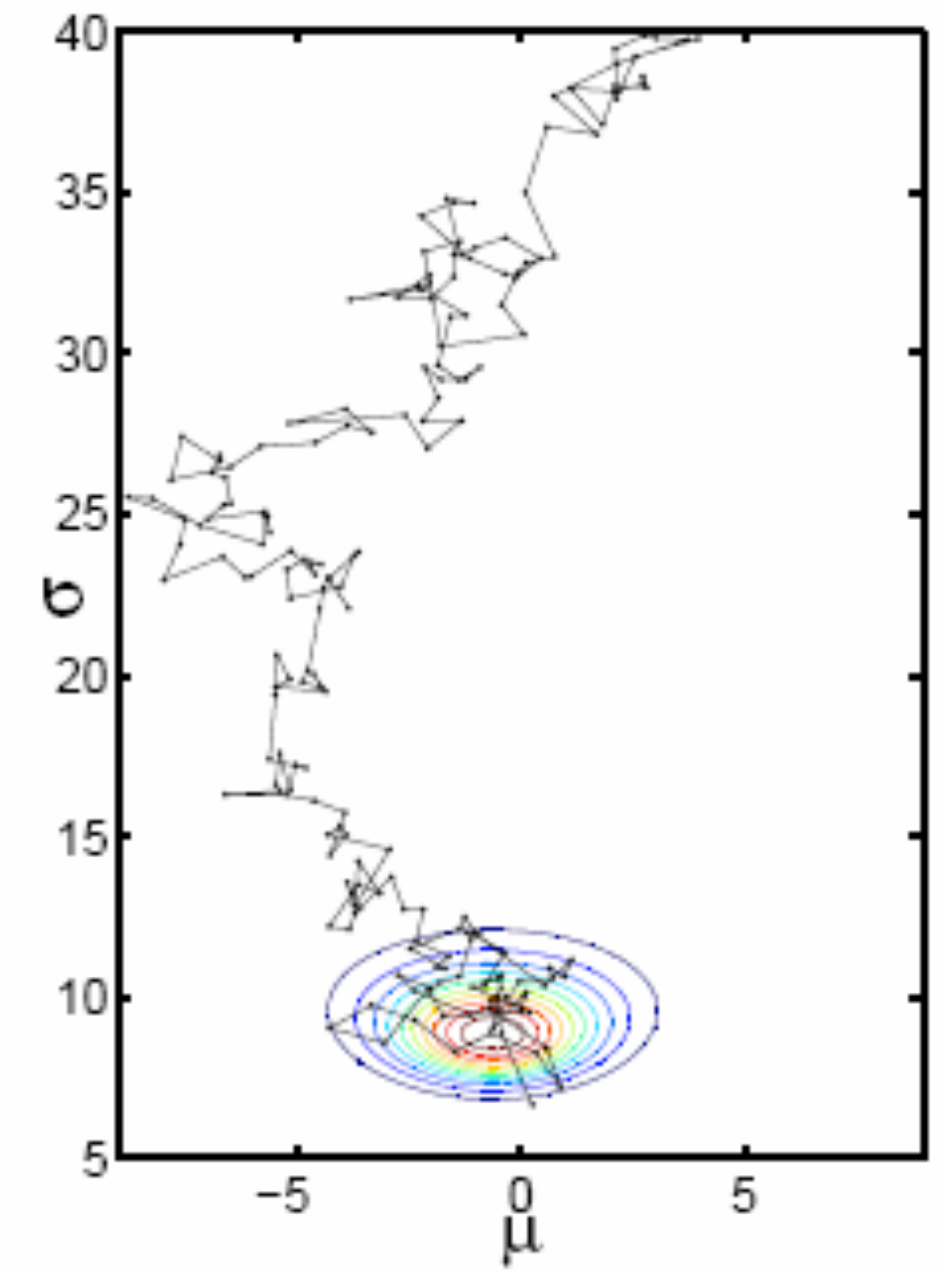
Pseudo code for Metropolis-Hastings Algorithm

```
alpha = likelihood2 / likelihood1;
if alpha > 1:
    jump to the new state;
else:
    if alpha > rand();
        jump to the new state;
    else:
        remain in the same state;
```

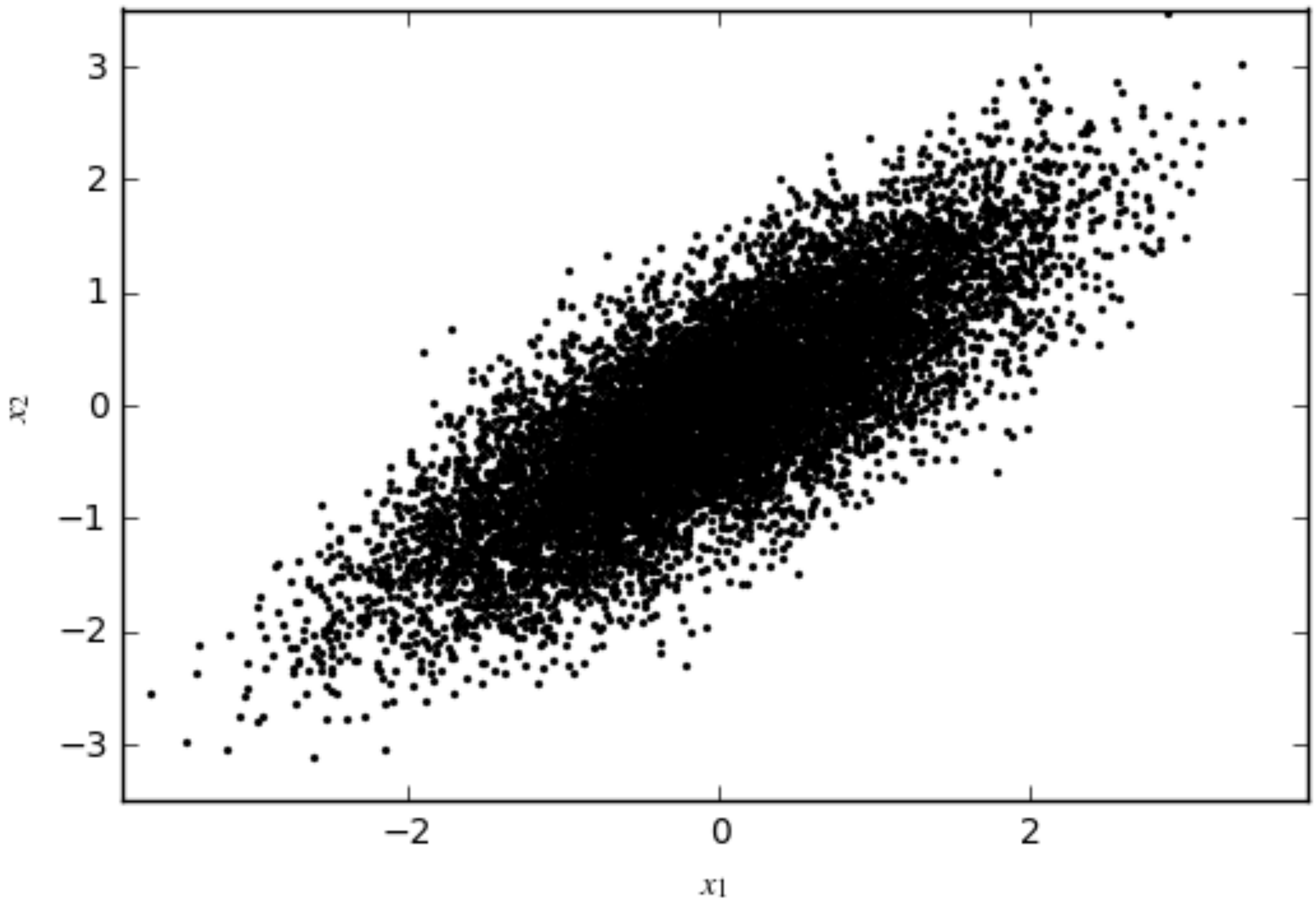
MCMC Chains

The chain will take some time to stabilize. This is called the **burn-in phase**.





