USTH MASTER SPACE

Earth Observation - Astrophysics - Satellite Technologies

USTH

VIETNAM FRANCE UNIVERSITY

Lecture 4: Principal components analysis (PCA) and Bayesian methods

Hoàng Đức Thường

Department of Space and Applications (DSA), USTH

Hanoi, 2021

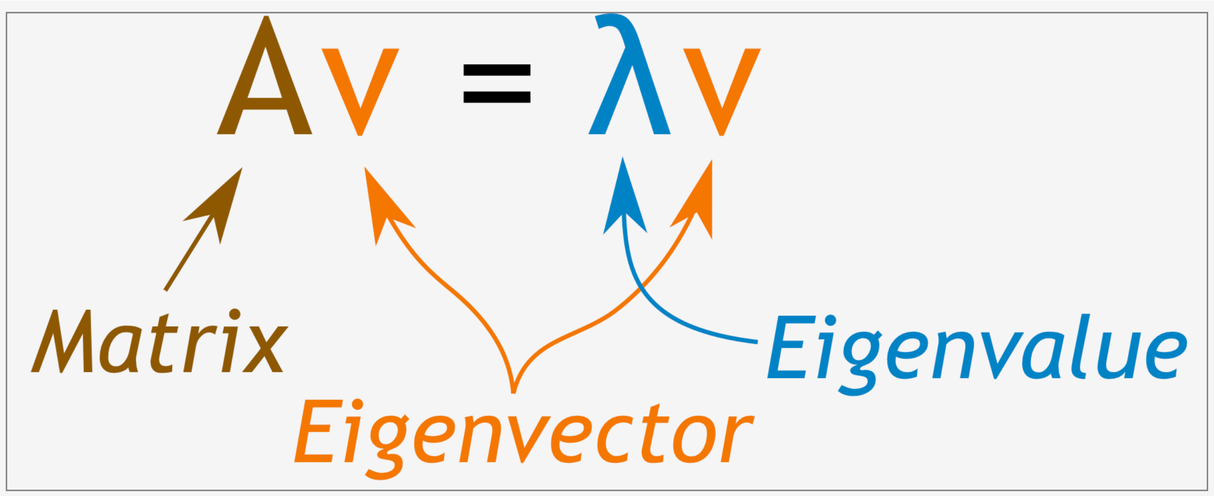
USTH MASTER SPACE
Earth Observation - Astrophysics - Satellite Technologies

USTH
VIETNAM FRANCE UNIVERSITY

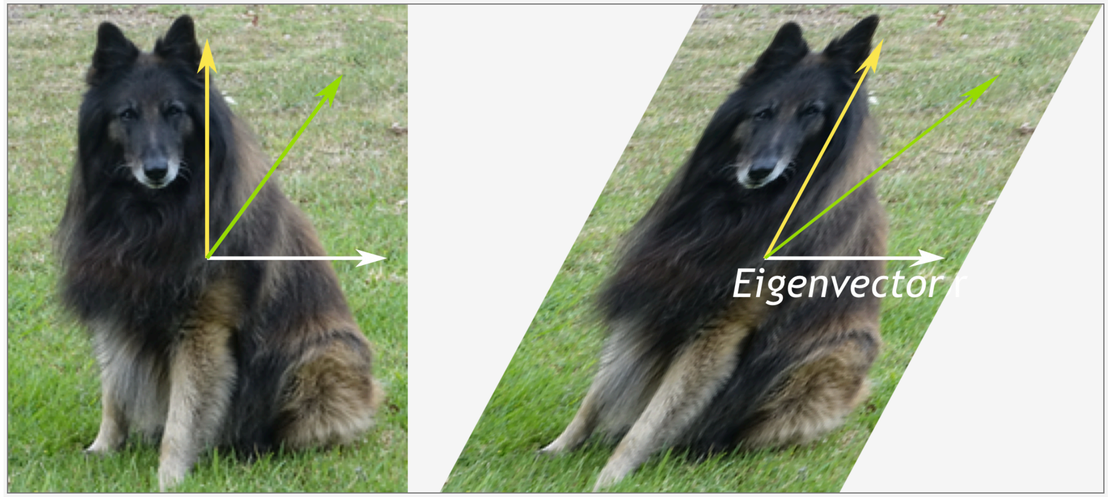
Principal components analysis (PCA) and Bayesian methods

In this class I will focus on PCA dimensionality reduction to model variables and Bayesian statistic applications.

Matrix: Eigenvector, eigenvalue



An eigenvector **does not change direction** in a transformation, the eigenvalue is the scale of the stretch.



To solve the eigenvector problem for $n \times n$square matrix $\mathbf{A}$:

- Compute the determinant of $\mathbf{A} - \lambda\mathbf{I}$. With λ subtracted along the diagonal.
- Find the roots by solving: $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ to find eigenvalues: λ.
- For each eigenvalue λ, solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ to find an eigenvector $\mathbf{x}$.

source: https://www.mathsisfun.com/algebra/eigenvalue.html

Example 4 $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ is already singular (zero determinant). Find its λ's and $\boldsymbol{x}$'s.

When A is singular, $\lambda = 0$ is one of the eigenvalues. The equation $A\boldsymbol{x} = 0\boldsymbol{x}$ has solutions. They are the eigenvectors for $\lambda = 0$. But $\det(A - \lambda I) = 0$ is the way to find *all* λ's and $\boldsymbol{x}$'s. Always subtract λI from A:

Subtract λ from the diagonal to find $$A - \lambda I = \begin{bmatrix} 1-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix}. \tag{4}$$

Take the determinant "$ad - bc$" of this 2 by 2 matrix. From $1 - \lambda$ times $4 - \lambda$, the "ad" part is $\lambda^2 - 5\lambda + 4$. The "bc" part, not containing λ, is 2 times 2.

$$\det \begin{bmatrix} 1-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix} = (1-\lambda)(4-\lambda) - (2)(2) = \lambda^2 - 5\lambda. \tag{5}$$

Set this determinant $\lambda^2 - 5\lambda$ to zero. One solution is $\lambda = 0$ (as expected, since A is singular). Factoring into λ times $\lambda - 5$, the other root is $\lambda = 5$:

$\det(A - \lambda I) = \lambda^2 - 5\lambda = 0$ yields the eigenvalues $\lambda_1 = 0$ and $\lambda_2 = 5$.

Now find the eigenvectors. Solve $(A - \lambda I)\boldsymbol{x} = \boldsymbol{0}$ separately for $\lambda_1 = 0$ and $\lambda_2 = 5$:

$$(A - 0I)\boldsymbol{x} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ yields an eigenvector } \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \text{ for } \lambda_1 = 0$$

$$(A - 5I)\boldsymbol{x} = \begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ yields an eigenvector } \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ for } \lambda_2 = 5.$$

Source: http://math.mit.edu/~gs/linearalgebra/ila0601.pdf

Covariance Matrix

- Variance and standard deviation are operated in one dimension dataset. **Covariance *always* measures between *2 dimensions*** how much the dimensions vary from the mean with respect to each other.
- If we calculate the covariance in one dimension, it is variance.
- $Cov(x, y) = Cov(y, x)$

$$Var(x) = \sigma^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)}{N-1}$$

$$Cov(x, y) = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{N-1}$$

Covariance Matrix

- If we have 3 dimensional dataset (x,y,z) we can measure covariance between $cov(x, y)$, $cov(x, z)$, and $cov(y, z)$. The measurement between $cov(x, x)$, $cov(y, y)$, and $cov(z, z)$ are variance.

- The way to estimate all values of covariance is to calculate all dimension and put in a square matrix.

