DATA221 Intro Machine Learning
04 Gaussians

William Trimble

Winter 2023

THE UNIVERSITY OF CHICAGO

Plan

- Least Squares Adjustment (demo)
- Definitions
- Anatomy of the Gaussian distribution
- Fitting one-dimensional Gaussian mixture distribution

Problems 1 and 2 on homework present different types of attacks on inference:

- Q2: Direct

You have a likelihood for the data given the parameters

P (x | λ)

P (childHeight $|\mu, \sigma$)

Likelihood sum is over the data points; one term for each observation

- Q2: Indirect

You have a likelihood for a statistic based on the data given the parameters (for instance a histogram or a moment)

P (z | λ) =

Poisson($\int_{Z_i}^{Z_{i+1}} Exp'l(x;\lambda)dx \ * N_{total}$)

Likelihood sum is over the histogram bins. One point for each bin.

Problems 1 and 2 on homework present different types of attacks on inference:

- Q2: Direct

You have a likelihood for the data given the parameters

P (x | λ)

P (childHeight $|\mu, \sigma$)

Likelihood sum is over the data points; one term for each observation

- Q2: Indirect

You have a likelihood for a statistic based on the data given the parameters (for instance a histogram or a moment)

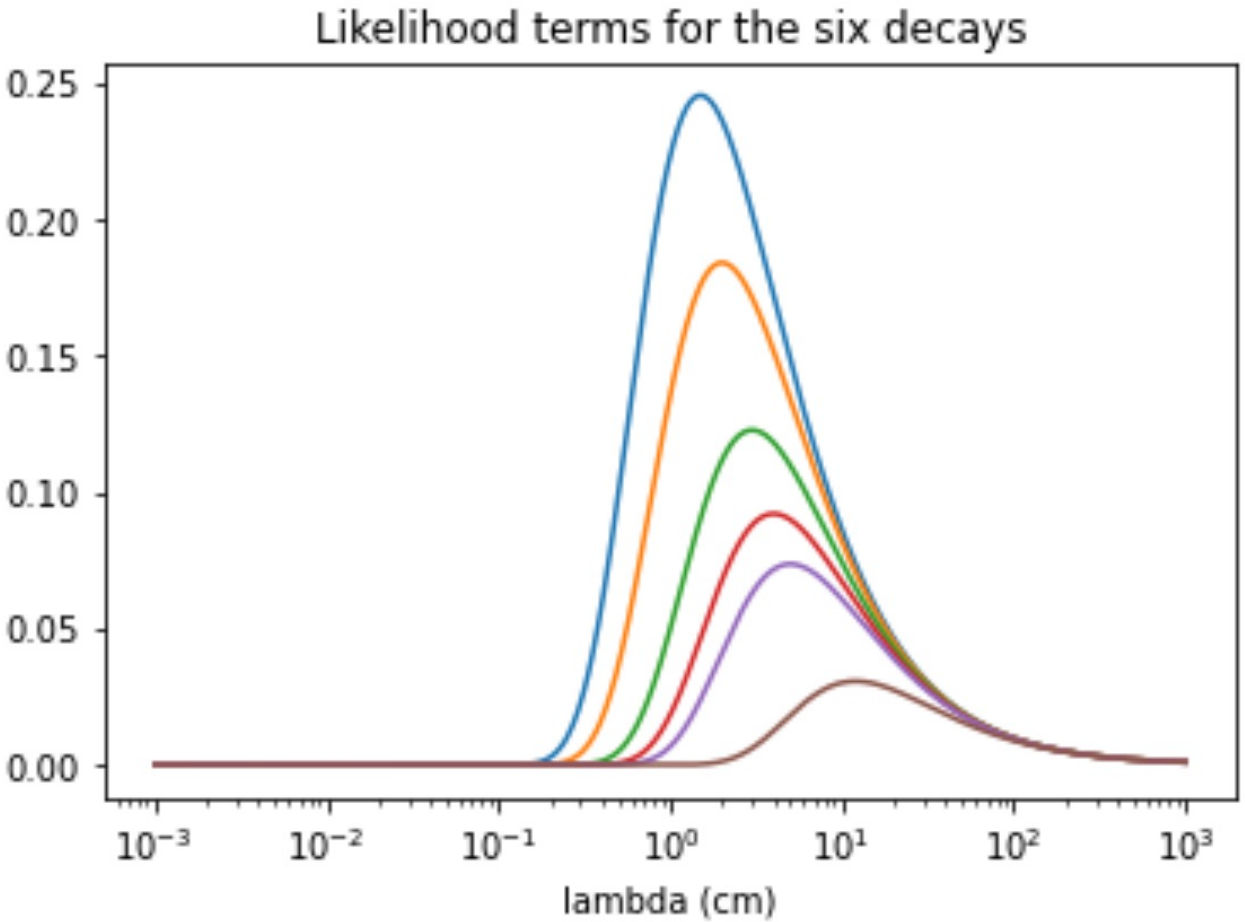
P (z | λ) =

Poisson($\int_{Z_i}^{Z_{i+1}} Exp'l(x;\lambda)dx \ * N_{total}$)

Likelihood sum is over the histogram bins. One point for each bin.

Do you have probability for the values of the data?

Newsom, Phys Rev 48 790



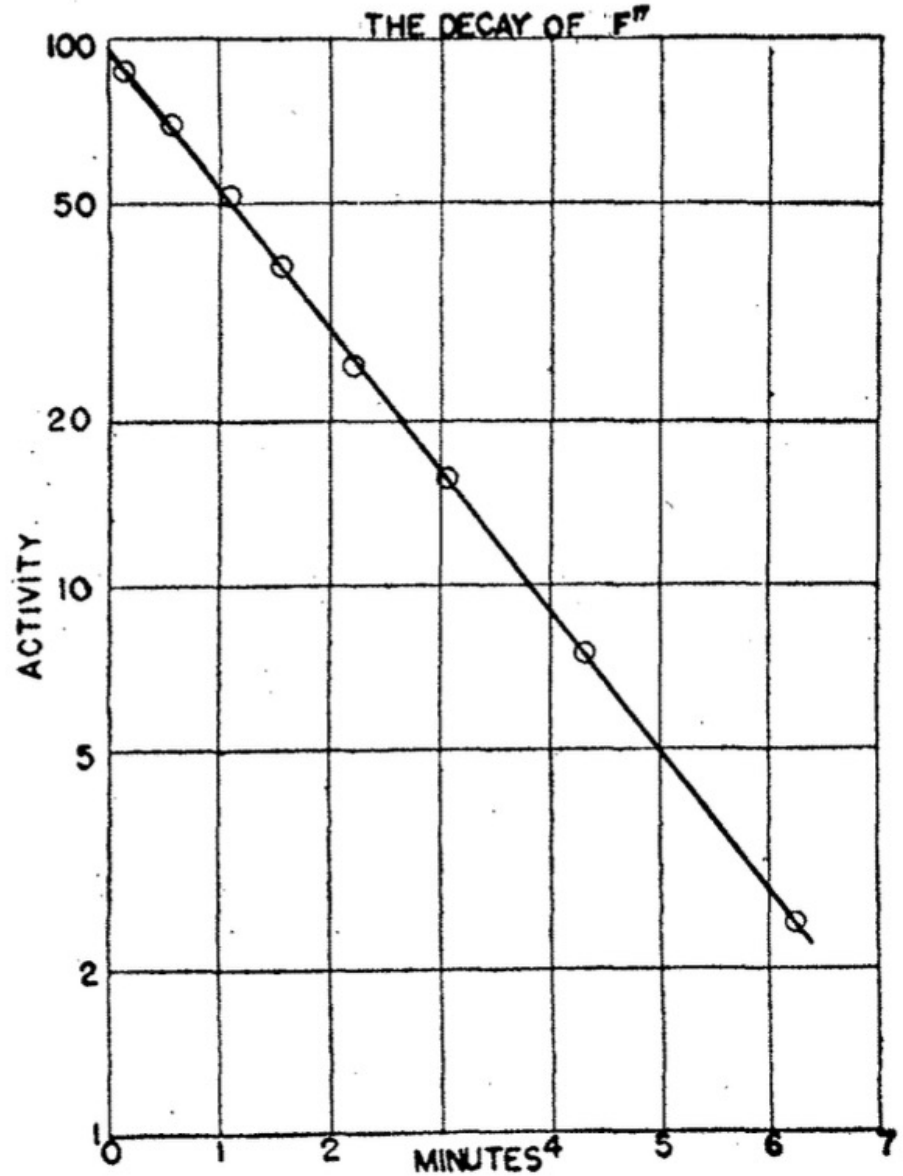


FIG. 1. A logarithmic plot of the decay of F^{17}; the half-life of this substance is found to be 1.16 minutes.