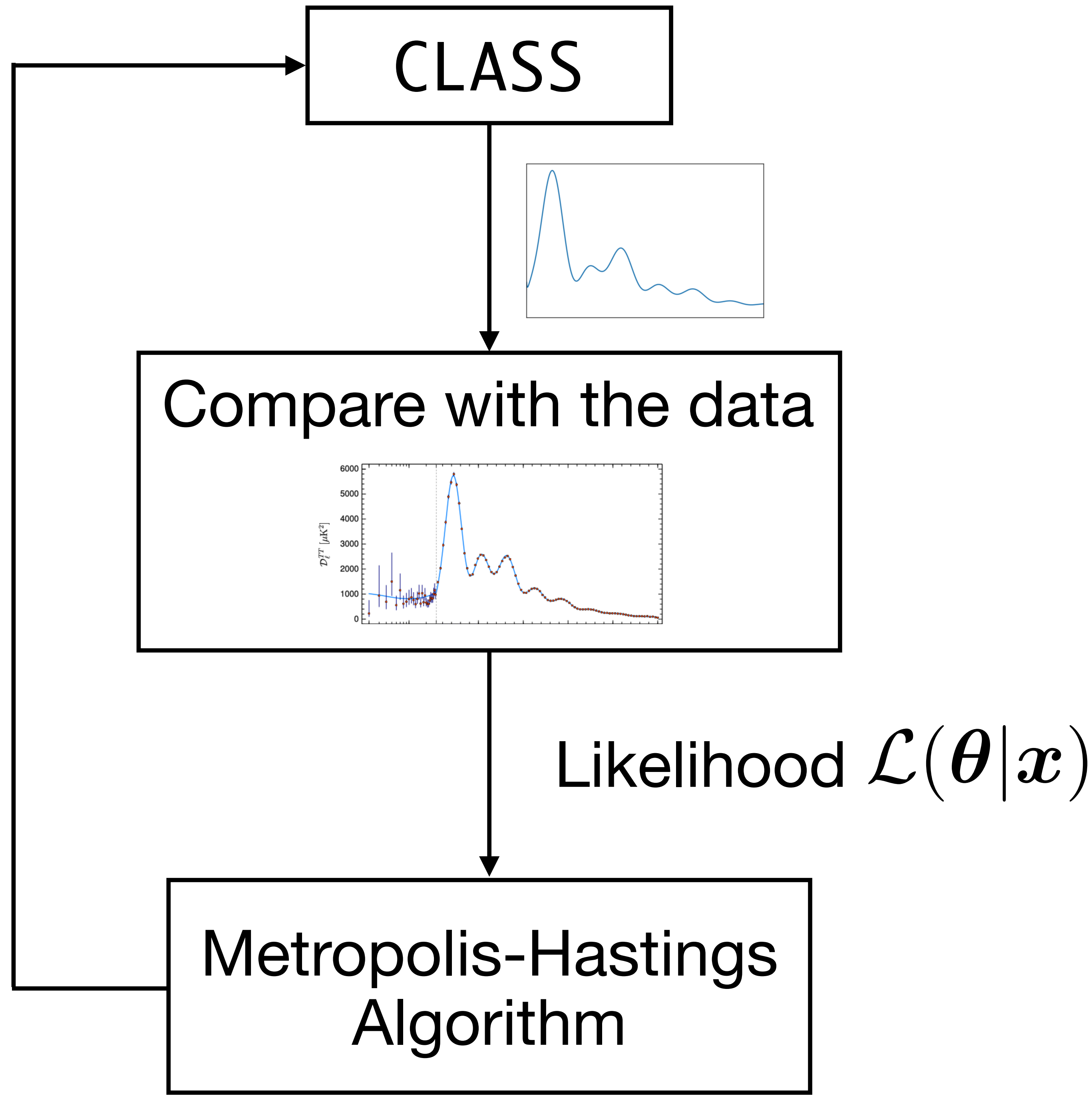
MCMC Chains



Markov Chain Methodology

Input Parameters

$$\left\{\Omega_{\mathrm{M}}, \Omega_{\mathrm{B}}, H_0, A_{\mathrm{s}}, n_{\mathrm{s}}, \tau\right\}$$



Convergence Test

- Operationally, effective convergence of Markov chain simulation has been reached when inferences for quantities of interest **do not** depend on the **starting point** of the simulations.
- We will need to cut the **burn-in** phase - usually the first half of the chains.
- It is advisable to have **many chains** and make a comparison between them.

Gelman-Rubin Convergence Test

- We will need to compute the estimated mean and compare the variance.

- For m number of MC chains, Define between-chain variance as

$$B/n = \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_{j.} - \bar{\theta}_{..}\right)^2$$

where θ_{jt} is the t^{th} of the n iteration of θ in chain j. The variance between chains is

$$W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}\left(\bar{\theta}_{jt} - \bar{\theta}_{j.}\right)^2$$

Gelman-Rubin Convergence Test

- We can calculate the weighted variance $\hat{\sigma}^2$ as

$$\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{B}{n}.$$

- The **Gelman-Rubin diagnostic** $\hat{R}$ is a method to assess the convergence of MCMC chains.