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

1. Principal components analysis (PCA)

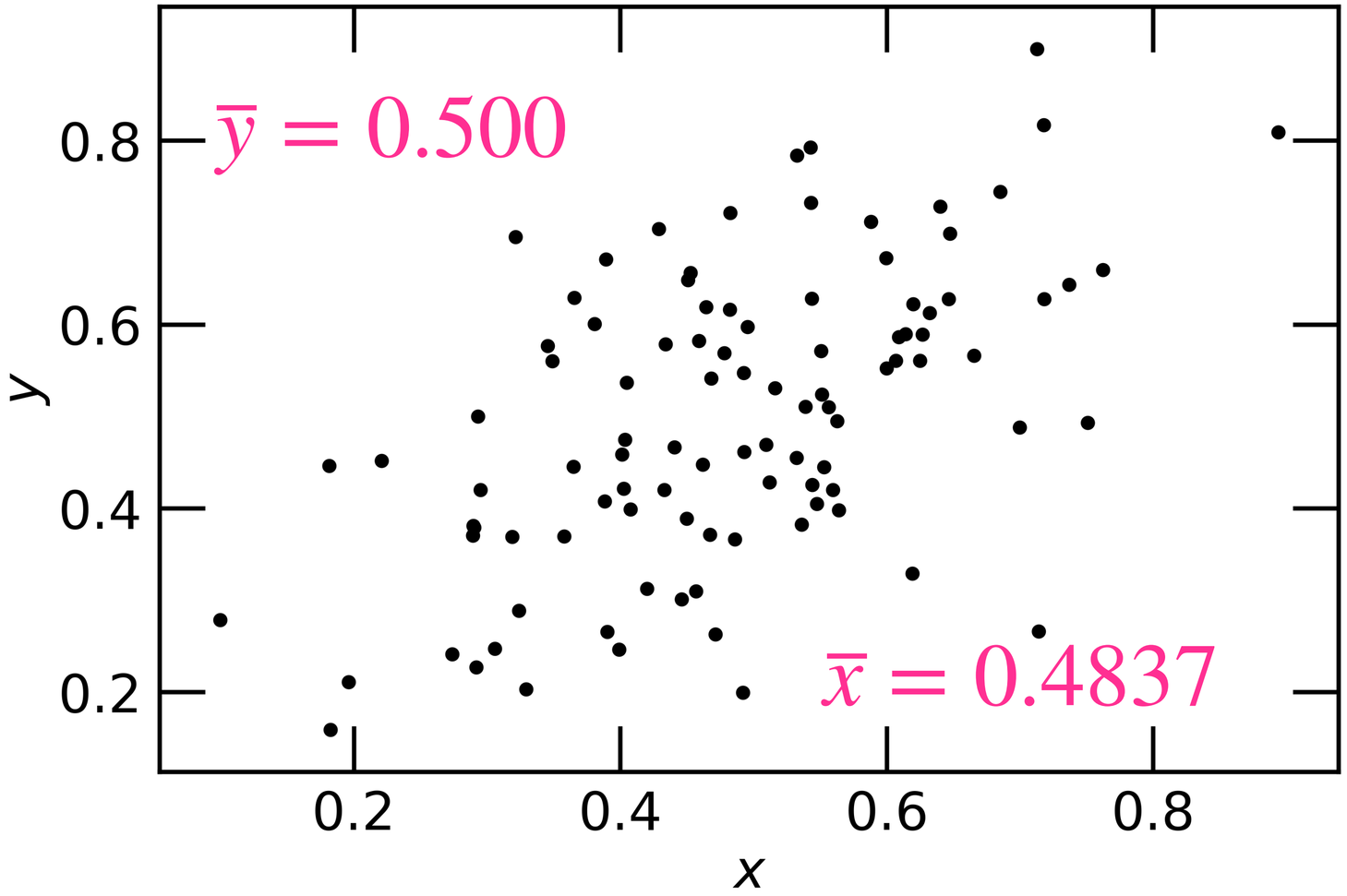
- **Principal component analysis** (PCA) is a statistic procedure that uses the correlations between the variables to identify *which combinations of variables capture most information about the dataset. In a simple words, PCA reducing the number of dimensions.*

- **Geometrically**, it identifies the **directions** in which the cloud of variables is most elongated.

- **Mathematically**, it determines the ***eigenvectors*** of *the covariance matrix* and sorts them in importance according to their corresponding ***eigenvalues.***

1. Principal components analysis (PCA)

- We use to have datasets with many variables for each object. Example: (magnitude, size, types of stars), (temperature, humidity, pressure of an area).



To be simple, we are taking example of 2 dimensional variables x, y.

In fact, the cloud data can be multidimensional variables.

1. Principal components analysis (PCA)

PCA procedure for our dataset (x,y) is following steps:

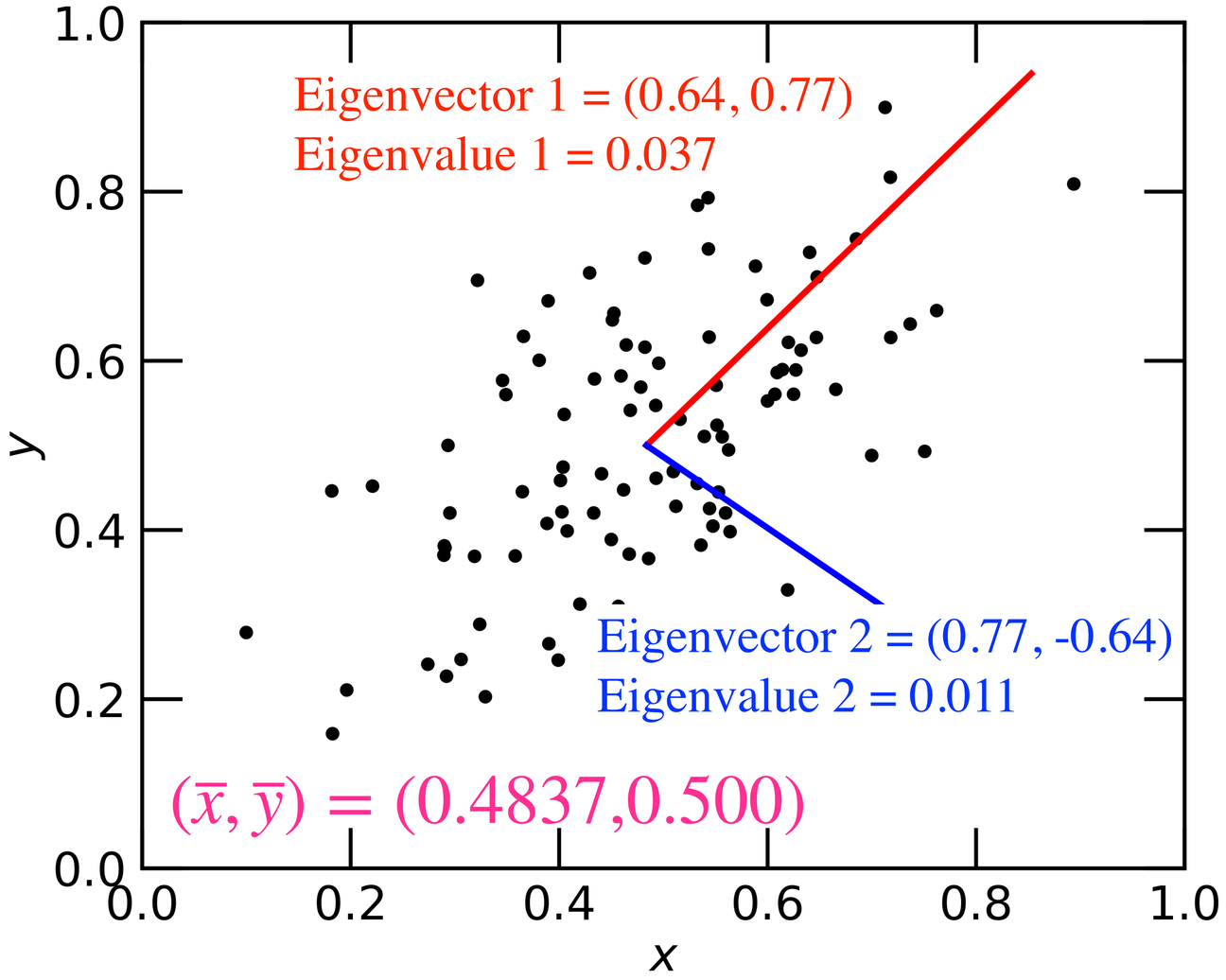
- **Step:** Subtract data to its mean
- **Step:** Calculate the **covariance matrix** of (x,y): $C = \begin{pmatrix} 0.021 & 0.013 \\ 0.013 & 0.026 \end{pmatrix}$

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{pmatrix} \qquad Cov(x,y) = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

- **Step:** Determine the **eigenvalues** and **eigenvectors** of the covariance matrix C:
  - *eigenvalues are* λ_1=0.037, λ_2 = 0.011 and
  - *eigenvectors* $\overrightarrow{v_1}$=(0.64, 0.77) *and* $\overrightarrow{v_2}$=(0.77, -0.64), *respectively.*
- **Step**: Express the data points in the **basis of the eigenvectors** – new co-ordinates are (PCA_1, PCA_2): $\overrightarrow{x} = (x,y) = (\bar{x},\bar{y}) + PCA_1\overrightarrow{v_1} + PCA_2\overrightarrow{v_2}$.

1. Principal components analysis (PCA)

- Eigenvectors and dataset in the same plot. The lengths of the vectors is scaled to promotional to square root of eigenvalues.



