Bayesian Modeling
Introduction and examples

Yixin Wang

Preliminary Draft.
Please do not distribute.

1. Introduction and examples

1. What is Bayesian inference?

Bayesian inference is a major framework for statistical inference. vs frequentist inference

In general, statistical inference is the (computational) process of turning data into some form of data summarization, prediction and understanding. interpretation estimation, hypothesis testing

Bayesian inference, or more broadly speaking, Bayesian statistics, is often contrasted with a competing framework known as frequentist (or classical) statistics.

In this course, we refer to statistical inference and statistical learning interchangeably.

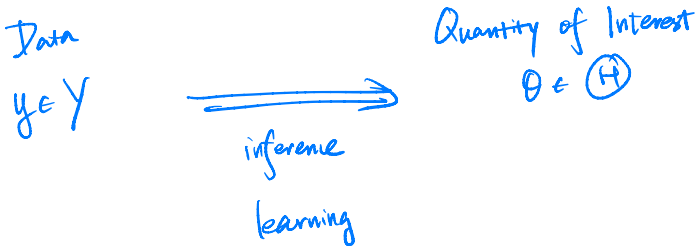
Bayesian inference
probabilistic inference
probabilistic machine learning → large-scale Bayesian inference
probabilistic modeling

There are two main players in statistical inference: data and quantity of inferential interest.

The data are represented by a variable y taking values in some suitable space $\mathcal{Y}$.

The quantity of interest is denoted by θ taking values in another space Θ.

Typically θ represents some characteristic of the data population that we wish to understand.



For inference to be possible, there must be a "linkage" between θ and the observed data y.

This linkage is formalized by a sampling model (statistical model) for which θ is viewed as the model parameter: the true θ that is responsible for generating the observed data y is unknown.

As such, θ encodes our understanding of the data population.

It is the quantity of interest.

quantity of interest θ

Example 1.1. Suppose we are interested in the prevalence of an infectious disease in a city. Data y are obtained from a random sample of individuals from the city, namely, the total number of people in the sample who are infected. Of interest is θ, the fraction of infected individuals in the city. Thus, $\Theta = [0, 1]$, while $Y = \{0, 1, 2, 3, \ldots, \}$.

data (count) $y \in \mathbb{Z}$ → prevalence $\theta \in [0,1]$

inference

linkage = $\theta \sim y$

$\theta = 0$, $y = 10$ is it plausible?

cannot generate the data

Q: what values of θ can generate the data y?

are more likely to

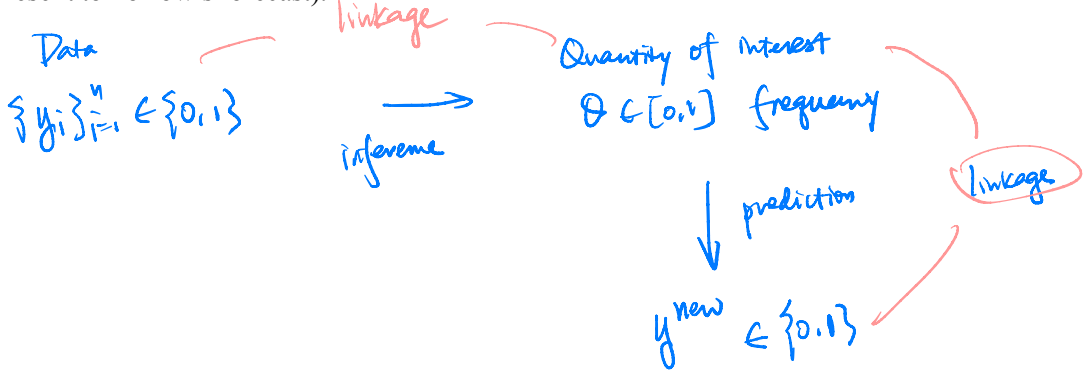
Example 1.2. y represents a collection of heights sampled from a population, θ the typical height. Here, $\Theta = \mathcal{Y} = \mathbb{R}^+$.