$$\hat{R} = \frac{m+1}{m} \frac{\hat{\sigma}^2}{W} - \frac{n-1}{mn}.$$

- The standard convergence is when

$$\hat{R} - 1 < 0.01$$

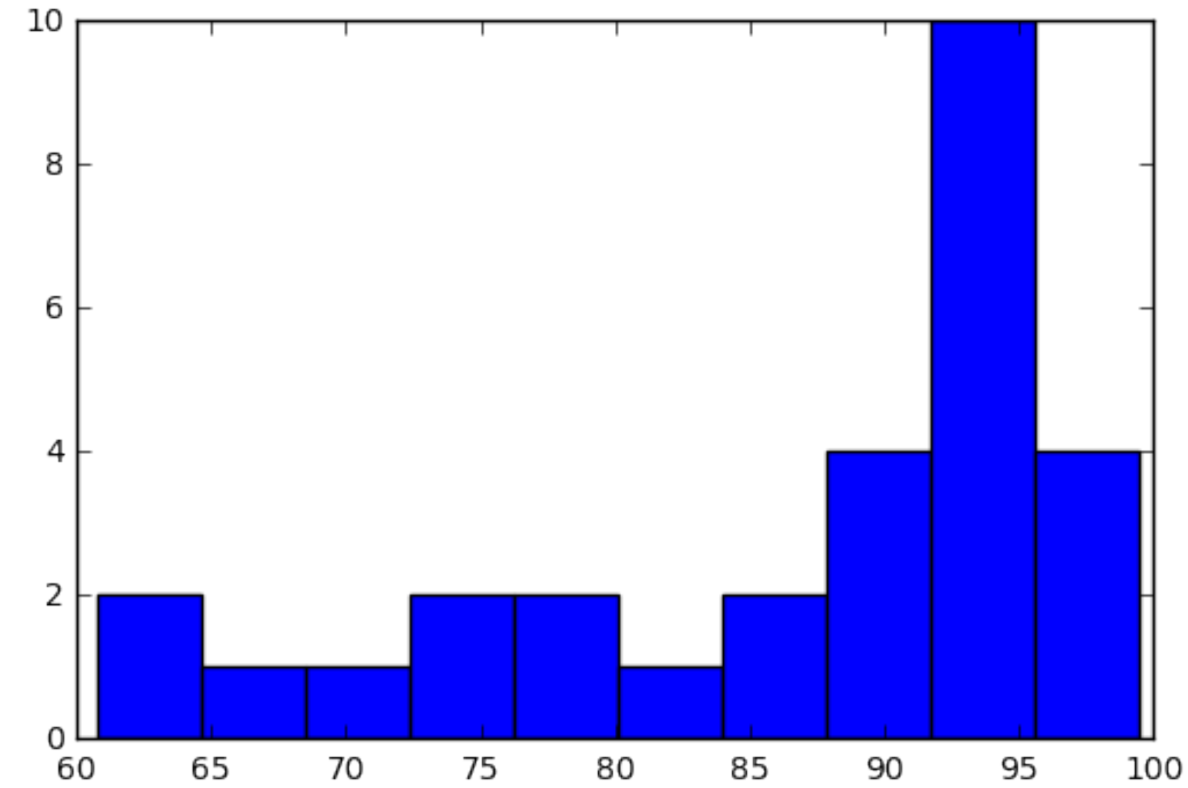
Histogram

- Histogram is a common way to make sense of **discrete data**

Data

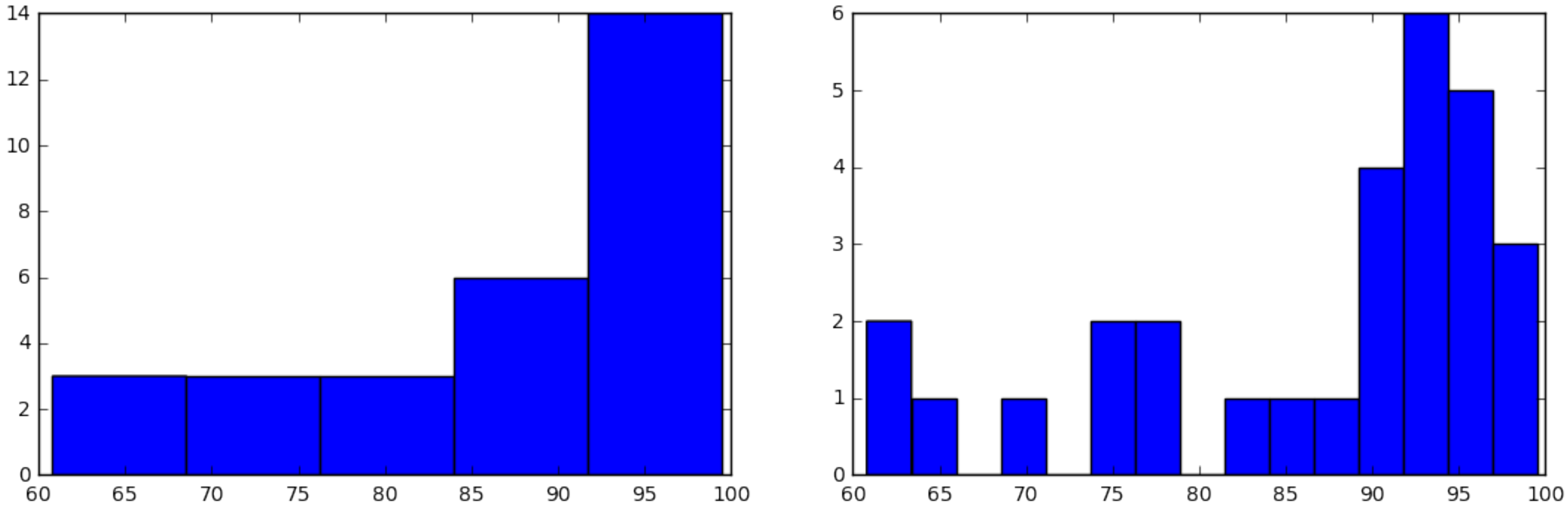
93.5, 93, 60.8, 94.5,
82, 87.5, 91.5, 99.5,
86, 93.5, 92.5, 78,
76, 69, 94.5, 89.5,
92.8, 78, 65.5, 98,
98.5, 92.3, 95.5, 76,
91, 95, 61.4, 96, 90

Histogram



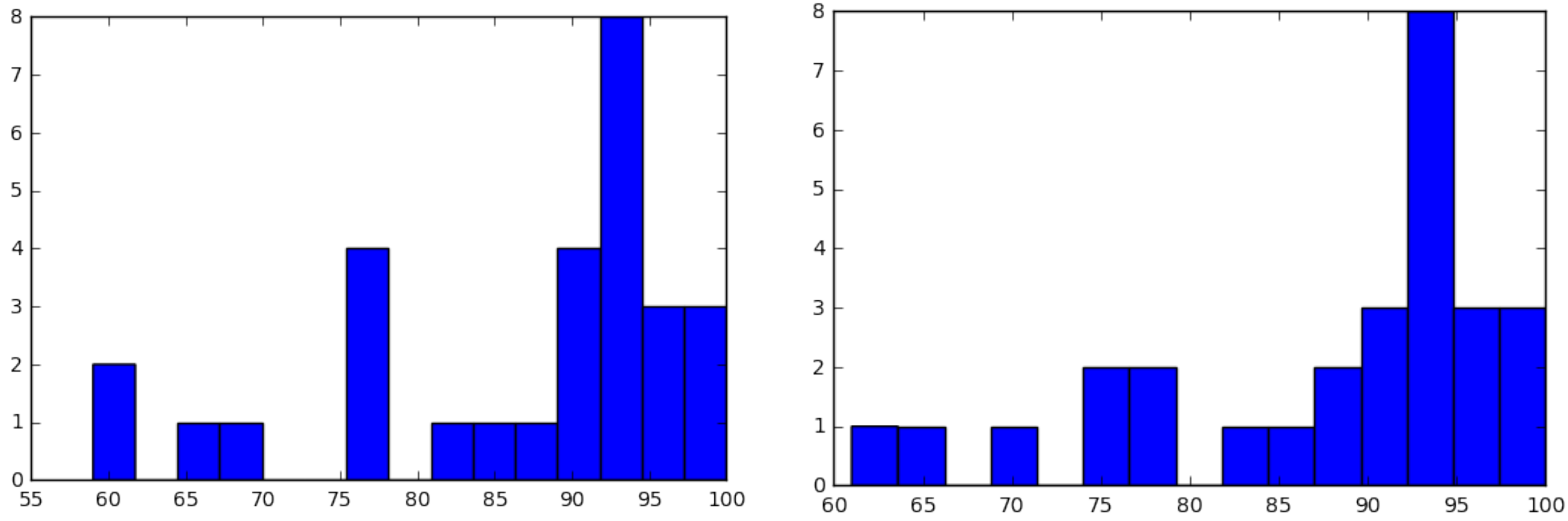
Histogram

- The same data could generate **different histograms** depending on the **number of bins** used.