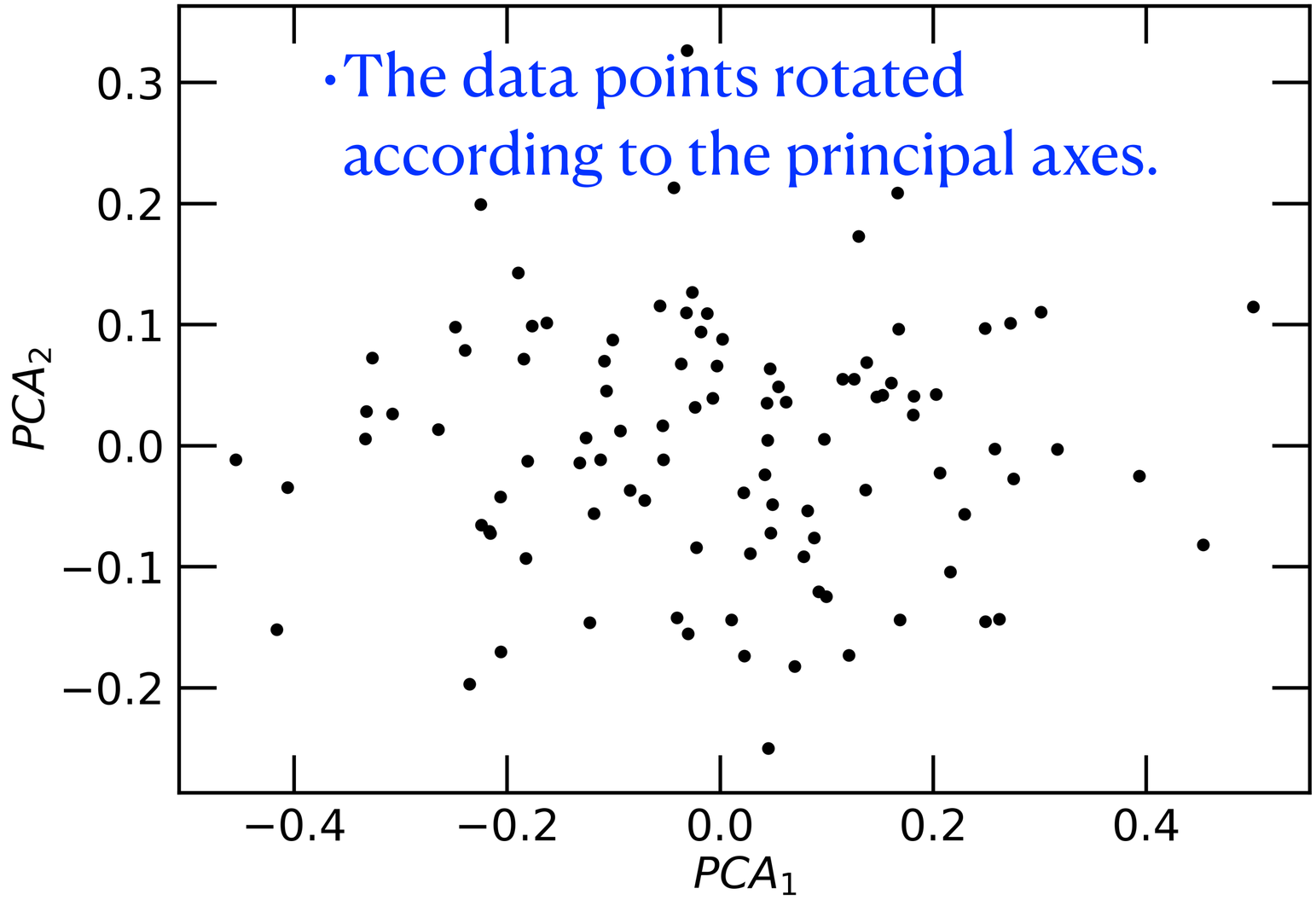
- The eigenvectors indicate the direction of the data points as "principal axes" => PCA axes.

- The size of eigenvalues means the spread (variance) of data along each principal axis.

- In this example: size of eigenvalues = 2. There are 2 axes of data spreading.

1. Principal components analysis (PCA)

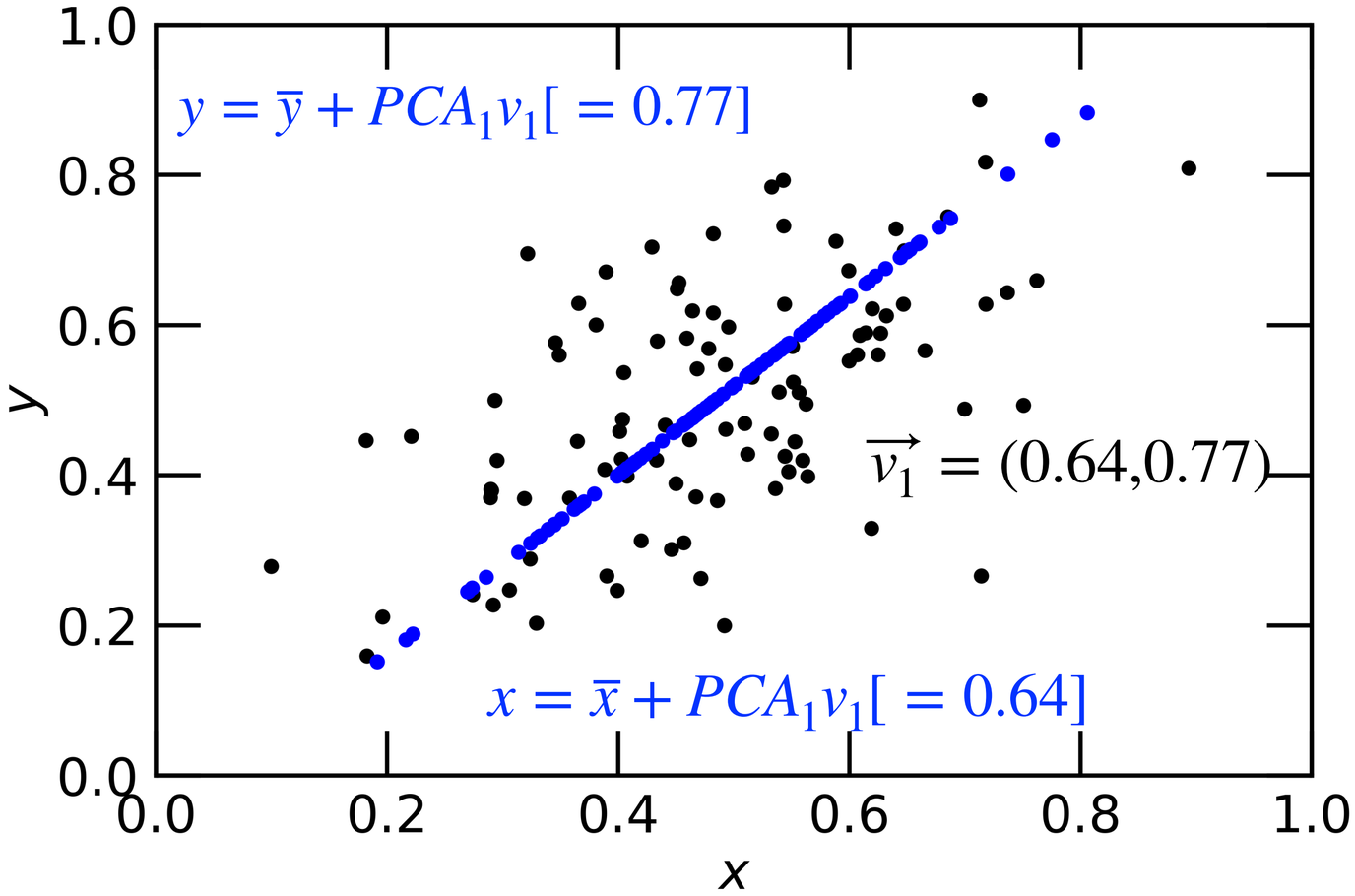
- The principal component values of each data point in the new transformed coordinate, or we can say we projected data point to eigenvectors coordinate.



1. Principal components analysis (PCA)

- The powerful of the PCA method is applied in **reducing dimension** of data. In the other words, approximating a dataset with fewer number of variables.

- Indeed, we can illustrate our dataset by reconstructing a model using 1 principal component. **The best principal component is the eigenvectors that has highest eigenvalue.** $(x, y) = (\bar{x}, \bar{y}) + PCA_1\overrightarrow{v_1}$.



- The blue data points are approximation of the original black data point.

- We can reduce 2 dimensional data to a model of a vector, or so-call data compression.

- The variance is the eigenvalue λ_1.