The simplest case of parameterization ambiguity: distribution of a single parameter

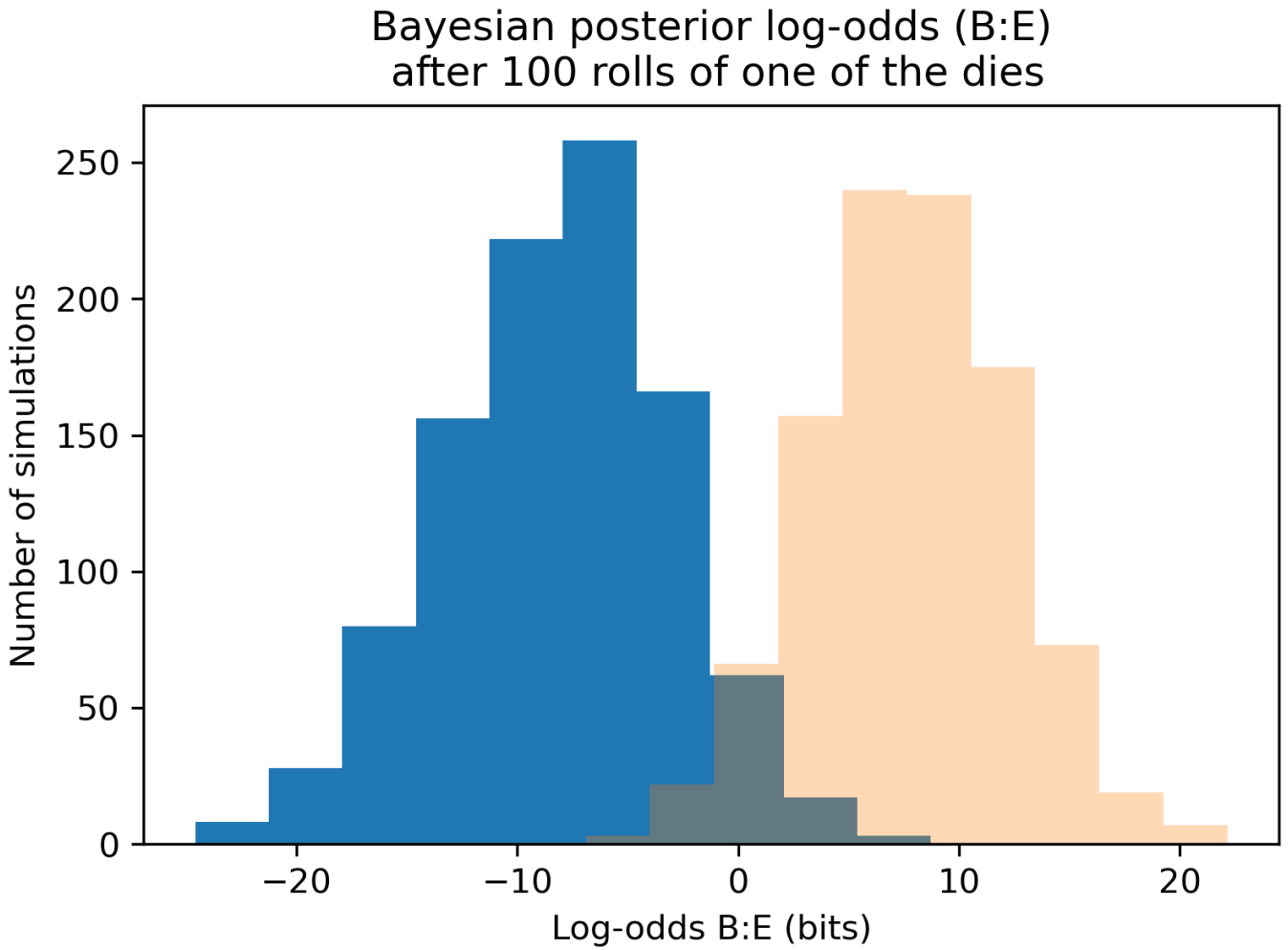
MacKay Ch3:

$$P(x \mid \lambda) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} / Z(\lambda) & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases}$$

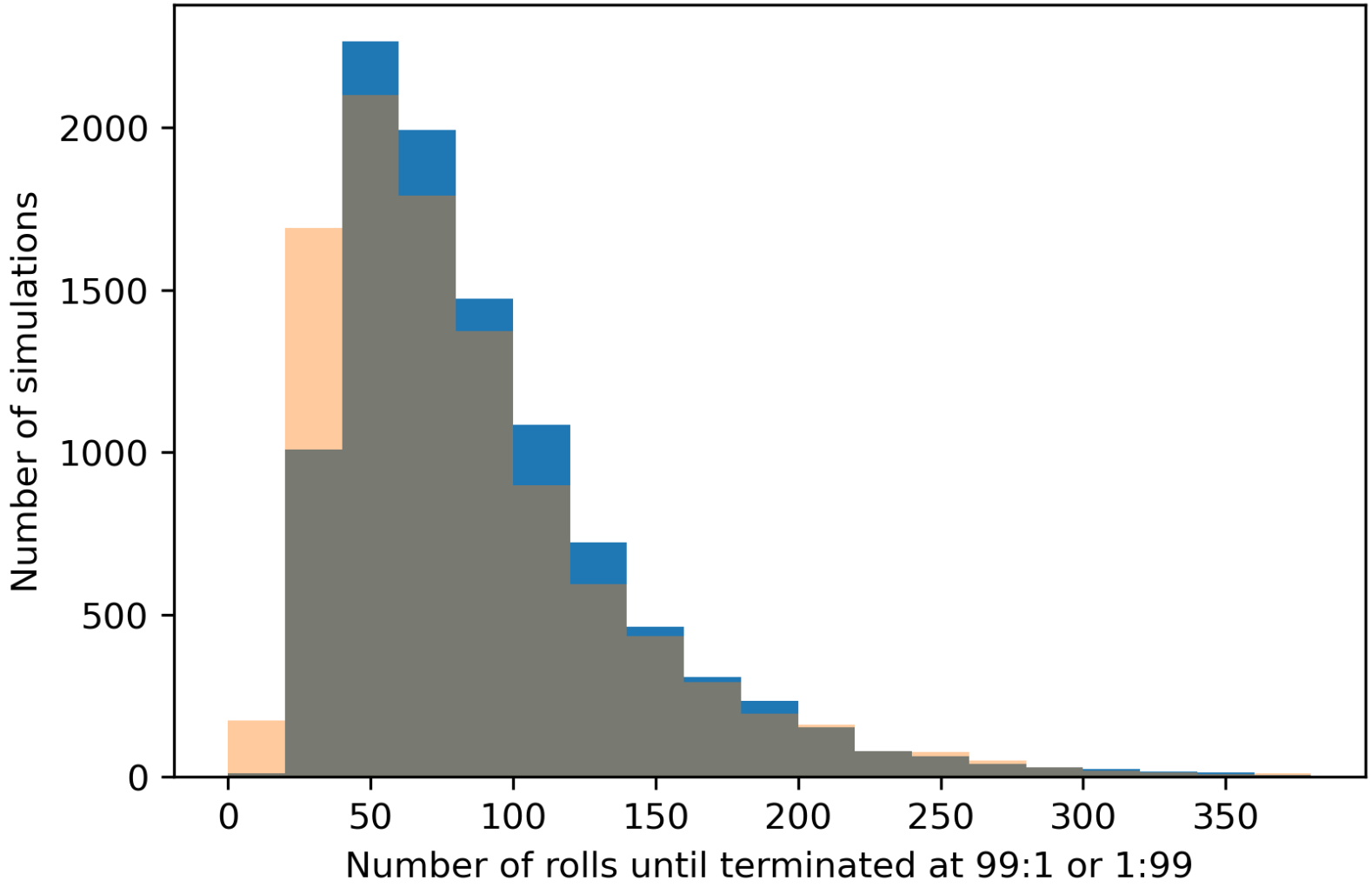
Wikipedia “Exponential distribution” $\lambda e^{-\lambda x}$ decay rate

Wikipedia “Alternate parameterization” $\frac{1}{\beta} e^{-x/\beta}$ decay length / decay time