Histogram

- The same data can generate different histograms depending on the starting point of the **left edge of the bins.**



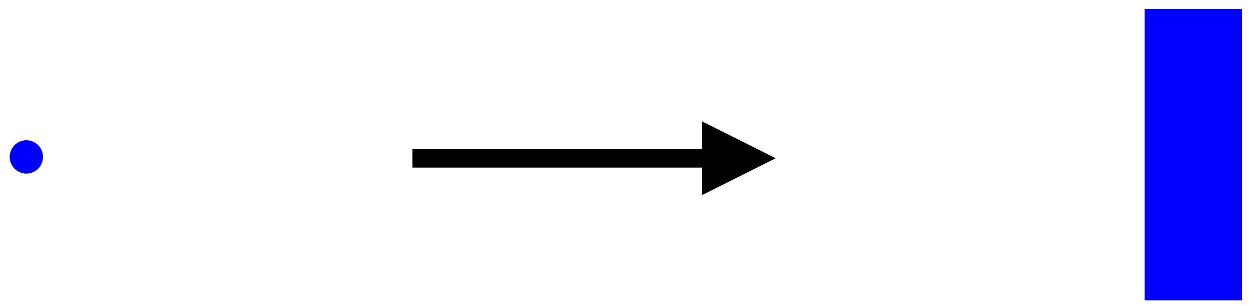
Histogram

Drawbacks of Histogram

- Not smooth
- Dependence on width of the bins
- Dependence on the end points of bins