2. Bayesian method

- To understand Bayesian statistic, we need to understand the **conditional probability.**
- $P(A \mid B)$ means **the probability of A on the condition that B has occurred**.
- P(phở | Monday) = 1, P(noodle/Tuesday) = 0.
- The important **Baye's theorem**:

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

- We can write in a join probability:

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

2. Bayesian method

- **Example:** If a patient has disease COVID-19, the chance of a certain medical test being positive is 95%. According to a survey, 10% of the population have the disease (COVID-19), and the test records a false positive 5% of the time due to the kit test. *If I receive a positive test, what is your probability of having COVID-19?*

- **First**, we have:

  The probability of a positive test will have COVID-19: $P(+|COVID) = 0.95$,

  The probability of population have the disease: $P(COVID) = 0.1$,

  The probability of false test: $P(+|No\ COVID) = 0.05$.

- **Second**, there are two cases of a **truly positive** $P(+)$,
  - The first case is falling in the 10% population and 95% probability of positive test,

    $P(+) = P(+|COVID) * P(COVID) = 0.95 \times 0.1$
  - The second case is falling in the 90% population and 5% of failure of kit test.

    $P(+) = P(+|No\ COVID) * P(No\ COVID) = 0.90 \times 0.05$

- We want to calculate, if I have a positive test, the probability of have COVID19 truly: $P(COVID|+)$

2. Bayesian method

- **Example:** If a patient has disease COVID-19, the chance of a certain medical test being positive is 95%. According to a survey, 10% of the population have the disease (COVID-19), and the test records a false positive 5% of the time due to the kit test. *If I receive a positive test, what is your probability of having COVID-19?*

- We want to calculate, if I have a positive test, the probability of have COVID19 : $P(COVID|+)$

- Apply Baye's theorem: $P(COVID|+) = \frac{P(+|COVID)P(COVID)}{P(+)}$

$$P(COVID|+) = \frac{0.95 \times 0.1}{0.95 \times 0.1 + 0.90 \times 0.05} = 0.6785$$

- Interpretation: If the test kit is correct 95%, the probability of having COVID-19 after a positive test is 67.86 %. This is because of the fraction of the population have the disease is 10%. If we assume, there is only 1% of population have COVID-19. Then the probability of having COVID-19 after a positive test is small as 16.1%.

$$P(COVID|+) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.99 \times 0.05} = 0.161$$

2. Bayesian statistics

- Baye's theorem applies for science:

Posterior probability of the model

Likelihood function of the data

Prior probability of the model

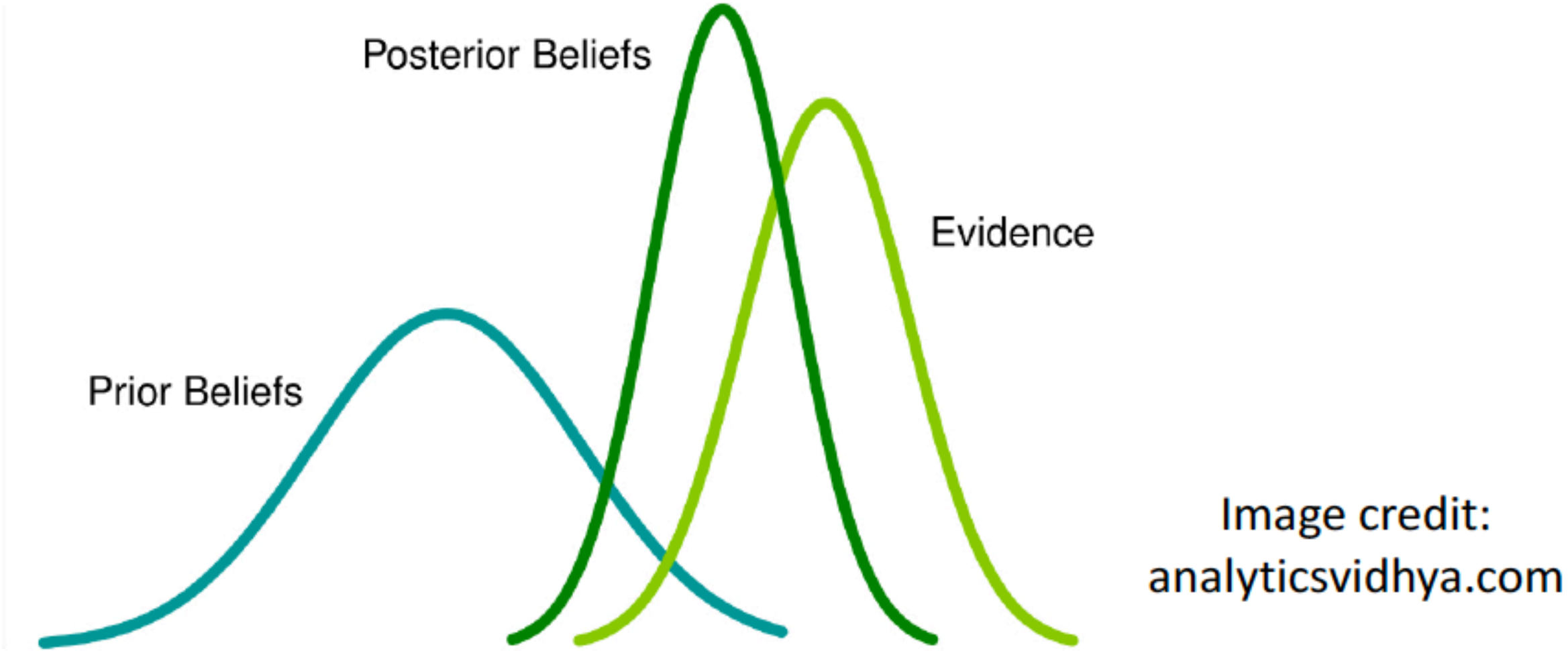
$$\text{Prob}(\text{model}\,|\,\text{data}) = \frac{\text{Prob}(\text{data}\,|\,\text{model})\text{Prob}(\text{model})}{\text{Prob}(\text{data})}$$

Evidence of an event/phenomena that we observed.

$$\text{Prob}(\text{model}\,|\,\text{data}) \propto \mathscr{L}(\text{data}\,|\,\text{model})\text{Prob}(\text{model})$$

2. Bayesian statistics

- Posterior: the probability of A on the condition that B (evidence) has occurred



2. Bayesian statistics

- Posterior probability $p(\theta, X)$ is the *probability* of the parameters θ given the evidence X.
- In contrast, Likelihood function $p(X \mid \theta)$ is the *probability* of the evidence X given the parameters θ.
- Prior probability distribution is an unknown quantity as parameters of a model. Example: A Bayesian can argue that there is a prior probability of 10% of population has COVID-19. This probability should be updated every time we have new data. The prior probability is both the strong and weak point of Bayesian. The prior should be stated. If it is unknown we can just use a uninformative (wide) prior.
- The prior probability is a logical necessity when evaluating the probability of a model.

2. Bayesian statistics: The uniform prior

- In the absence of information, a uniform (or constant) prior (distribution) is often supposed. This is equivalent to a fitting range.

- Let's take the example of fitting a parameter a to an observed data $p(a, data)$. In the previous lecturer, we actually assumed a *uniform prior (constant)* when determining the probability distribution of parameter "a" in χ^2 fitting:

$$P(a \mid data) \propto \mathscr{L}(data \mid a)P(a)$$

Constant, due to uniform distribution

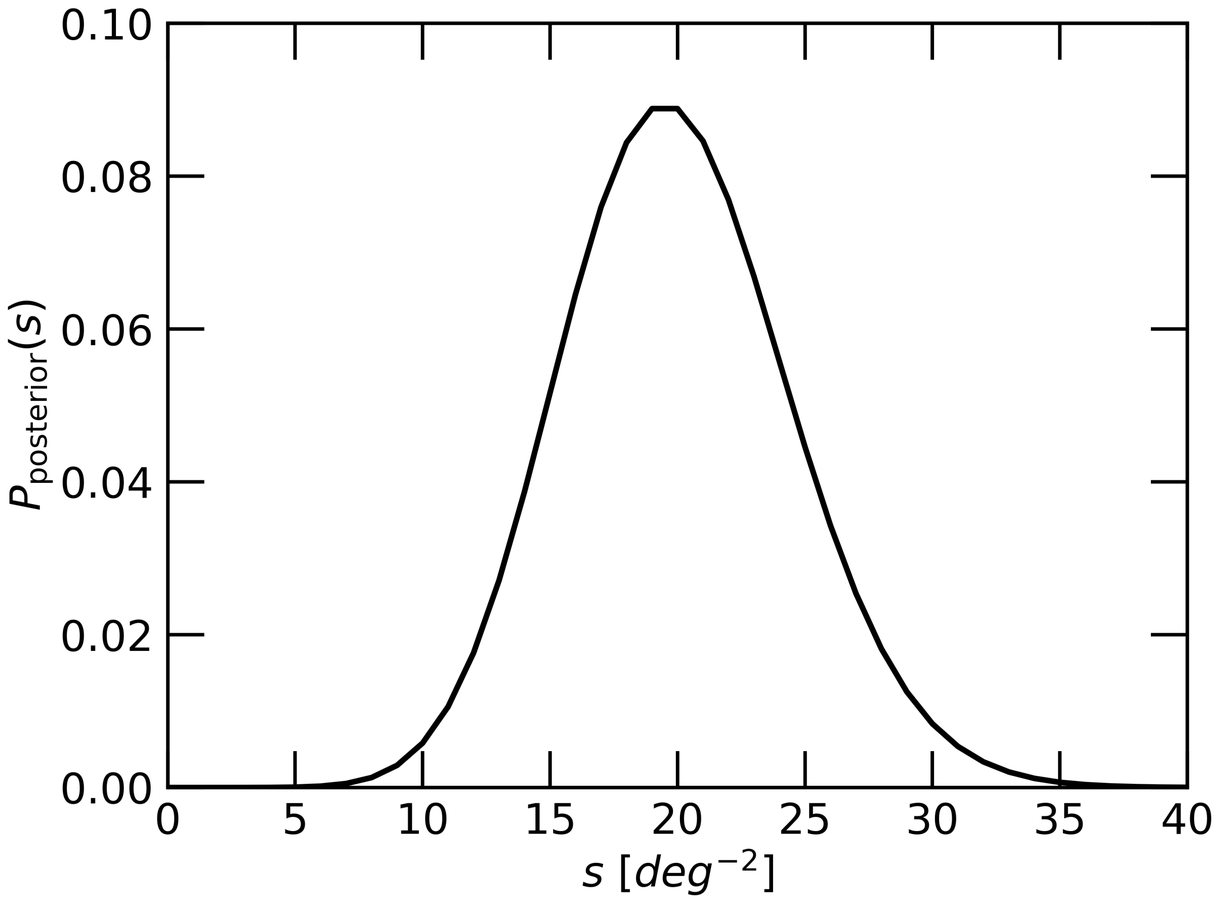
- Because we will normalize the posterior $\int P(a \mid data)da = 1$ for observed data, then we do not need the denominator.

- Assuming Gaussian variables, the likelihood: $\mathscr{L}(data \mid a) \propto exp(-\chi^2/2)$

- Then the posterior $P(a \mid data) \propto exp(-\chi^2/2)$

2. Bayesian statistics: The uniform prior

Example: A 1-degree survey finds 20 quasars (our data D) or the density of quasars is 20 per deg^2. What is the posterior probability distribution for the quasar number density s?



- We are counting random events compared to a mean (=20), so the **Poisson distribution** applies.
- Mean number of quasars in area: A deg^2 = 20 A
- $P(0) = exp(-20A) = 0.01 \ \ if\ A = 0.23deg^2$

- **Using bayesian statistic:**
  - The posterior probability distribution

    $P(s \mid D) \propto P(D \mid s)P(s)$
  - It is equivalent as a Poisson distribution:

    $\text{Poisson}(\text{n} = 20 \mid \text{s})[\text{Uniform prior}]$

2. Bayesian statistics: The uniform prior

Example: A satellite observes 100 galaxies, 30 of which have elliptical shape. What is the posterior probability distribution of the elliptical shape fraction p assuming (a) a uniform prior, (b) p has a Gaussian distribution with mean 0.35 and r.m.s. 0.05?

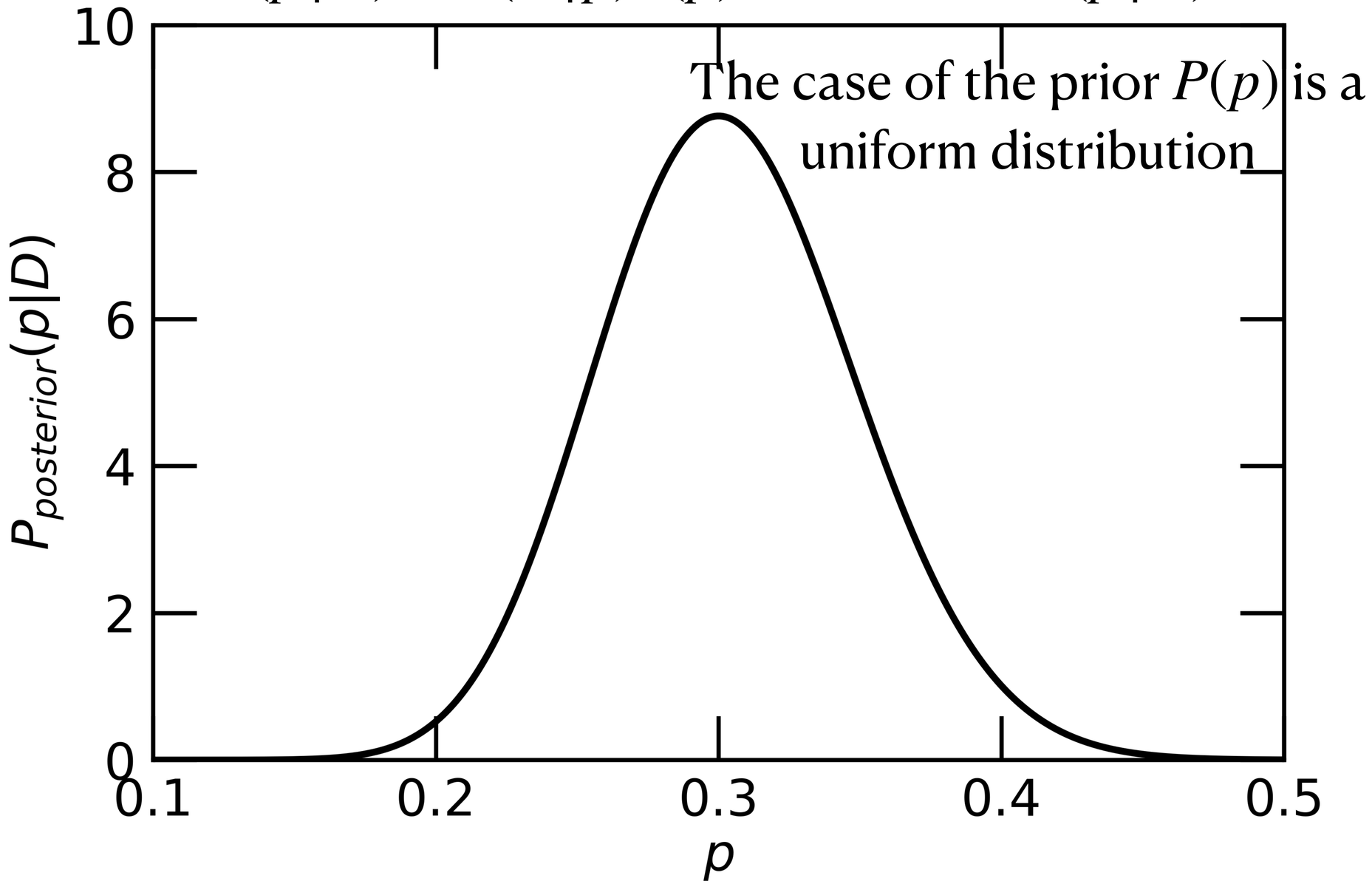
- $P(p \mid D)$ is the probability distribution of p given the data D, the quantity we want to determine.
- $P(D \mid p)$ is the probability of the data for a given value of p (likelihood). As we studied in the probability lecture, the probability distribution is given by the Binomial distribution:

$$P_{Binomial}(n = 30 \mid N = 100, p)$$

- $P(p)$ is the prior in p which we take as a uniform distribution between $p = 0$ and $p = 1$.
- Use Baye's theorem $P(p \mid D) \propto P(D \mid p)P(p)$ to determine $P(p \mid D)$.