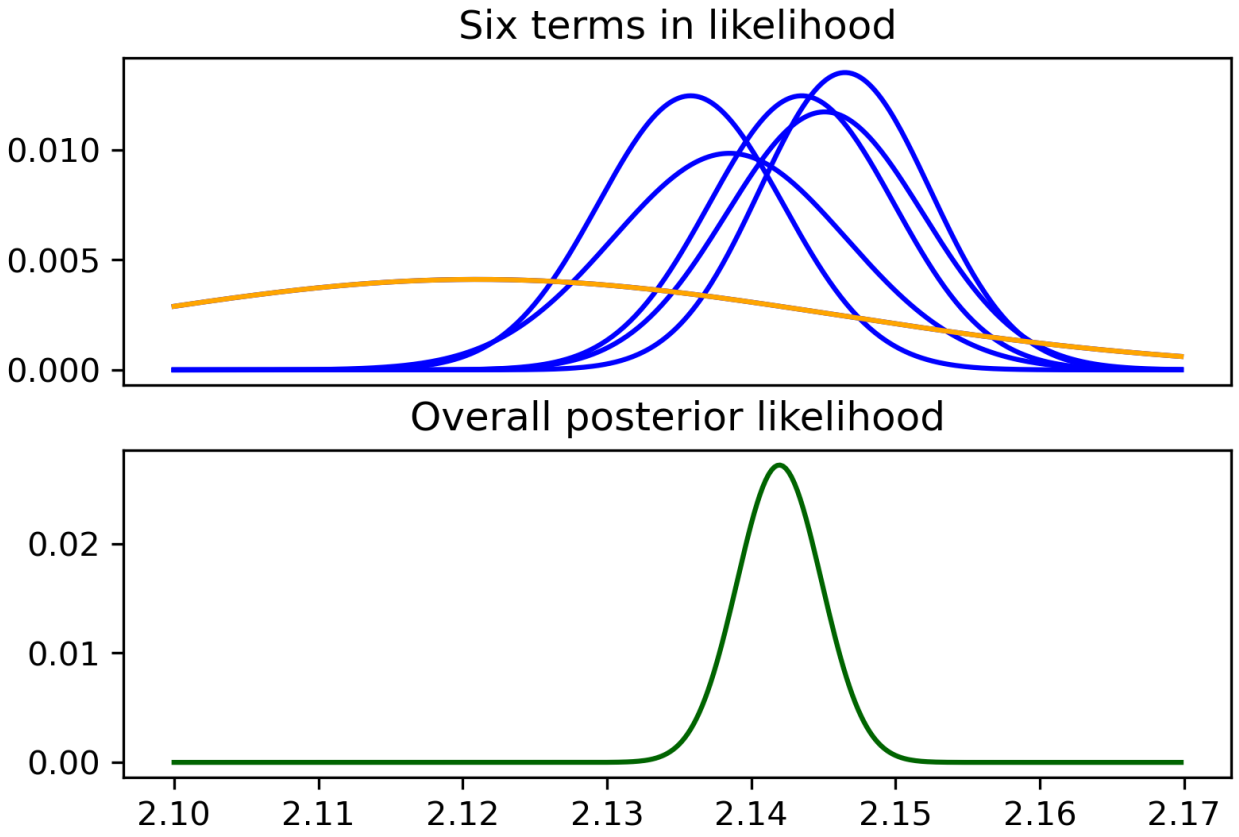
Dice problem



Dice problem



Least-squares adjustment with normal posteriors



Posterior(x) = $P_1(x)\ P_2(x)\ P_3(x)\ P_4(x)\ P_5(x)$ Prior(x)

This is the simplest Bayesian inference problem I can think of with a one-dimensional, continuous variable; the probability density at every stage is a well-behaved normal distribution.

The log-likelihoods are parabolas.

Caution: this weighted-average approach takes the standard deviations as known with certainty.

`LEASTSQADJUST.ipynb`

Least-squares adjustment

Prior(x) = N (m_1, s_0) = $\frac{1}{\sqrt{2\pi {s_0}^2}} e^{\frac{(x-m_0)^2}{2{s_0}^2}}$

Likelihood $_1$ (x) = N (m_1, s_1) = $\frac{1}{\sqrt{2\pi {s_1}^2}} e^{\frac{(x-m_1)^2}{2{s_1}^2}}$

In one dimension, imagine I have a variable x.

I want the best number I can get for P (x| DATA) (the posterior)
but all I have are a basket full of p(DATAi | x). (the likelihoods)
Bayes' rule to the rescue!
P(x | DATA) = P(x) P(DATA | x) / p(DATA)

Least-squares adjustment

Prior(x) = N (m_1, s_0) = $\frac{1}{\sqrt{2\pi s_0^2}} e^{\frac{(x-m_0)^2}{2s_0^2}}$

Likelihood $_1$ (x) = N (m_1, s_1) = $\frac{1}{\sqrt{2\pi s_1^2}} e^{\frac{(x-m_1)^2}{2s_1^2}}$

Likelihood $_2$ (x) = N (m_2, s_2) = $\frac{1}{\sqrt{2\pi s_2^2}} e^{\frac{(x-m_2)^2}{2s_2^2}}$

Likelihood $_3$ (x) = N (m_3, s_3) - $\frac{1}{\sqrt{2\pi s_3^2}} e^{\frac{(x-m_3)^2}{2s_3^2}}$

…

Posterior = Prior * L1 * L2 * L3 …

Take prior to be constant for now; take logs of all the likelihoods:

Prior(x) = constant

Likelihood $_1$ (x) = N (m_1, s_1) = $\frac{1}{\sqrt{2\pi s_1^2}} e^{\frac{(x-m_1)^2}{2s_1^2}}$

Likelihood $_2$ (x) = N (m_2, s_2) = $\frac{1}{\sqrt{2\pi s_2^2}} e^{\frac{(x-m_2)^2}{2s_2^2}}$

Likelihood $_3$ (x) = N (m_3, s_3) - $\frac{1}{\sqrt{2\pi s_3^2}} e^{\frac{(x-m_3)^2}{2s_3^2}}$

…

Posterior = Prior * L1 * L2 * L3 …

Least-squares adjustment

log Prior(x) = $\log C$ = C

LLikelihood_1 (x) = log N (m_1, s_1) = $-\frac{1}{2}\log(2\pi {s_1}^2) - \frac{(x-m_1)^2}{2{s_1}^2}$

LLikelihood_2 (x) = log N (m_2, s_2) = $-\frac{1}{2}\log(2\pi {s_2}^2) - \frac{(x-m_2)^2}{2{s_2}^2}$

LLikelihood_3 (x) = log N (m_3, s_3) = $-\frac{1}{2}\log(2\pi {s_3}^2) - \frac{(x-m_3)^2}{2{s_3}^2}$

…

log Posterior = log Prior + LL1 + LL2 + LL3 …

Least-squares adjustment

log Prior(x) = $\log C$ = C

LLikelihood_1 (x) = log N (m_1, s_1) = $-\frac{1}{2}\log(2\pi {s_1}^2) - \frac{(x-m_1)^2}{2{s_1}^2}$

LLikelihood_2 (x) = log N (m_2, s_2) = $-\frac{1}{2}\log(2\pi {s_2}^2) - \frac{(x-m_2)^2}{2{s_2}^2}$

LLikelihood_3 (x) = log N (m_3, s_3) = $-\frac{1}{2}\log(2\pi {s_3}^2) - \frac{(x-m_3)^2}{2{s_3}^2}$

…

While these are straightforward functions of the s_i, these are just parabolas as a function of x. Sum of parabolas is a …

Least-squares adjustment

log Prior(x) = $\log C$ = C

LLikelihood_1 (x) = log N (m_1, s_1) = $-\frac{1}{2}\log(2\pi {s_1}^2) - \frac{(x-m_1)^2}{2{s_1}^2}$

reorganize terms in powers of x

log Posterior = $-\frac{1}{2}\sum\frac{1}{{s_i}^2}\ x^2 + \frac{2}{2}\sum\frac{m_i}{{s_i}^2}\ x - \frac{1}{2}\sum\frac{{m_i}^2}{{s_i}^2}$

$\frac{\text{d log(posterior)}}{dx} = -\sum\frac{1}{{s_i}^2}\ x + \sum\frac{m_i}{{s_i}^2} = 0$

$$x_{max} = \frac{\sum\frac{m_i}{{s_i}^2}}{\sum\frac{1}{{s_i}^2}} \qquad s_{tot} = \frac{1}{\sqrt{\sum\frac{1}{{s_i}^2}}}$$

Ta-da! (Undergrad-level combination of measurements)

$x_{max} = \frac{\sum \frac{m_i}{s_i^2}}{\sum \frac{1}{s_i^2}}$ $s_{tot} = \frac{1}{\sqrt{\sum \frac{1}{s_i^2}}}$

When you have a collection of normally-distributed likelihoods with known means and standard deviations

(These represent measurements and someone's judgement of the uncertainty attached to each measurement.)