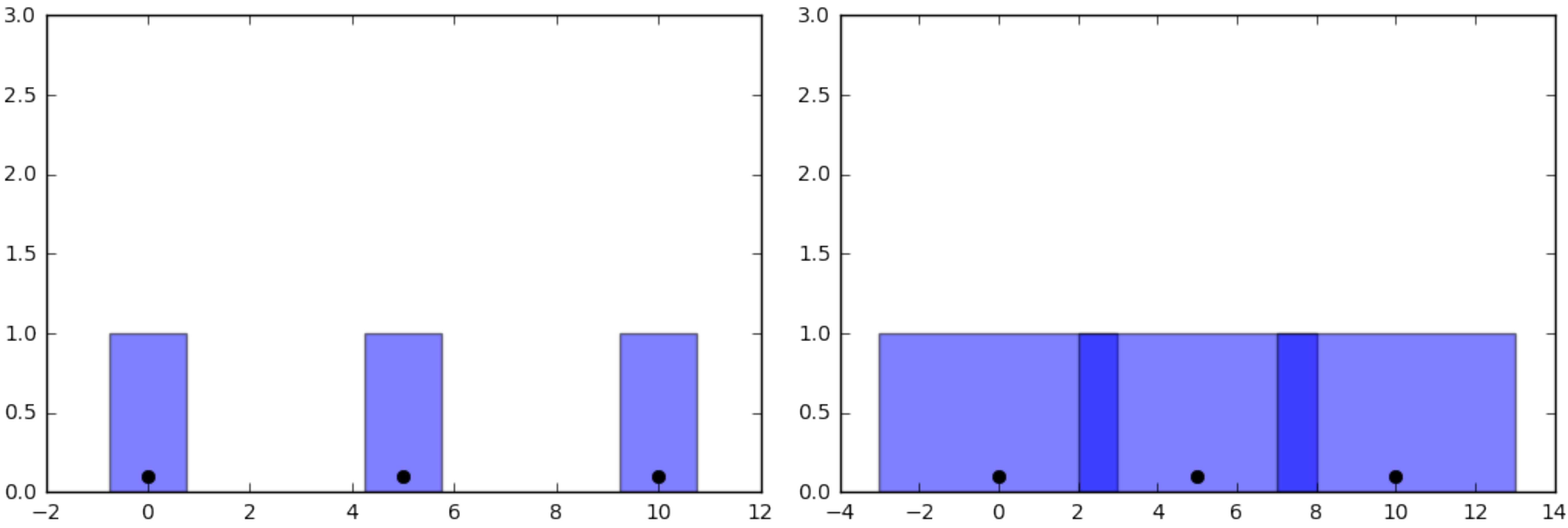
Kernel Density Function

- If we instead replace the data point by a **kernel function**



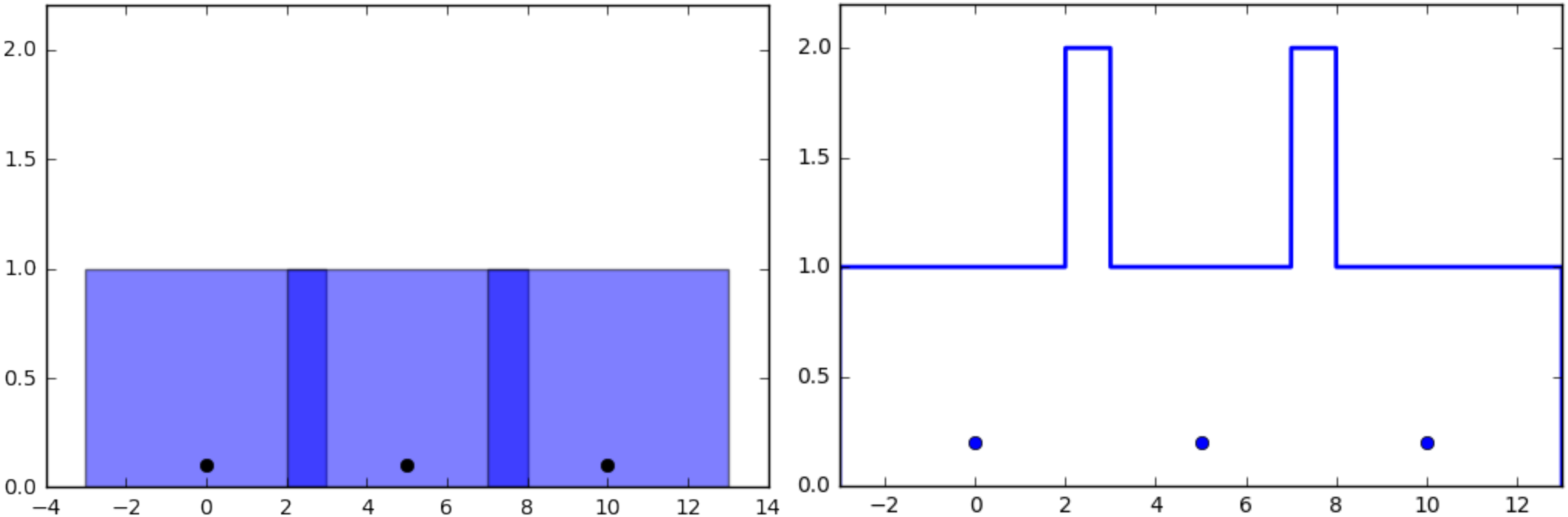
Kernel Density Function

- For simplicity suppose we have only three data points **0, 5, 10**



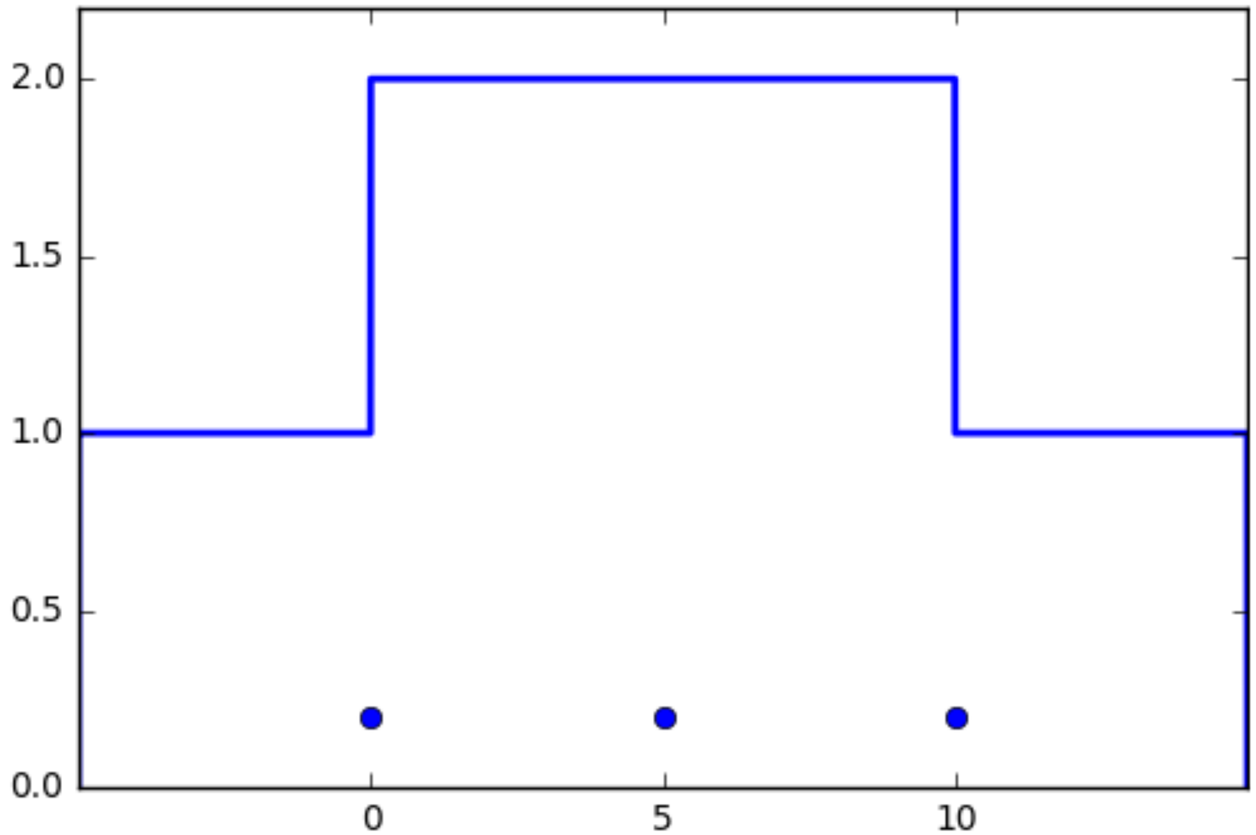
Kernel Density Function

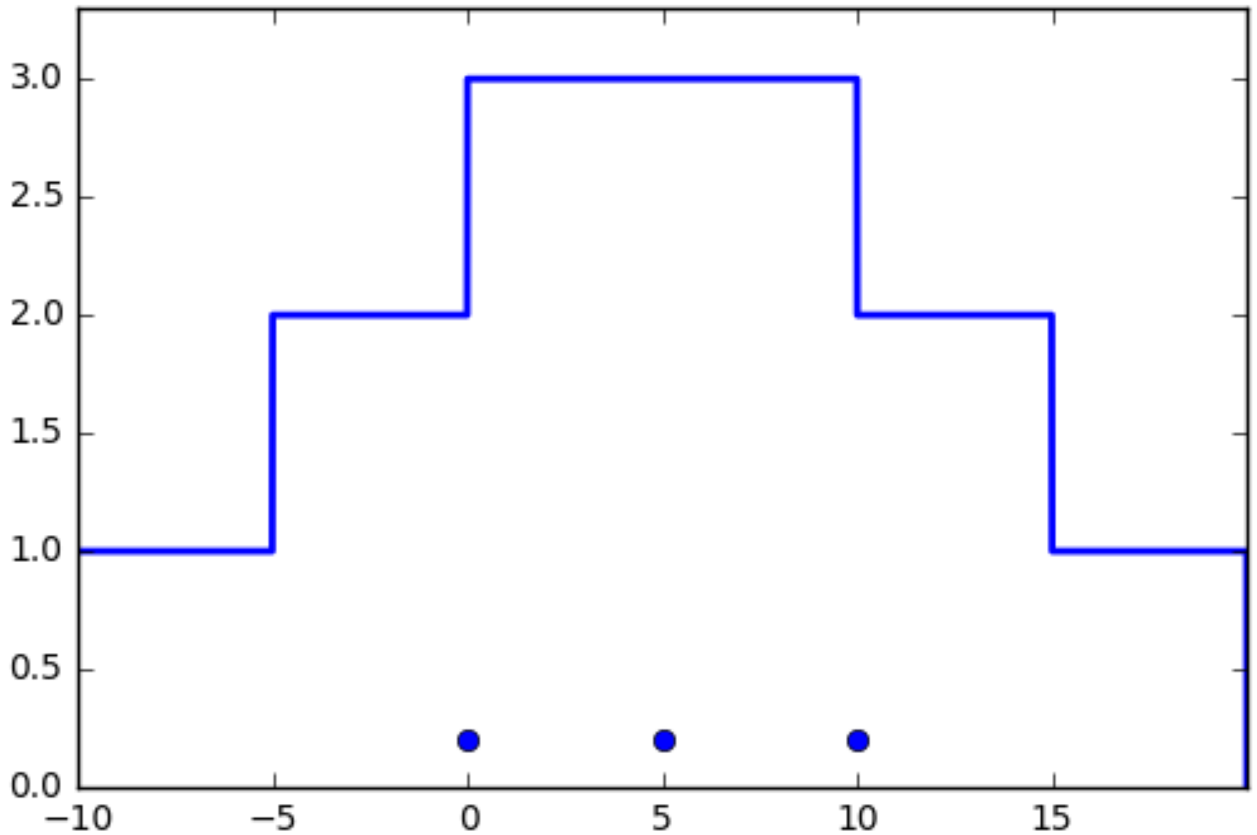
- For simplicity suppose we have only three data points
  0, 5, 10