Example 1.3. y represents polling data, θ is a categorical valued variable that tells us which candidate wins an election.

y ← data

θ → quantity of interest

Example 1.4. y is a sequence of of binary values that record whether a given day is rainy or not. θ may be taken to represent the frequency of rainy days, i.e., the cloudiness of a location. We may also want to predict if it is going to rain tomorrow or not (in this case, we may introduce another binary random variable to represent tomorrow's forecast).



Example 1.5. "Supervised learning". A less obvious example, y is the collection of data pair of the form (u, v), where v is the binary class label the represents the "class" of the corresponding u. θ is a mathematical quantity related to the classifier, a function which maps u to v that we wish to obtain on the basis of a training data set.

data

quantity of interest

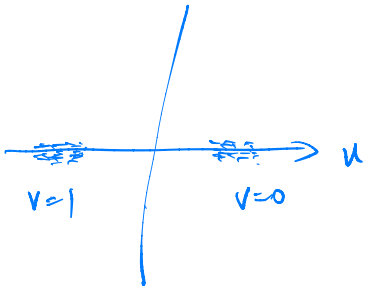
fail this class

when submit final project

how the time stamp of submitting final project u is related to pass or fail of this class. v

θ ⟸ relationship/mapping from u to v.

$\{(u_i, v_i)\}_{i=1}^{n}$



$\theta = \mathbb{1}$(u takes large values)

classify based on u.

Supervised: quantity of interest is about the θ relationship between two sets of variables.

(input, labels)

Example 1.6. ”Unsupervised learning”. A clustering problem involves subdividing a collection of data points represented by y into ”clusters”, which can be represented by θ.

data

quantity of interest

$y \in \mathbb{R}^2$, n datapts

$y = (y_1, \dots, y_n)$

$y_i \in \mathbb{R}^2$

$\theta = (\theta_1, \dots, \theta_n)$

θ_i

↳ cluster assignment of datapt i.

$\theta_i \in \{A, B, C\}$

① What are the $K=3$ major clusters in the data set?

The cluster center locations?

② For each data point, which cluster do they belong to?

Which data points belong to the same cluster?

Unsupervised: no labels

not supervised

data

quantity of interest

Example 1.7. "Who is who". y represents a collection of photos available in the Internet. θ represents identity of all individuals that appear in such photos.

$y = (y_1, \dots, y_n)$ n photos

$\theta = (\theta_1, \dots, \theta_n)$ θ_i: identity of individuals in photo i.

Example 1.8. "Generative AI". What are the y's? What are the θ's?

language model: (ChatGPT)

prompt → output.

① θ: relationship between prompt & output — supervised.

y: (prompt, output)

② θ: [prompt concatenated with output]

topics of sentences

or length of sentences generated by ChatGPT

or word distributions of ---

unsupervised

y: [prompt + output]

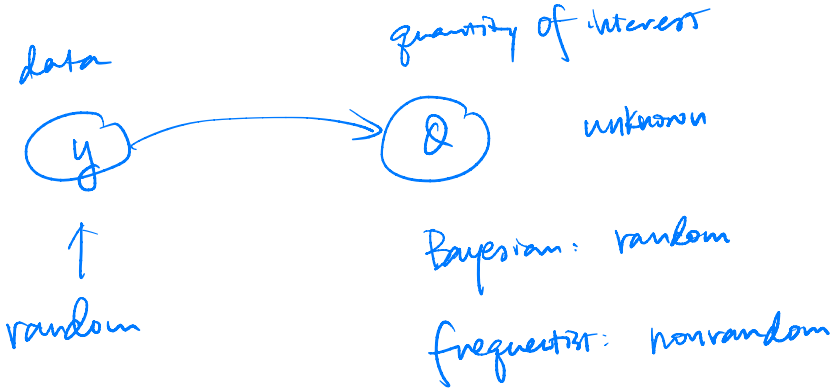
lots of sentences.

In practice and in our times, y has become increasingly complex, and so is the ambition of the data modeler and statistician, who want to infer increasingly complex quantity of interest θ.