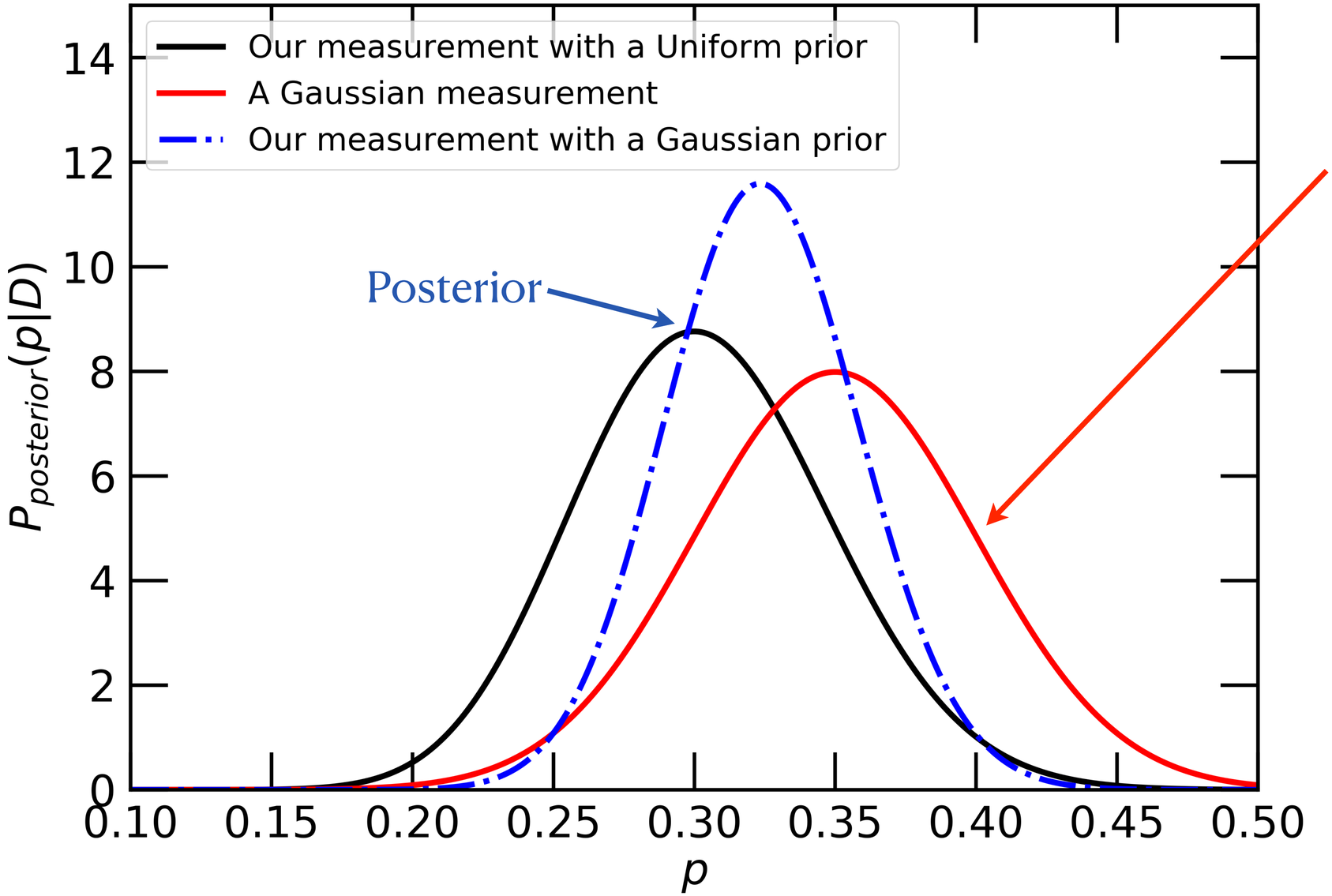
2. Bayesian statistics: The uniform prior

- Use Baye's theorem $P(p \mid D) \propto P(D \mid p)P(p)$ to determine $P(p \mid D)$.



2. Bayesian statistics: The Gaussian prior

- Use Baye's theorem $P(p \mid D) \propto P(D \mid p)P(p)$ to determine $P(p \mid D)$.



P(p) has Gaussian distribution with mean 0.35 and r.m.s. 0.05

Bayesian statistics allows us to combine previous measurement using the prior.

2. Bayesian statistics: Marginalization

- Now if we have have already determined the 2D posterior probability distribution of a model of 2-parameter, $P_{2D}(a, b)$.

- What is the probability distribution for parameter a, in case of all possible values of parameter b? [This is called as **marginalization** of parameter b].

1. For Gaussian variables: $Likelihood \propto exp(-\chi^2/2)$. (recall chi-square distribution)

2. Convert the 2D χ^2 grid into a 2D probability grid: $P_{2D}(a, b) \propto exp(-\chi^2/2)$.

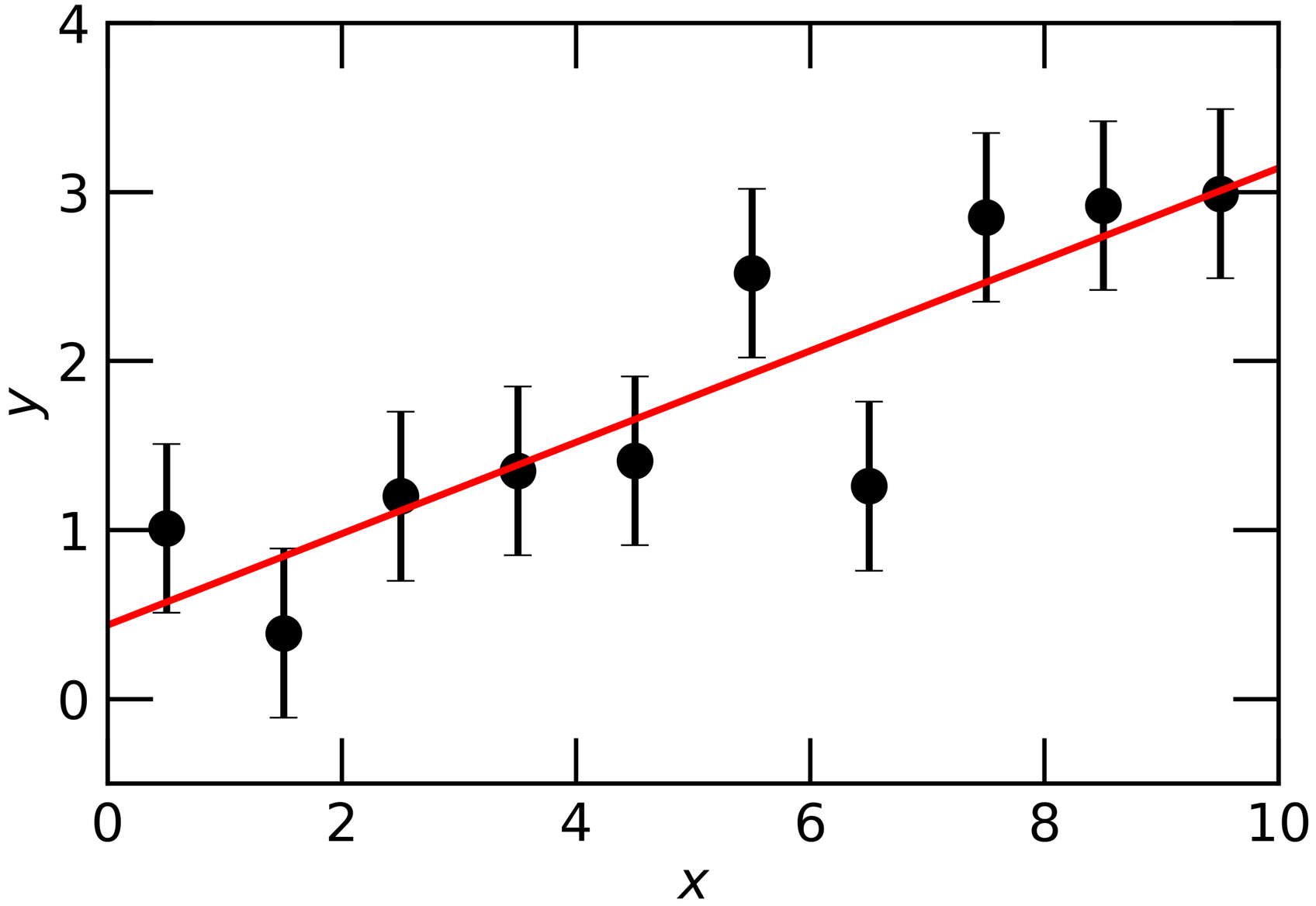
3. Normalize the grid: $\sum_b P_{2D}(a, b) = 1$