Kernel Density Function

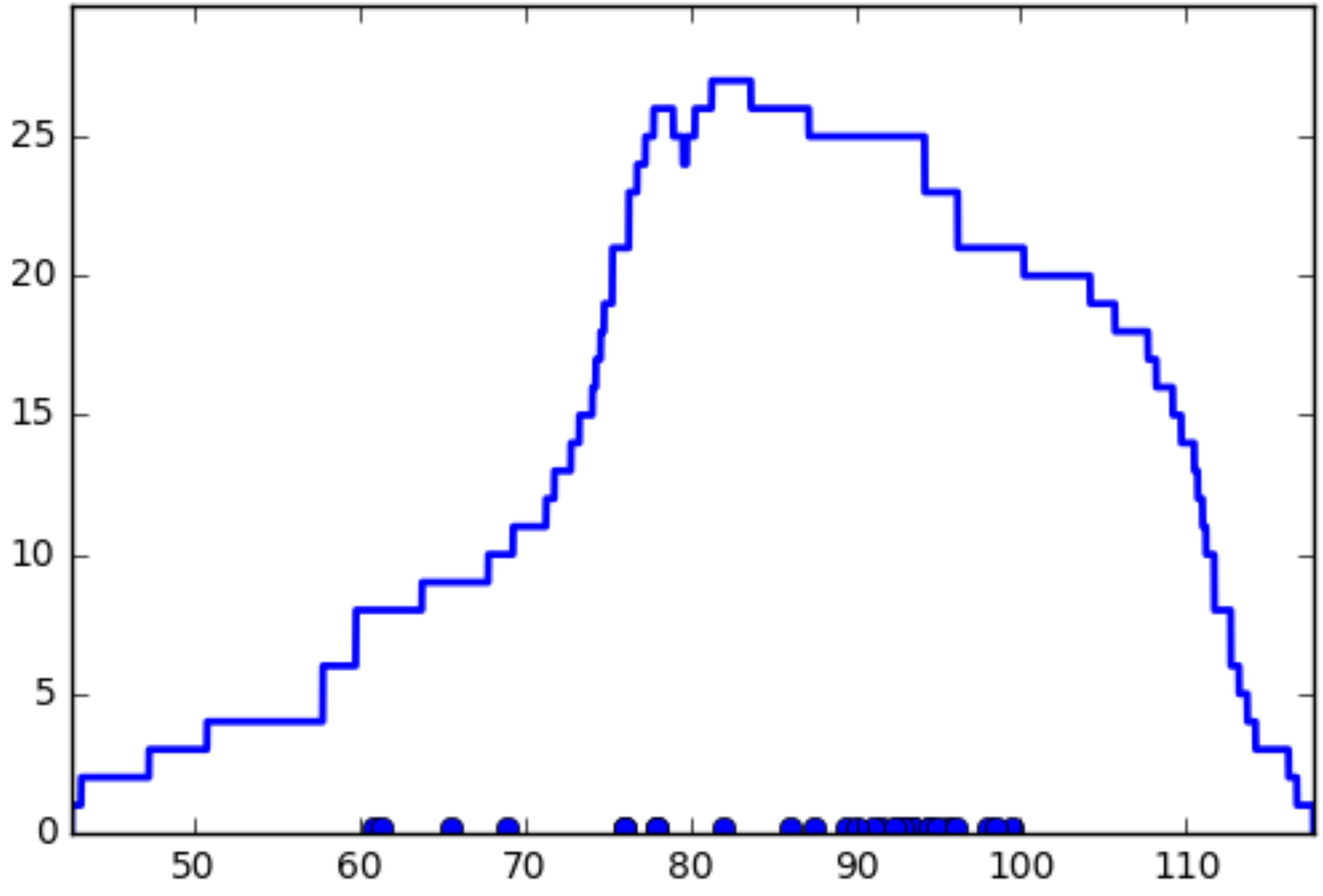
- For simplicity suppose we have only three data points
  0, 5, 10





Kernel Density Function

- With our previous data



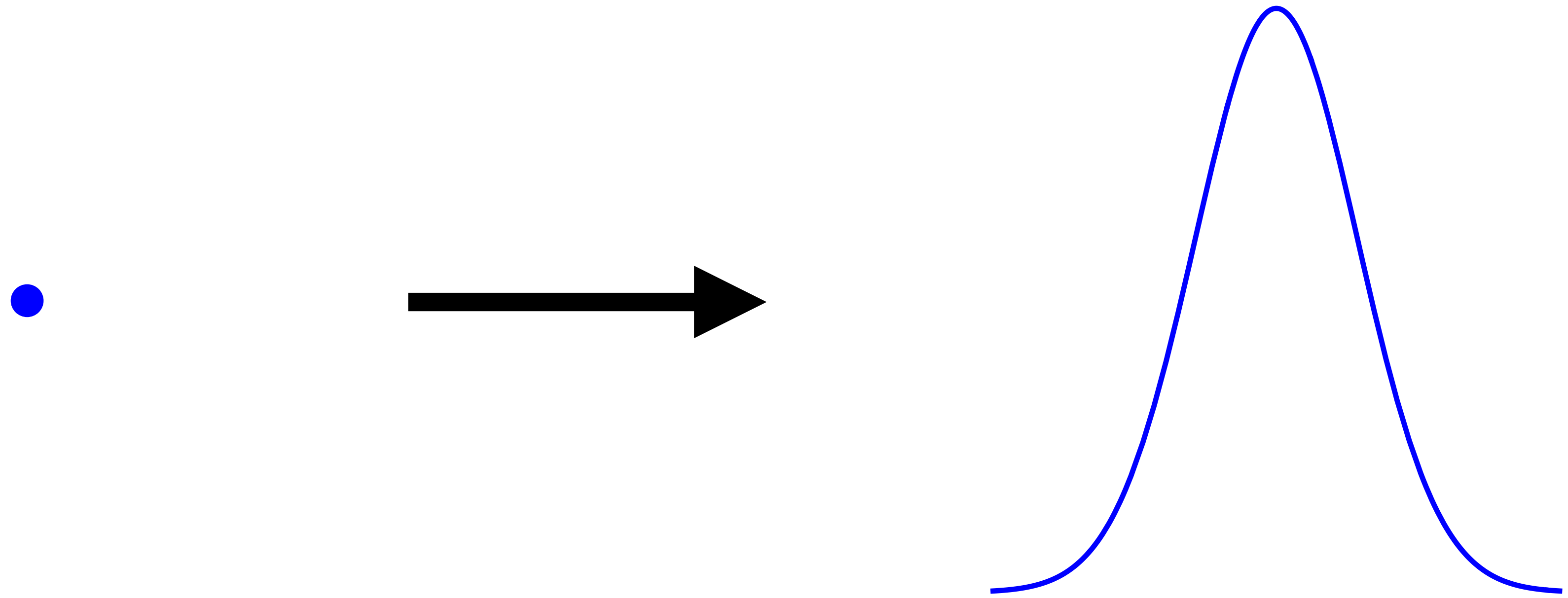
Kernel Density Function

- By using the kernel density function, our histogram will no longer depends on the width of the bins and the end points of the bins

- However, the distribution is still **not smooth**.