For both frameworks of Bayesian and frequentist statistics, data y are always considered to be realizations of some random variable denoted by Y. [1]

The nature of the unknown θ is a different matter: frequentist methods treat θ as unknown but nonrandom.

Bayesian methods always assume θ to be random.



[1] In these notes, we will try to adhere to the convention that random variables are upper cases, unless denoted by Greek letters. The numerical value of the random variable, say Y, is denoted in lower cases, y.

The randomness of the unknown can be viewed as the most distinguishing feature of Bayesian statistics.

The ramifications are both deep and strong.

This course is an applied Bayesian analysis course, so we will not get into the deeper theoretical foundations of Bayesian statistics.

Instead, we focus on Bayesian methods and applications. Nonetheless, such ramifications of the Bayesian choice will be felt strongly.

1.2 Bayes' rule

The idealized form of Bayesian inference begins with a numerical formulation of the joint beliefs about y and θ, expressed in terms of probability distributions over Y and Θ. Here are the key ingredients:

1. For each numerical value $\theta \in \Theta$, *prior distribution* $p(\theta)$ describes our belief that θ represents the true population's characteristics.

2. For each $\theta \in \Theta$ and $y \in Y$, *sampling model* $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true.
   Once we obtain the data y, the last step is to update our beliefs about θ:

3. For each numerical value of $\theta \in \Theta$, *posterior distribution* $p(\theta|y)$ describes our belief that θ is the true value, having observed data set y.

The posterior distribution is obtained from prior distribution and sampling model via Bayes' rule

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

aka likelihood

Data $y \in \mathcal{Y}$

QoI $\theta \in \Theta$

joint $p(\theta, y)$

conditional: $p(y|\theta)$ sampling model

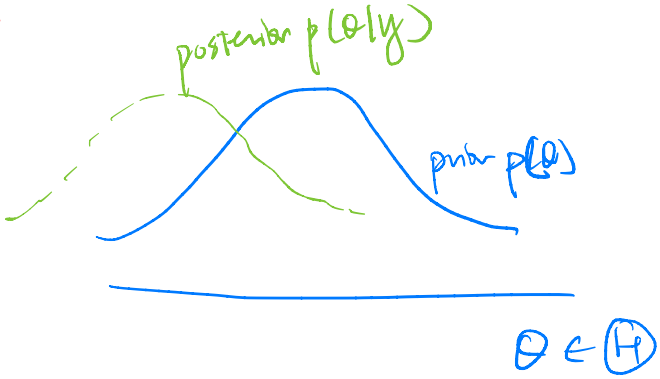
$p(\theta|y)$ posterior

chain rule

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

$$p(y) = \int p(y, \theta)\,d\theta$$

chain rule $= \int p(\theta)\,p(y|\theta)\,d\theta$



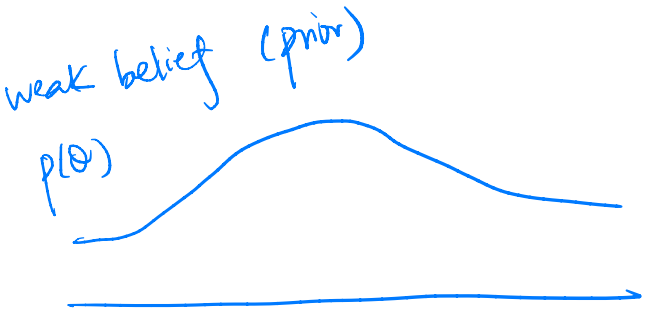
Note that Bayes' rule is a mathematical formula that allows one to "invert" arguments of conditional probabilities.