- **Marginalization** can be performed by integrating over one axis of the probability distribution:

$$P_{1D}(a) = \sum_b P_{2D}(a, b)$$

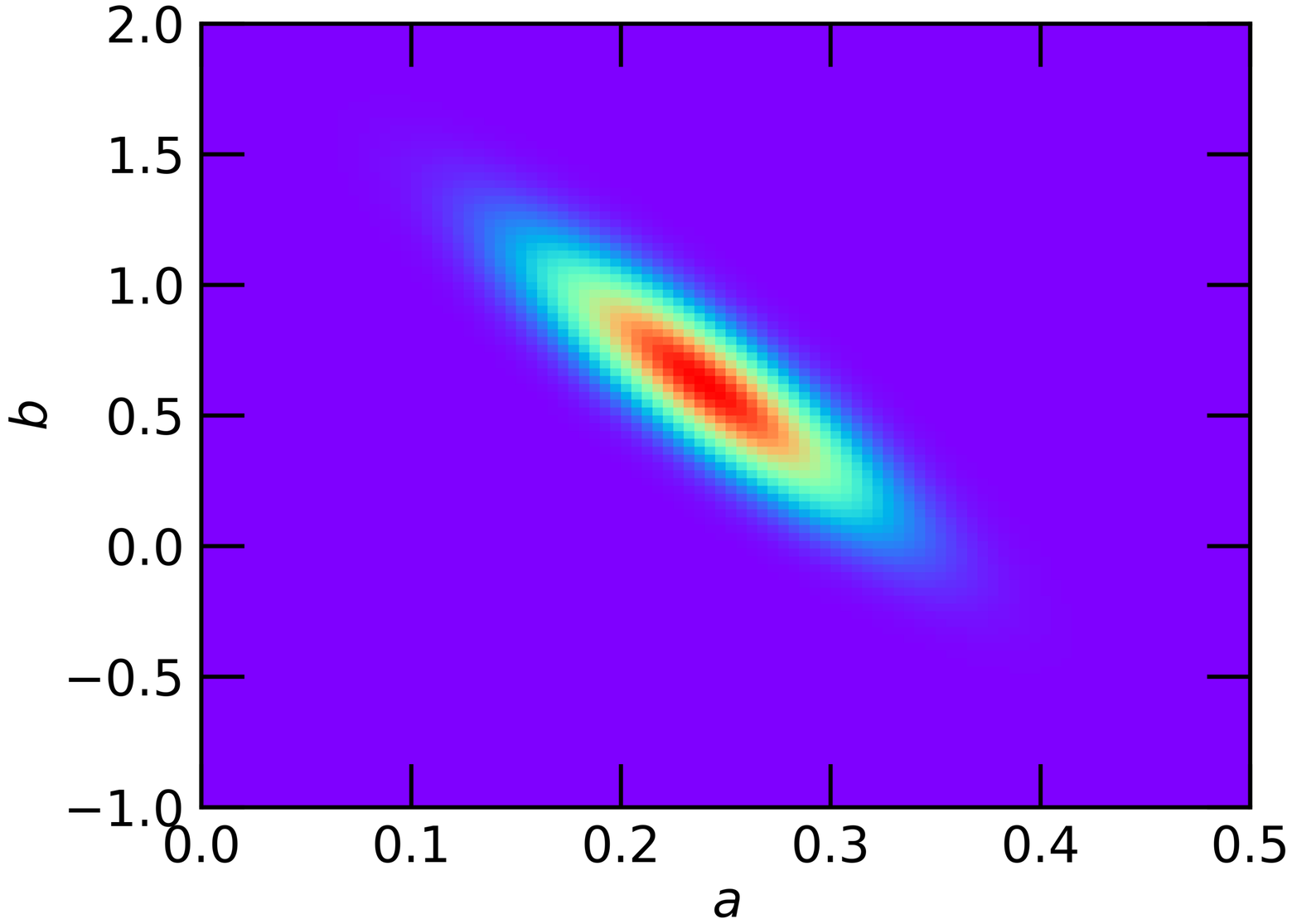
2. Bayesian statistics: Marginalization

- Lets apply the method for the straight model y = ax+b



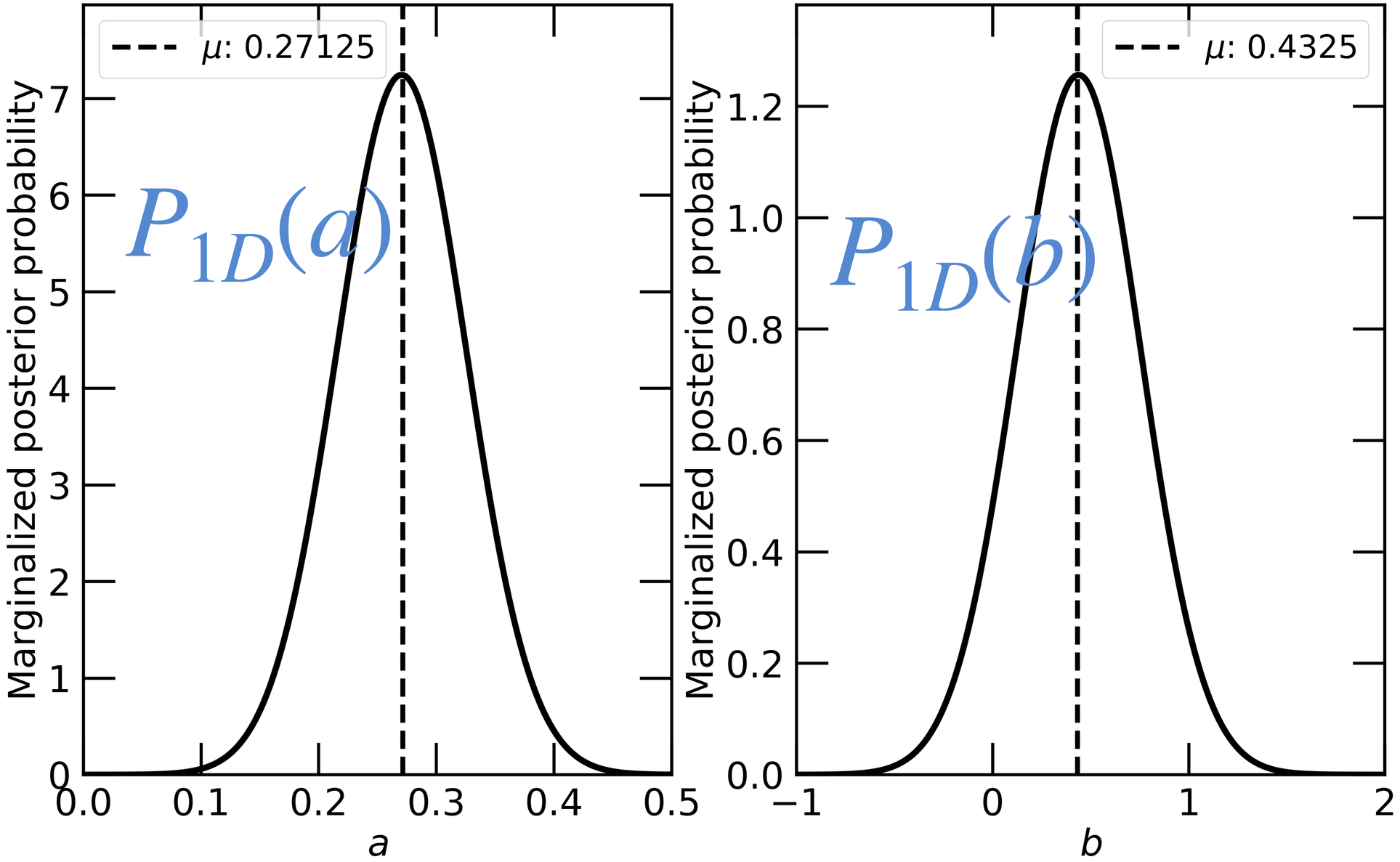
2. Bayesian statistics: Marginalization

Estimate the values of χ^2 over a grid of (a, b) and convert to 2D probability $P_{2D}(a, b) = exp(-\chi^2/2)$



2. Bayesian statistics: Marginalization

Marginalize to obtain the 1D posterior probability distributions for each parameter, $P_{1D}(a)$ and $P_{1D}(b)$