We have a procedure that turns likelihood functions into algebra

And the product of N normal distributions is a normal distribution that is much tighter.

Ta-da! (Undergrad-level combination of measurements)

$x_{max} = \frac{\sum \frac{m_i}{s_i^2}}{\sum \frac{1}{s_i^2}}$ $s_{tot} = \frac{1}{\sqrt{\sum \frac{1}{s_i^2}}}$ Posterior (x) = N (x_{max}, s_{tot})

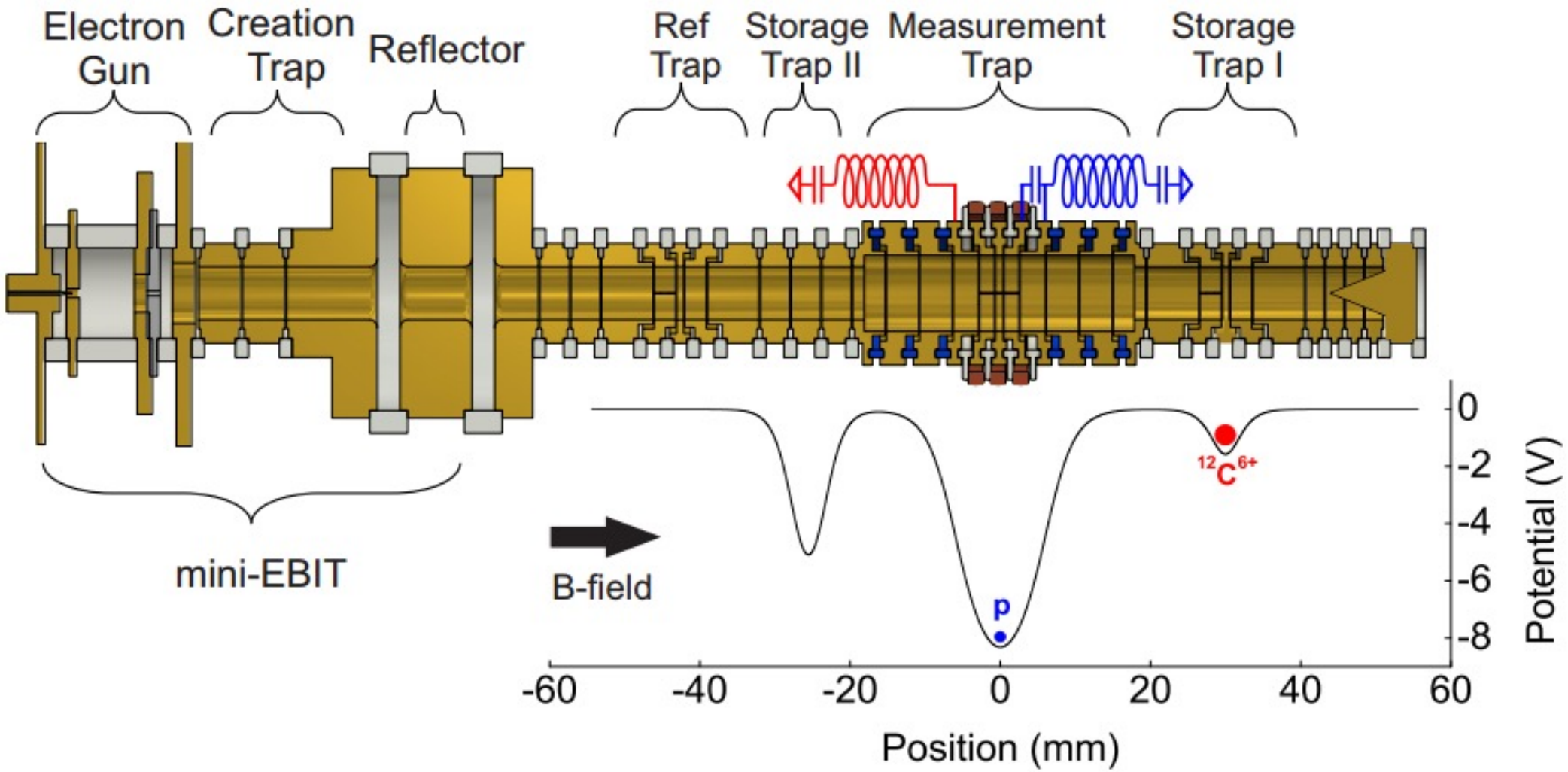
When you have a collection of normally-distributed likelihoods with known means and standard deviations

(These represent measurements and someone's judgement of the uncertainty attached to each measurement.)

We have a procedure that turns likelihood functions into algebra

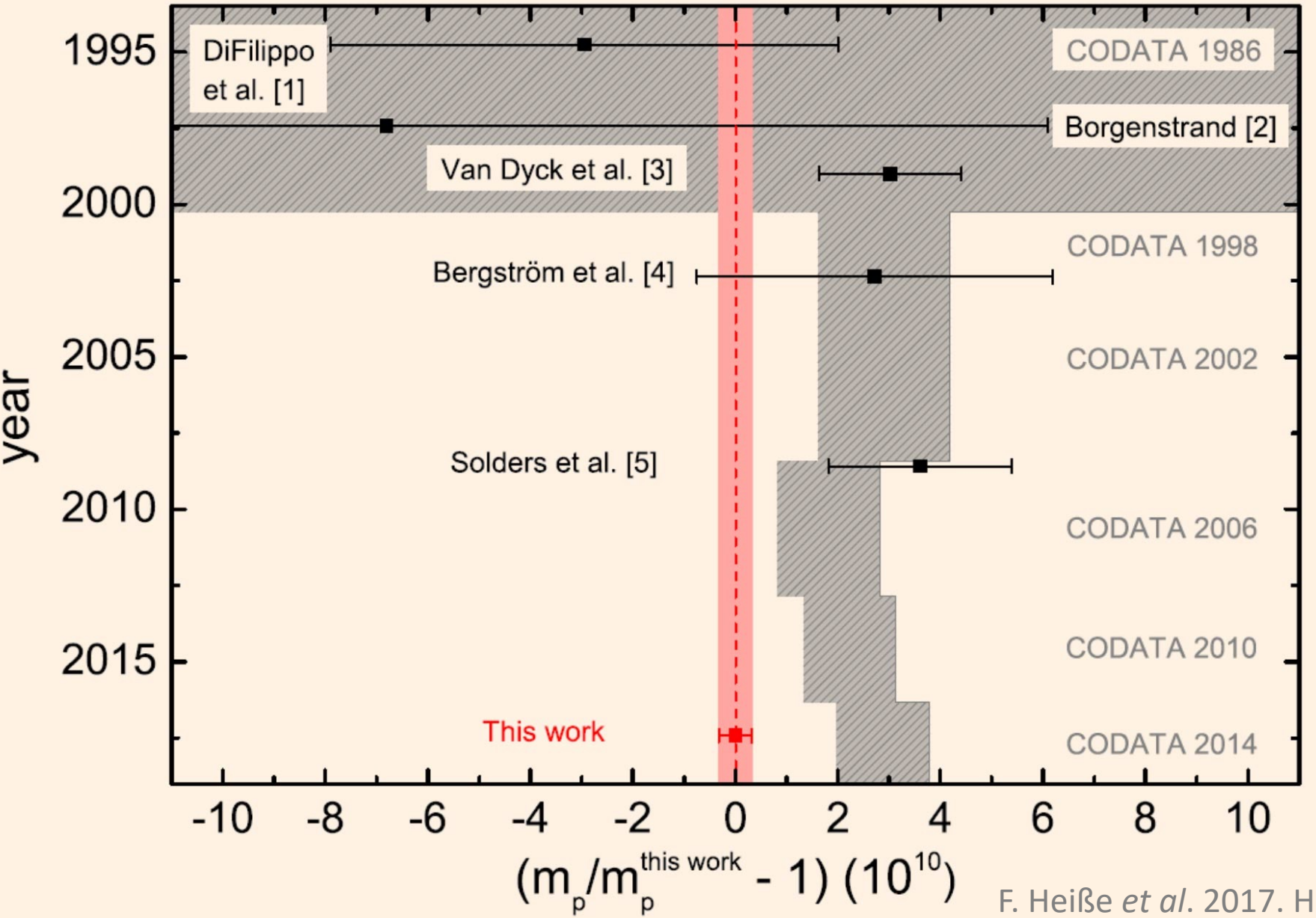
And the product of N normal distributions is a normal distribution that is much tighter.