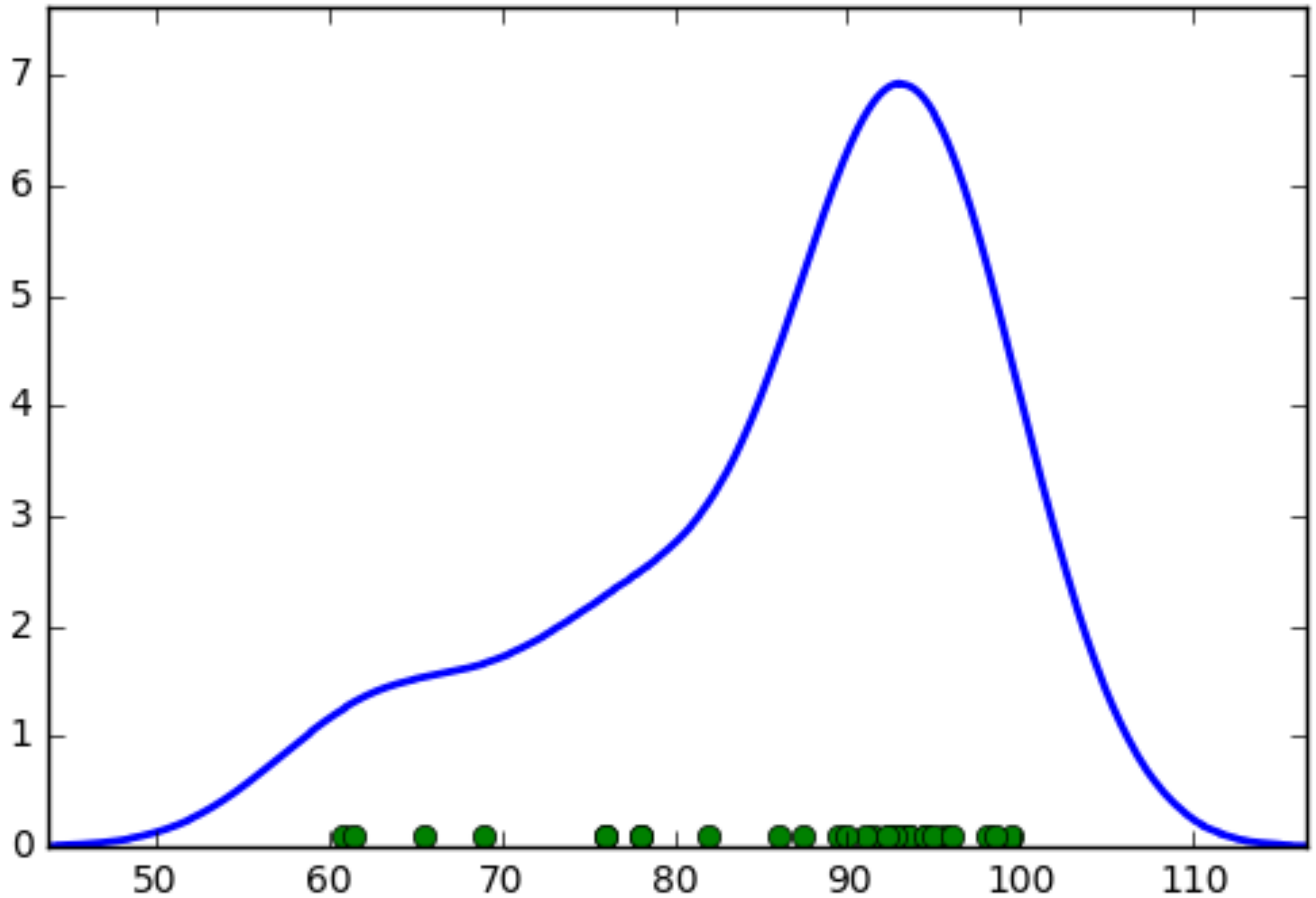
Kernel Density Function

- The plot is not smooth because we use a non-smooth kernel function.
- We can use a smooth kernel function; for example, the Gaussian function.



Kernel Density Function

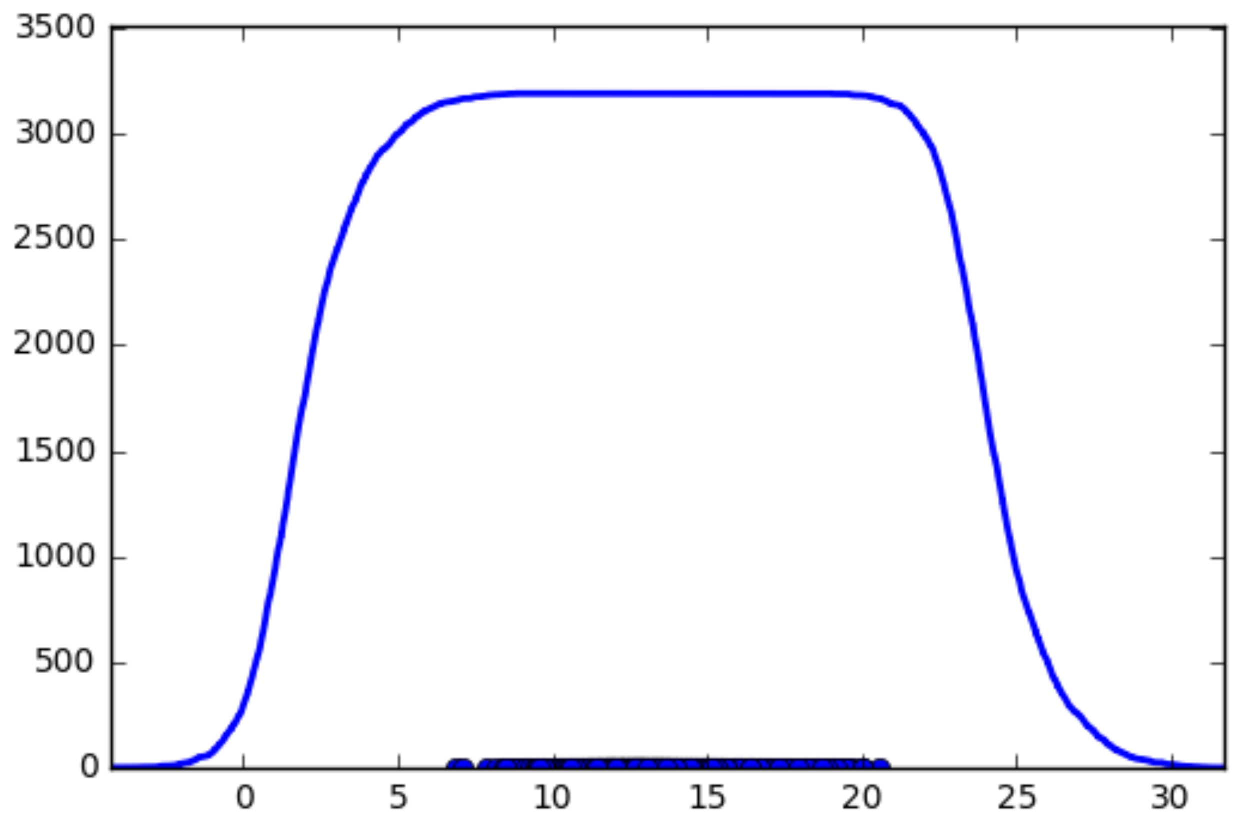
- We have a smooth distribution.



Kernel Density Function

- We oversmooth the distribution - the feature will be washed out.