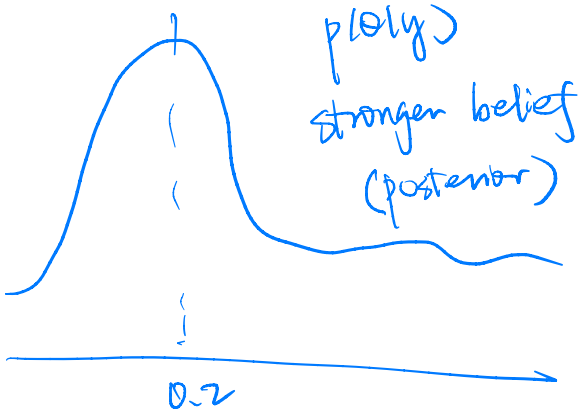
We have applied the Bayes formula for the purpose of statistical inference method named after its progenitor.

Implicit in the above description is a significant conceptual lift of Bayesian statistics: we treat the *a posteriori* "belief" about θ by adopting the conditional probability of θ given y.

The higher the value of the probability about a numerical value of θ, the stronger our belief about it. Athough "belief" may be a vague notion, probabilities and conditional probabilities are mathematically well-defined. Thus, we may speak of belief in a quantitively rigorous way.

Note also that Bayes' rule does not tell us what the truth θ should be; it tells us how our belief about θ changes after seeing new information.





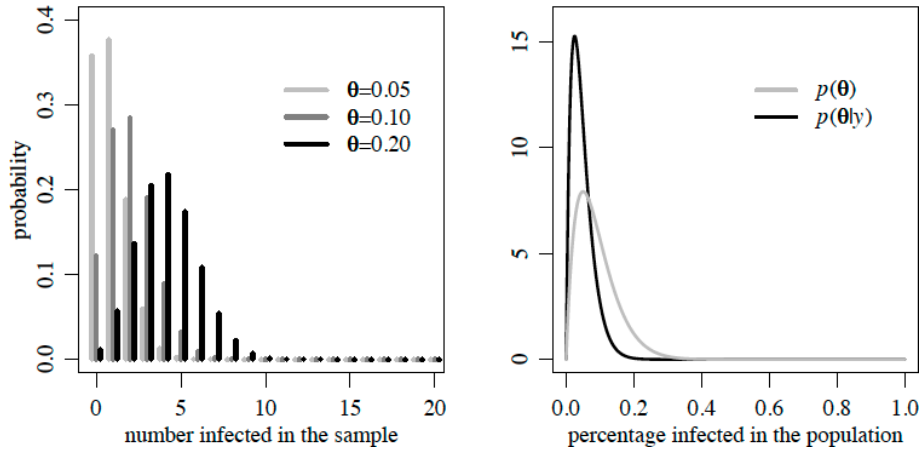


Figure 1.1: The plot on the left gives binomial(20, θ) distributions for three values of θ. The right side gives prior (gray) and posterior (black) densities of θ. This is Fig. 1.1 of PH.

1.3 Example: estimating the probability of a rare event

Continue on Example 1.1. Of interest is θ, the fraction of infected individuals in the city, so $\theta \in [0, 1]$. Data is the number of infected individuals out of a 20 sample. So $y \in \{0, \ldots, 20\}$.

Bay $p(\theta)$

$p(y|\theta)$ $p(y|\theta=0.05)$

$p(y|\theta=0.09)$

$p(y|\theta=0.2)$

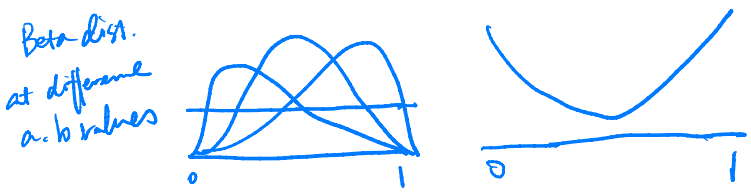
We need a sampling model. A reasonable choice is [why?]

$$Y|\theta \sim \text{binomial}(20, \theta)$$

$P(Y=y|\theta) = \binom{20}{y}\theta^y(1-\theta)^{20-y}$. In particular, $P(Y=0|\theta)=(1-\theta)^{20}$

To get a sense of this probability $P(Y = 0|\theta = 0.05) = 0.95^{20} \approx 0.36$. If $\theta = 0.1$ or $\theta = 0.2$, this number woud be 0.12 or 0.01, respectively.