Caution: this procedure does not know or care whether the measurements agree with each other.



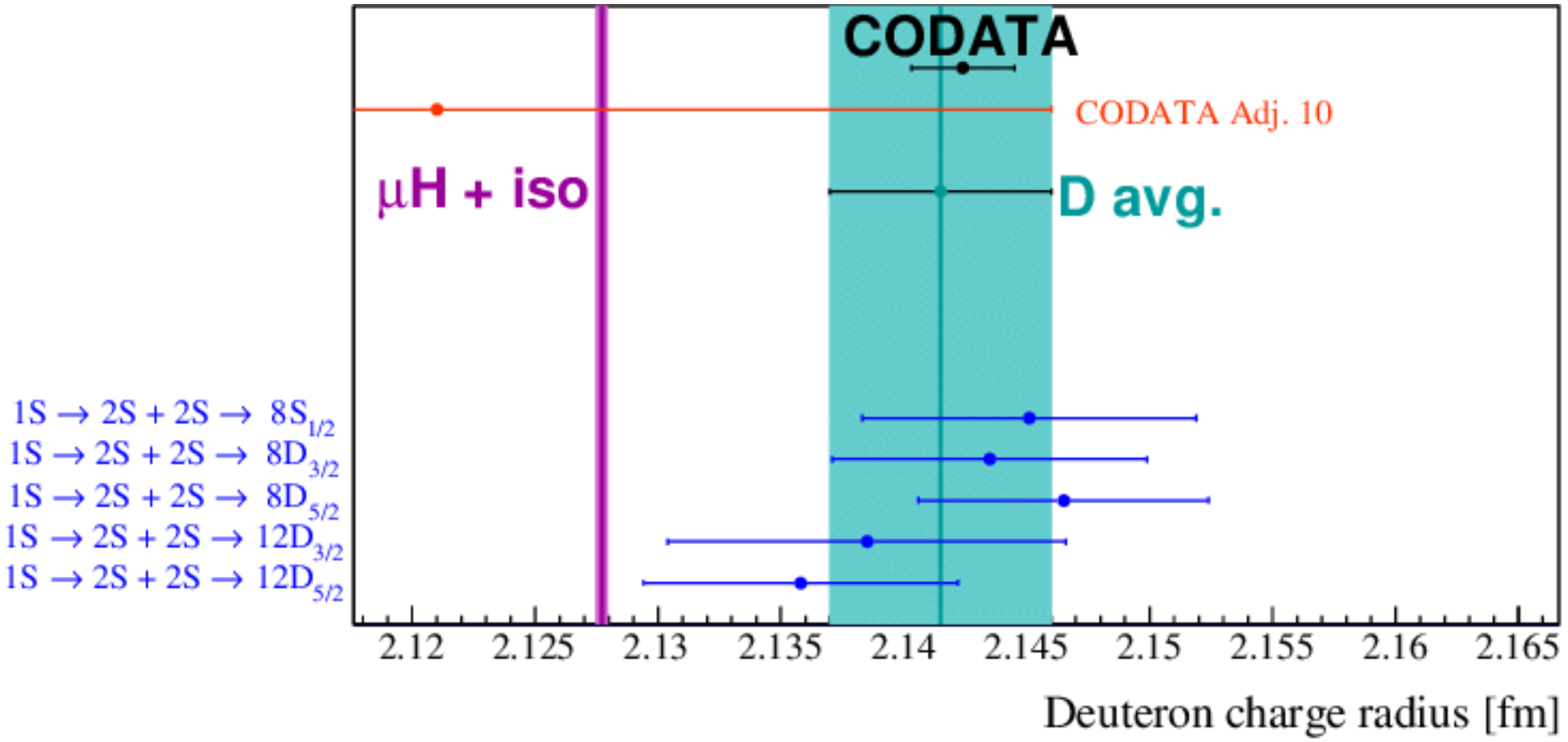
This is a balance that compares the mass of the proton to the mass of $^{12}C^{6+}$

F. Heiße *et al*. 2017. High-Precision Measurement of the Proton's Atomic Mass. *Phys. Rev. Lett* 119 (3): 033001; doi: 10.1103/PhysRevLett.119.033001

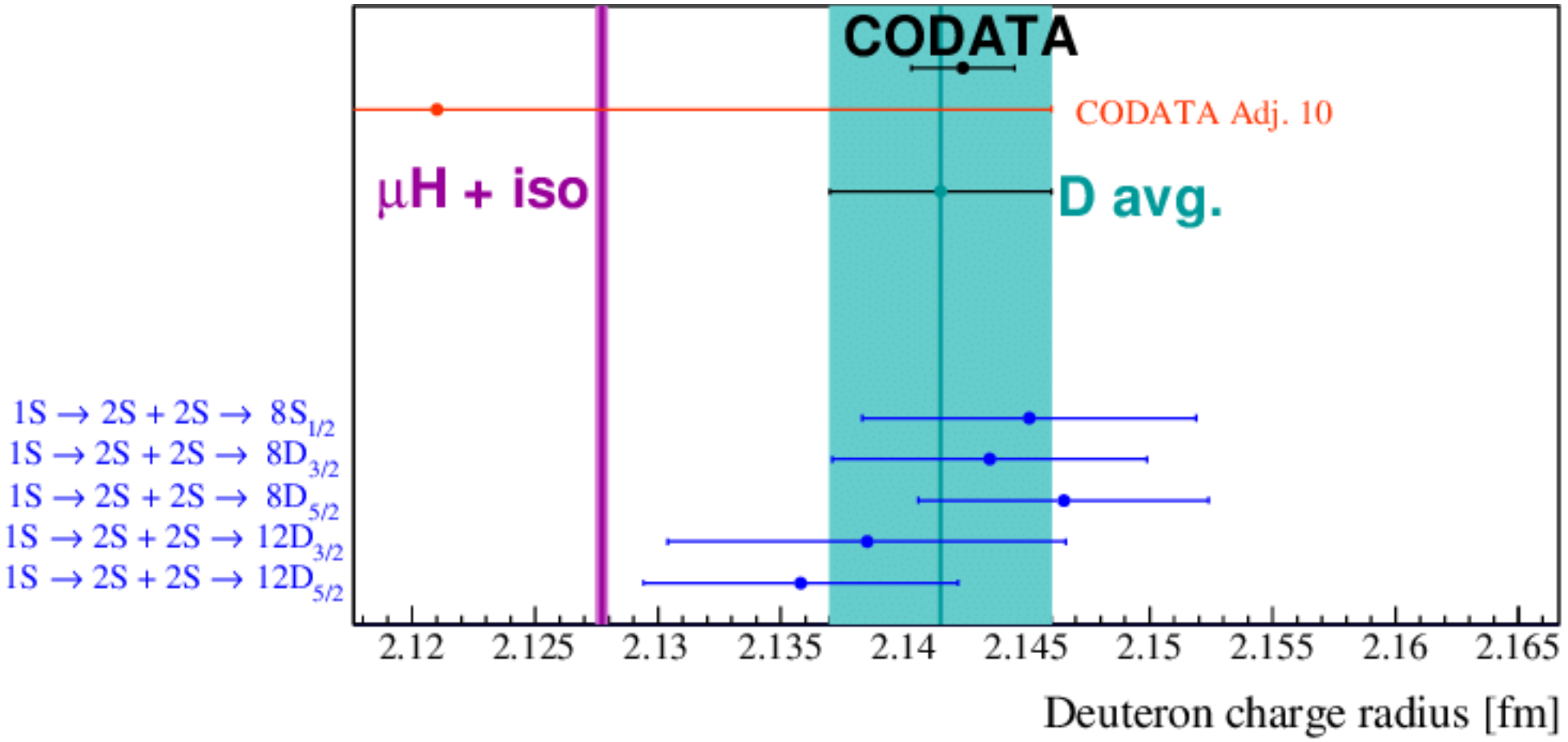


Multiple researchers, with advancing technology, have taken cracks at measuring the mass of the proton.