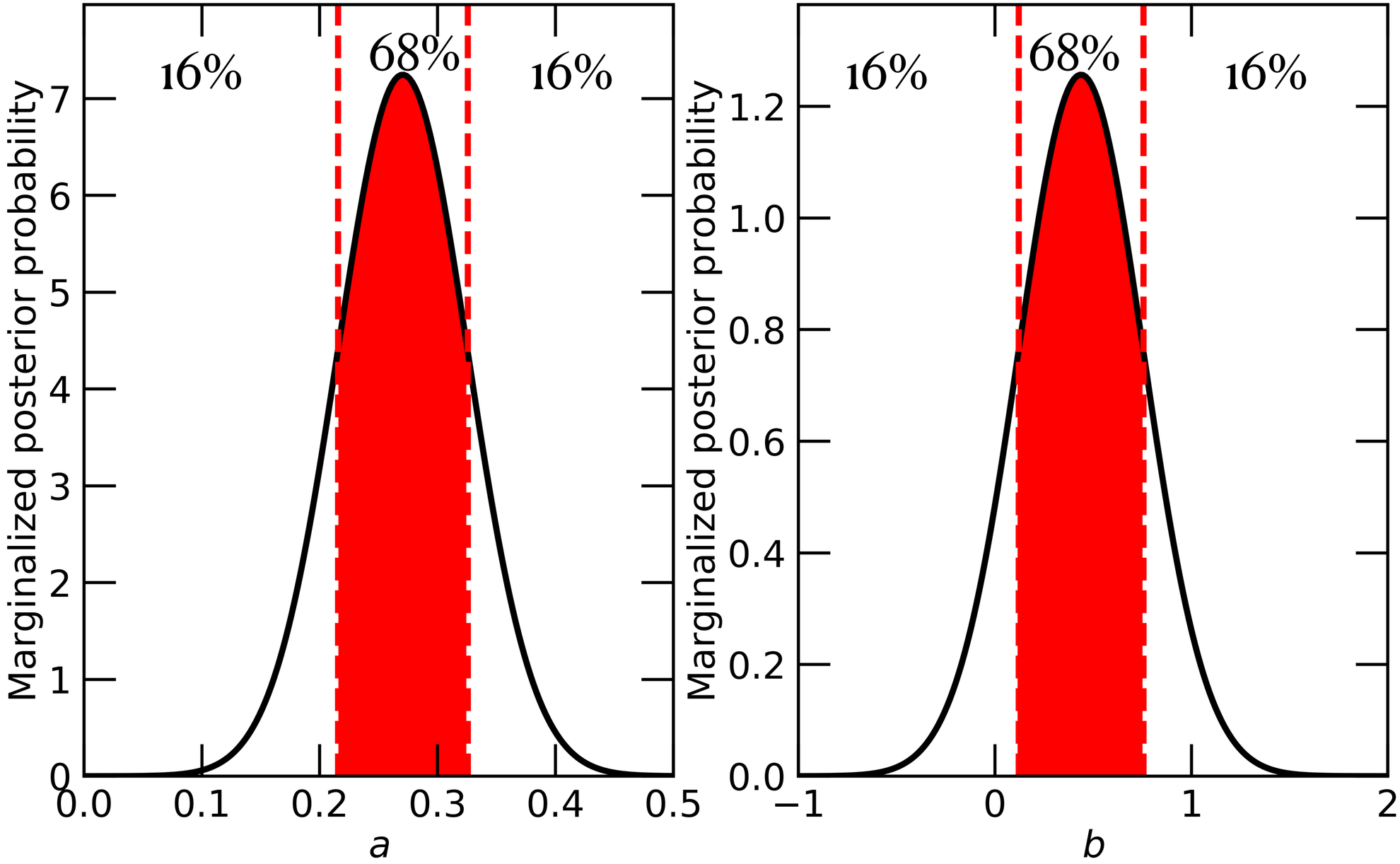


2. Bayesian statistics: Marginalization

- From the 1D posterior probability distribution, we can determine a **confidence level** of the parameter.
- Mean: $\mu_a = \int_{-\infty}^{\infty} aP_{1D}(a)da$
- Standard deviation: $\sigma_a^2 = \int_{-\infty}^{\infty} (a-\mu_a)^2 P_{1D}(a)da$
- If the 1D posterior probability distribution is Gaussian, the mean is the best fitting value and the standard deviation is 68% **confidence level**.
- Generally for different probability distribution, the confidence level is estimated by integration
- $$\int_{-\infty}^{a_{bot}} P_{1D}(a)da = 0.16 \quad ; \int_{a_{bot}}^{a_{top}} P_{1D}(a)da = 0.68 \, ; \int_{a_{top}}^{\infty} P_{1D}(a)da = 0.16$$

2. Bayesian statistics: Marginalization

Integrate under these distributions, we can identify the 68% confidence regions



2. Bayesian correlation testing

In the lecture statistic, we measure the correlation coefficient and Pearson correlation of two variables. Using Bayesian we can ask what is the posterior probability distribution for the correlation coefficient given the measurement of Pearson value.

$$P(\rho \mid r)$$

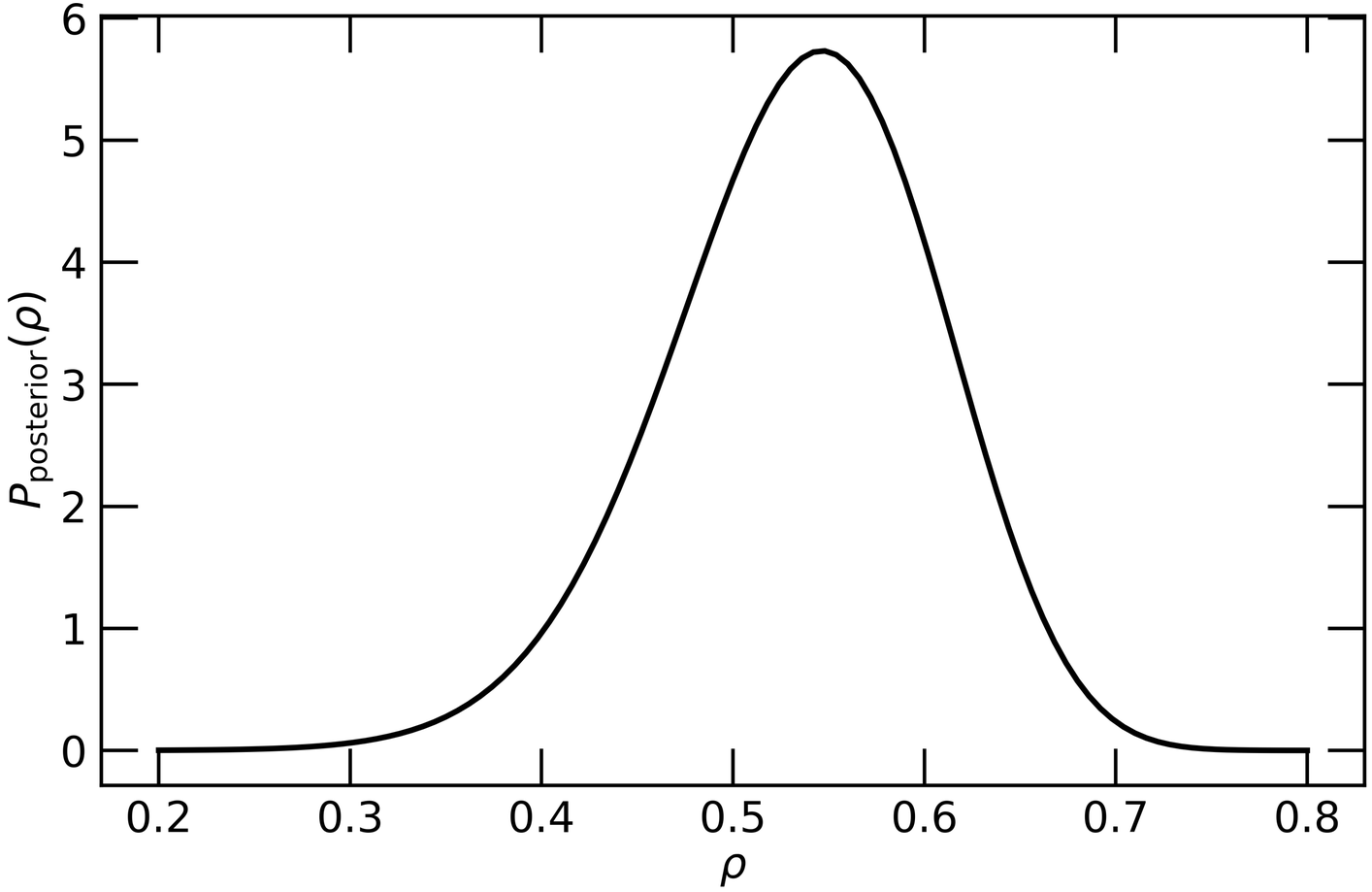
Assuming (x, y) data are drawn from N sample of a bivariate Gaussian distribution:

$$\mathrm{P(X,\,Y)} = \frac{1}{2\pi\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(\mathrm{X}-\mu_{\mathrm{X}})^2}{\sigma_{\mathrm{X}}^2} - 2\rho\frac{(\mathrm{X}-\mu_{\mathrm{X}})(\mathrm{Y}-\mu_{\mathrm{Y}})}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} + \frac{(\mathrm{Y}-\mu_{\mathrm{Y}})^2}{\sigma_{\mathrm{Y}}^2}\right)\right]$$

2. Bayesian correlation testing

The estimation of Bayes theorem for the bivariate Gaussian, marginalization over parameters:

$$P(\rho \mid r) \propto \frac{(1-\rho^2)^{(N-1)/2}}{(1-\rho r)^{N-3/2}} \left(1 + \frac{1}{N-1/2}\frac{1+\rho r}{8} + \ldots\right)$$



Thank you!

https://xkcd.com/