Oversmoothed

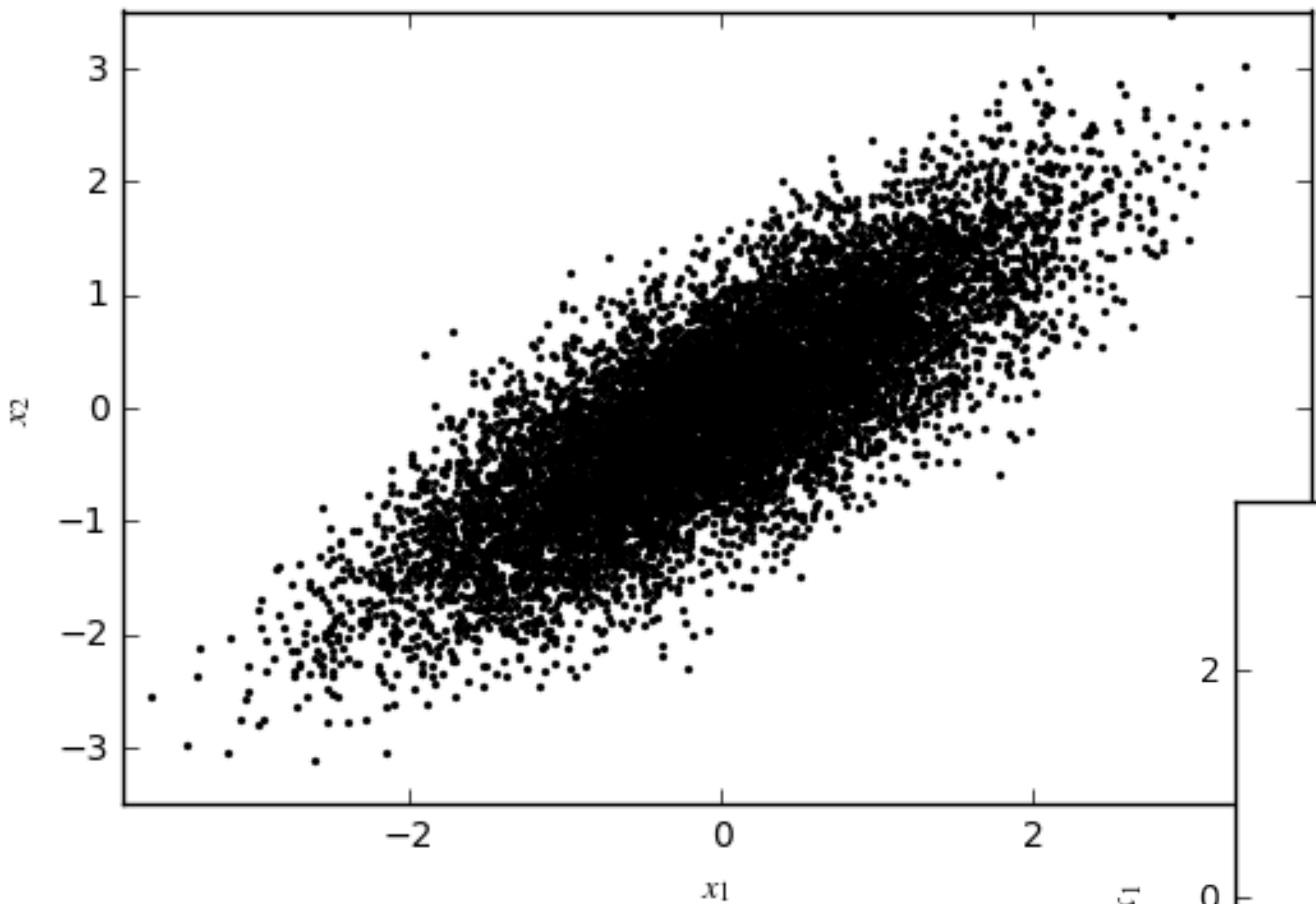


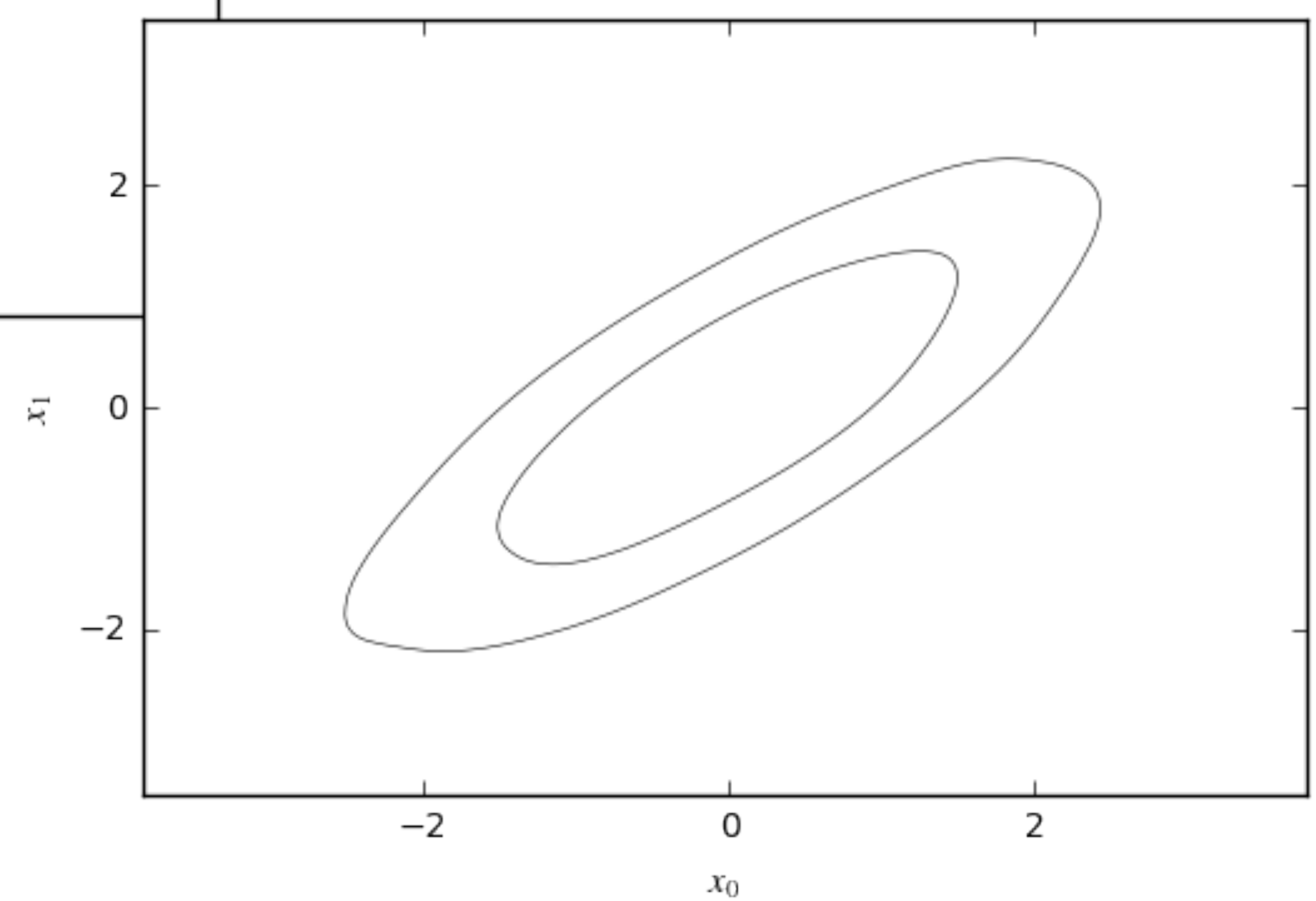
Kernel Density Estimator

- The optimal bandwidth has to be estimated
- A standard way to estimated the optimal bandwidth is to use **Sheather-Jones estimator.**

$$h = 1.06\hat{\sigma}_X N^{-1/5}$$

Kernel Density Estimator





Kernel Density Estimator