F. Heiße *et al*. 2017. High-Precision Measurement of the Proton's Atomic Mass. *Phys. Rev. Lett* 119 (3): 033001; doi: 10.1103/PhysRevLett.119.033001



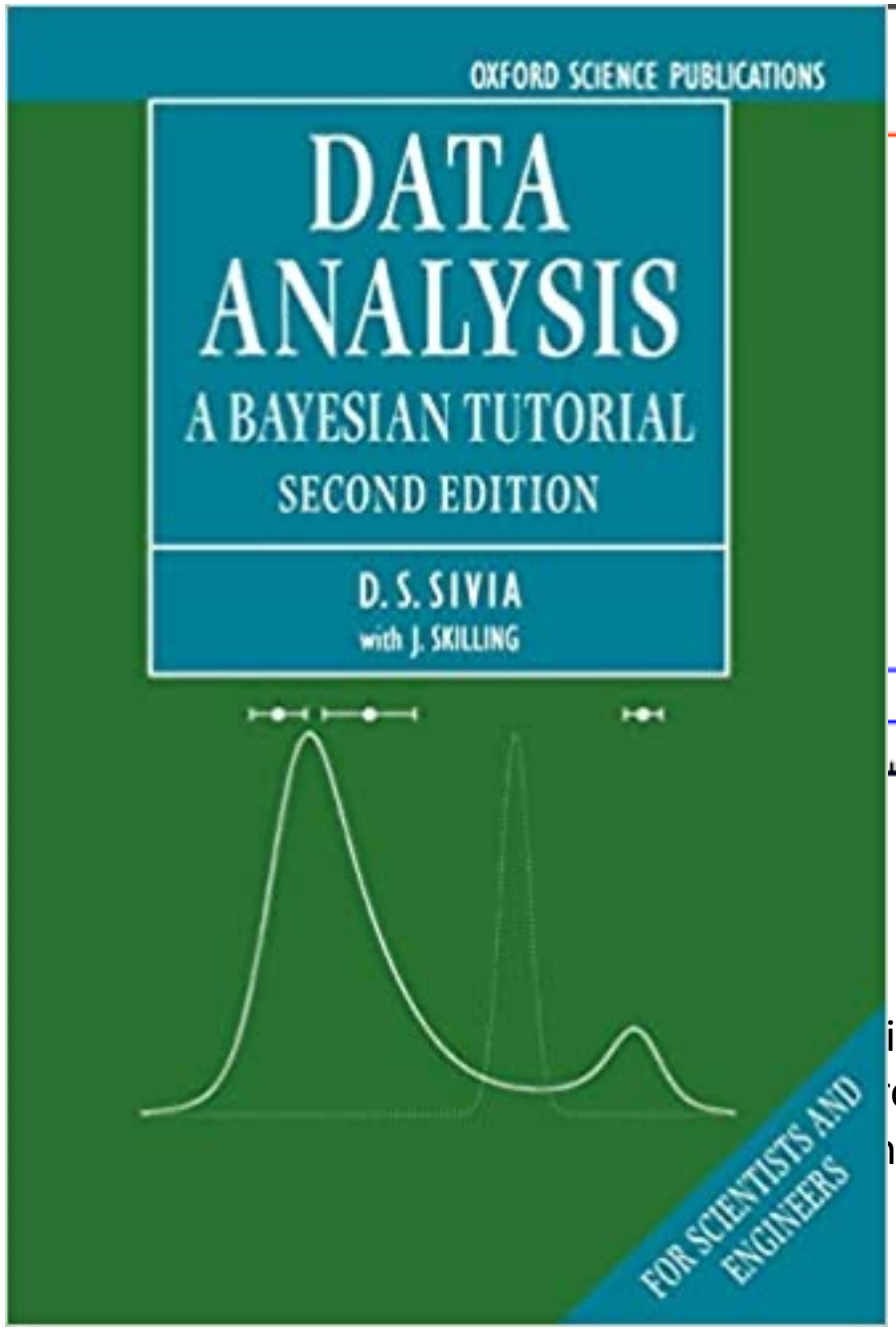
In a different measurement campaign, there are multiple (technically different) measurement approaches to the deuteron charge radius.

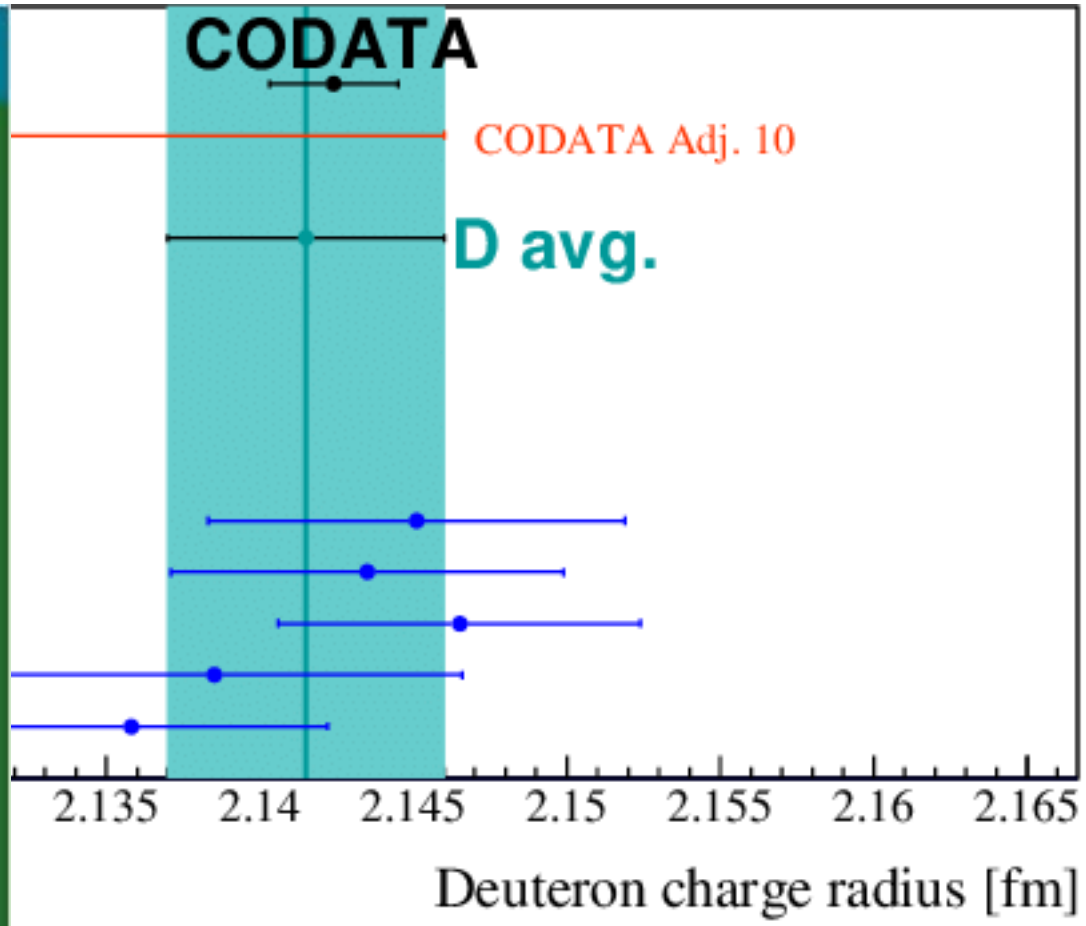


In a different measurement campaign, there are multiple (technically different) measurement approaches to the deuteron charge radius.

In a measurement-evaluation doctrine called “Least Squares Adjustment,” information about the physical constant (deuteron charge radius) is combined from multiple experiments using a very lightweight version of Bayesian inference.

Goes by the term “metaanalysis” in other fields.





In a different measurement campaign, there are multiple (technically different) measurement approaches to the deuteron charge radius.

ine called “Least Squares Adjustment,” information
on charge radius) is combined from multiple
t version of Bayesian inference.

other fields.

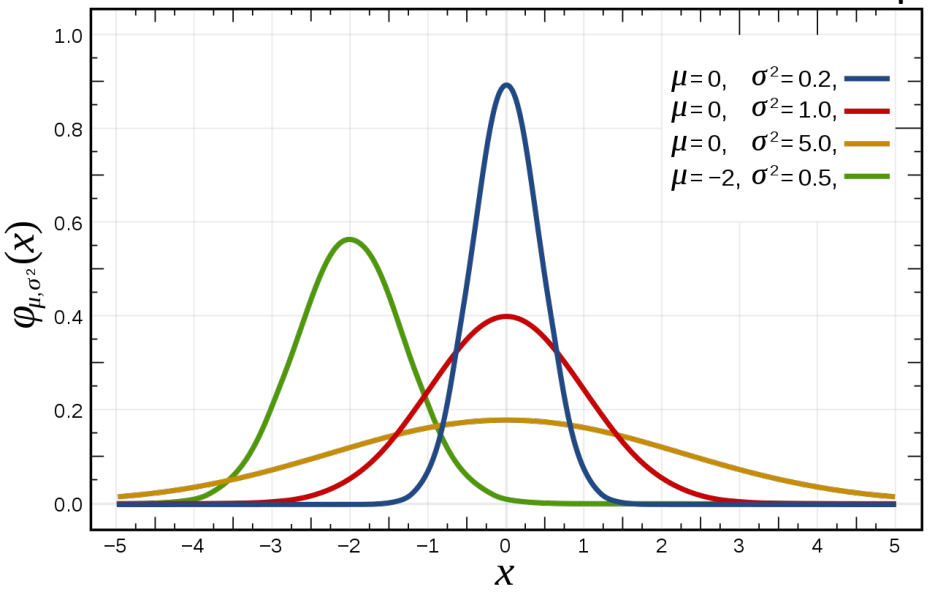
Anatomy of the normal distribution

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

"The normalizing constant"

peak density proportional to $1/\sigma$

x- dependence
shape and position
of distribution



Anatomy of the normal distribution

$$\log \mathcal{N}(x|\mu,\sigma) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

Constant term, not going to give us trouble

Term depends only on σ

Term depends on σ and μ

Anatomy of the normal distribution

$$\log \mathcal{N}(x|\mu,\sigma) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

Constant term, not going to give us trouble

Term depends only on σ

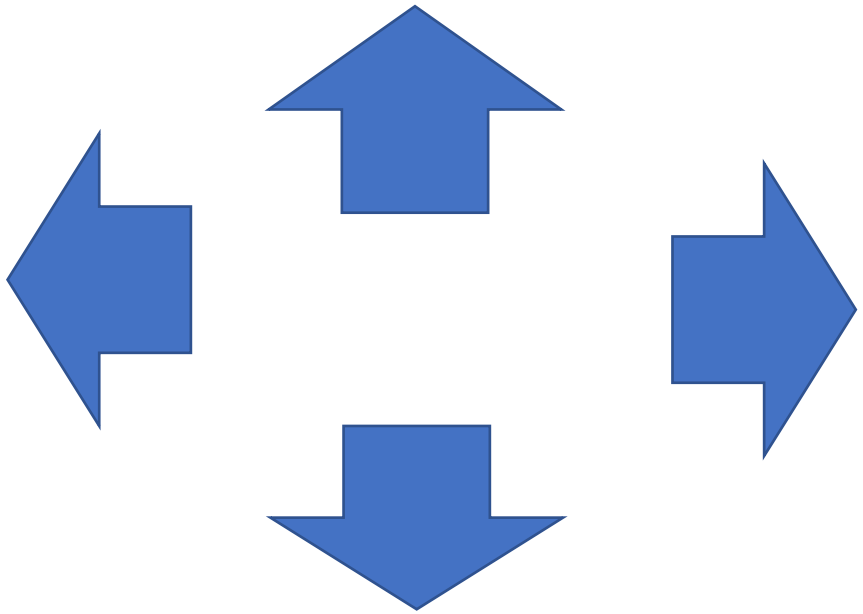
Term depends on σ and μ

Penalizes distributions that are too spread out

Tells us where the center is

Anatomy of the normal distribution

$$\log \mathcal{N}(x|\mu,\sigma) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$



When σ is too large, this term makes likelihood small

When σ is too small, this term penalizes the likelihood because the distribution will miss the data.

Normal(x_1) x Normal (x_2) ??

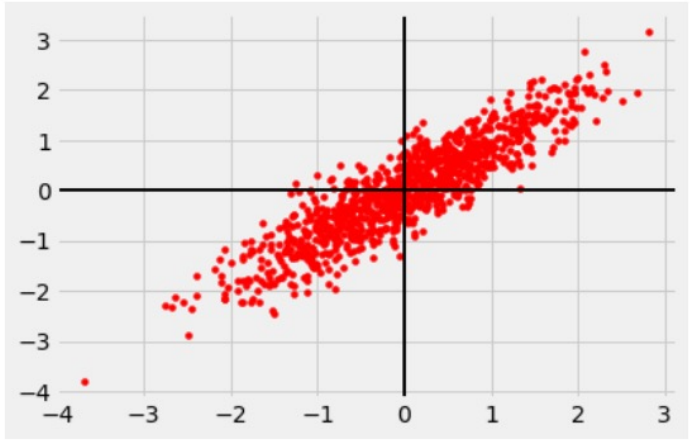
$$\mathcal{N}(x_1, x_2|\mu_1, \mu_2, \sigma_{11}, \sigma_{22}) = \frac{1}{(2\pi)} \frac{1}{\sigma_{11}\sigma_{22}} \exp -\frac{1}{2} \left(\frac{(x_1-\mu_1)^2}{\sigma_{11}} + \frac{(x_2-\mu_2)^2}{\sigma_{22}} \right)$$

If I take the product of two 1-D normal distributions, I get a perfectly good probability distribution, but it isn't quite flexible enough for my data-modeling needs.

This factorizes into the product of a term that depends only on x_1 and σ_1 and a term only in x_2 and σ_2.

This expression doesn't permit me to make distributions where the probability density of x_1 depends on the value of x_2

Multivariate normal



$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 \end{bmatrix}$$

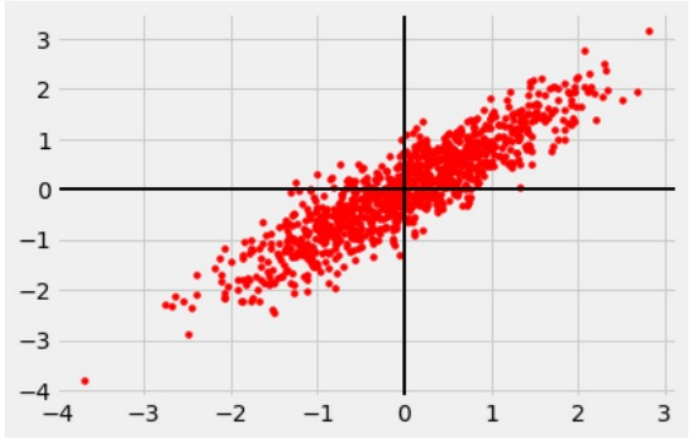
$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{(\sigma_{11}^2 \sigma_{22}^2 - \sigma_{12}^2)} \exp -\frac{1}{2}\left(\begin{bmatrix} x_0 - \mu_0, & x_1 - \mu_1 \end{bmatrix} \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 \end{bmatrix}^{-1} \begin{bmatrix} x_0 - \mu_0 \\ x_1 - \mu_1 \end{bmatrix} \right)$$

Constant term, not going to give us trouble

Term depends only on Σ

Two powers of x-μ

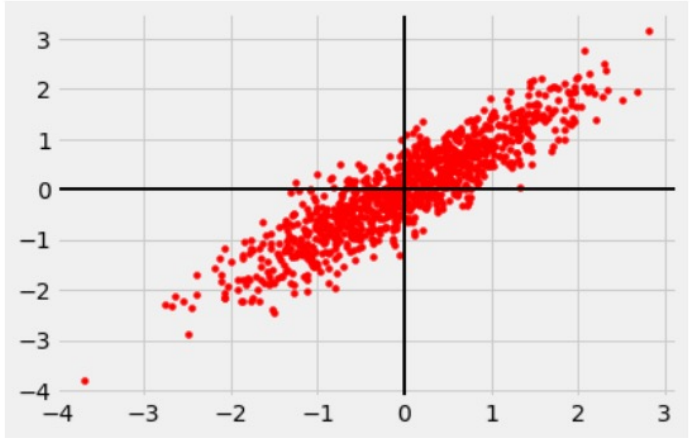
Multivariate normal



$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1D}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{D1}^2 & \cdots & \sigma_{DD}^2 \end{bmatrix}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{D/2}} \exp -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

Multivariate normal

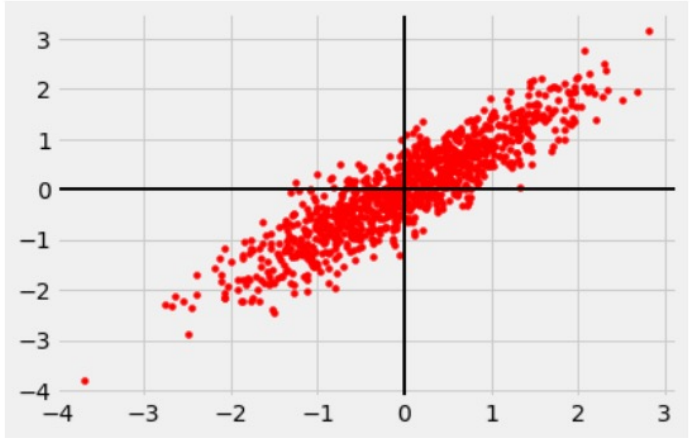


$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1D}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{D1}^2 & \cdots & \sigma_{DD}^2 \end{bmatrix}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{D/2}} \exp -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

How many parameters do I need for a D-dimensional multivariate Gaussian ?

Multivariate normal



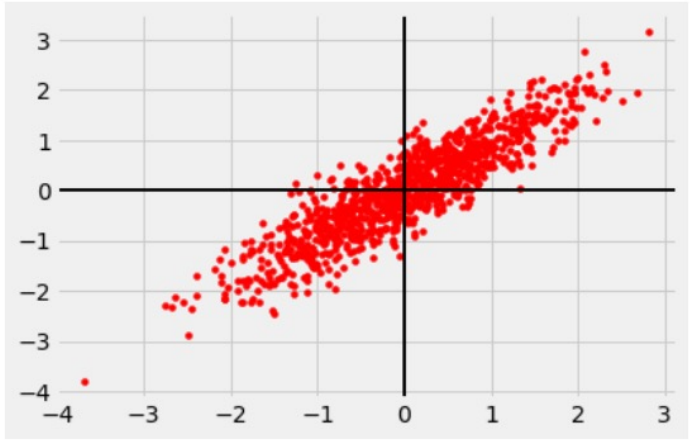
$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1D}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{D1}^2 & \cdots & \sigma_{DD}^2 \end{bmatrix}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{D/2}} \exp -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

How many parameters do I need for a D-dimensional multivariate Gaussian ?

μ : D Σ : $(D^2 + D)/2$

Multivariate normal



$$\mathcal{N}(\mathbf{x}|\mathbf{x_0},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{D/2}}\exp -\frac{1}{2}(\mathbf{x}-\mathbf{x_0})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x_0})$$

Constant term, not going to give us trouble

Term depends only on Σ

Two powers of $x\text{-}x_0$

Σ covariance matrix

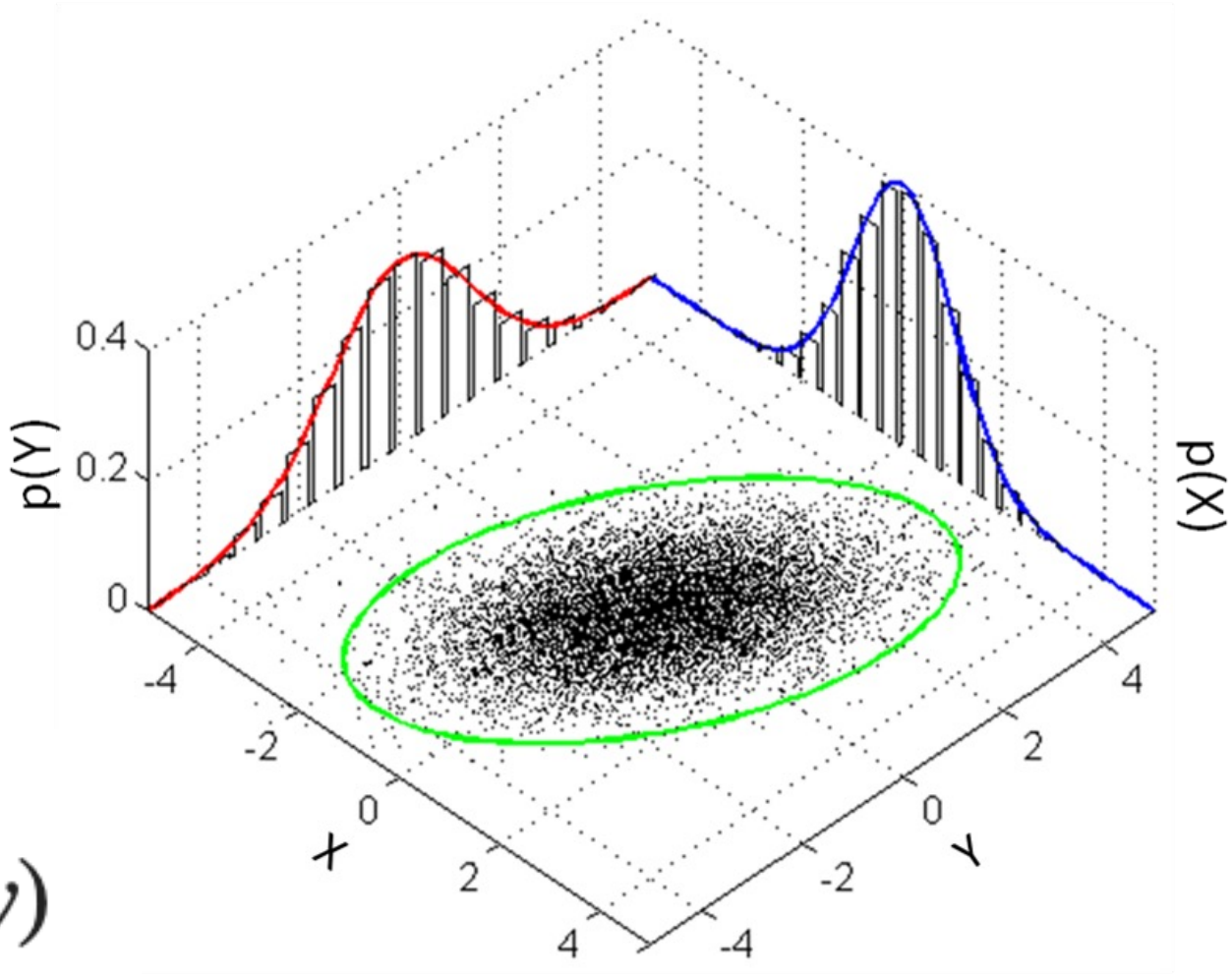
Σ^{-1} “precision” matrix

Jargon: Marginalization

To “marginalize” is to take a distribution of many variables and turn it into a distribution of fewer variables by integrating the product of the density and the prior probability of each of the unwanted variables.

When we are able to do this, this gives us the one-dimensional probability distribution (for the parameter of interest) that we seek, in essence averaging over all likely values of the variables not of interest.

$$P(x) = \int dy P(x|y)P(y)$$



Unsupervised learning

I don’t have labels, just data $\{x_i\}_{i=1}^{N}$

My model is a function that does something useful with new data.

Tasks may include

detecting outliers (points where the density is low; **density estimation**)

dimensionality reduction (produce coordinates in a low-dimensional space; compress the data while retaining its essence)

clustering (output is a cluster id)

Supervised learning

Supervised learning: the **dataset** is a a collection of
labeled examples $\{ x_i, y_i \}_{i=1}^{N}$

N examples
x is called a **feature vector** and
y is the set of **labels**

x, y can be in $\mathbb{R}^n$ or can be discrete, categorical labels, or some more complex kind of input and output

Goal of supervised learning is to produce a model that turns a feature vector into an approximation of the label.

Jargon

parameter = numbers that determine “the model.”

hyperparameter = number that controls the behavior of the learning algorithm

regression = prediction of a real-valued label

fitting = determining the parameters of the model

classification = prediction of a label from a finite set of labels

shallow learning algorithm – fits the parameters of the model from the features of the training data.

deep learning algorithm fits the parameters of the model from the output of layers that are fed by the training features.

Generative vs. discriminative

If I fit a probability distribution to my data, I can evaluate

P(newdata | model) and

P (label | model, newdata)

but I also can draw samples from the distribution. Models of this sort are called “generative” models because they permit me to “generate” new data consistent with the model parameters.

Models that just find the function of x that gives the right answer y (say, by fitting weights directly rather than finding weights from fitted distribution parameters) are called “discriminative” models.