* always plug in numbers into our sampling model to see if it makes sense / suitable.



Next, a prior is specified. A common choice is the beta distribution [why?] $\theta \sim$ beta(a, b). There are two parameters $a, b > 0$ that we need to set. But how?

The expectation under the beta prior is $a/(a + b)$. The mode of the beta prior is $(a - 1)/(a - 1 + b - 1)$. Previous studies from various parts of the country indicate that the infection rate in comparable cities ranges from about 0.05 to 0.20, with an average prevalence of 0.10. This suggests us to take $a = 2$, $b = 20$:

$$\theta \sim \text{beta}(2, 20).$$

This prior specification yields the following choice about the prior distribution:

$$\begin{aligned} \text{E}[\theta] &= 0.09 \\ \text{mode}[\theta] &= 0.05 \\ \Pr(\theta < 0.10) &= 0.64 \\ \Pr(0.05 < \theta < 0.20) &= 0.66. \end{aligned}$$

Beta(2, 20) roughly faithful to our prior knowledge.

You may still find reasons to be uncomfortable with this particular choice of prior parameters (what?), but we will get to them. Let us now apply the Bayes rule that enables one to go from the prior to the posterior distribution.

$p(\theta) \equiv \text{Beta}(\theta \mid a=2, b=20)$ prior

$p(y \mid \theta) \equiv \text{Binomial}(20, \theta)$ sampling model / likelihood

Bayes formula to obtain posterior $p(\theta \mid y)$

From prior to posterior By an application of the Bayes rule, we will find that: if $Y|\theta) \sim$ binomial(n, θ), $\theta \sim$ beta(a, b) then the conditional distribution is again a beta: $\theta|Y = y \sim$ beta$(a + y, b + n - y)$.

This is just an example of a general structural property called as "conjugacy" (the beta prior is conjugate to the binomial likelihood) that is widely exploited in Bayesian computation. We will study this property systematically in later lectures.

Suppose that in our specific study, we observed that in fact $Y = 0$, i.e., none of the sampled individuals was infected. [What do we make of this?] The posterior distribution of θ is therefore

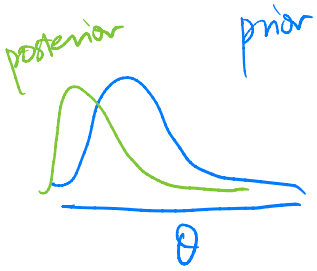
$$\theta|\{Y = 0\} \sim \text{beta}(2, 40).$$

Observe the *change* in shape from the prior distribution to the posterior distribution under the (new) observation $y = 0$ in Fig. 1.1: the mass of the posterior is shifted toward zero. This reflects the consequence of the "Bayes update"; by contrast a simple-minded approach is to set $\theta = 0$ in the presence of $y = 0$. The posterior is also more "peaked" than the prior. This reflects a general phenomenon: as more data are observed, our belief about θ becomes more concentrated, even if we start out with a prior belief that is more "defuse". In other words, the more data are observed, the less influential the role of the prior. This is a desirable property.

More quantitatively on this transformation:

$$\begin{aligned} \text{E}[\theta|Y = 0] &= 0.048 \\ \text{mode}[\theta|Y = 0] &= 0.025 \\ \Pr(\theta < 0.10) &= 0.93 \end{aligned}$$

In particular, we may say: our posterior belief in the presence of the observation that $\theta < 0.1$ is pretty high (> 0.93). How sensitive is this conclusion based on our prior specification?



Sensitivity analysis The Bayes update enables us to go from a beta(a, b) prior to a beta posterior, namely,

$$\text{beta}(a + y,\ b + n - y),$$

whose parameter incorporates the impact of the observed data y. In particular, we go from the prior mean $\theta_0 := a/(a + b)$ to the posterior mean

$$E(\theta \mid Y = y) = \frac{a+y}{(a+y)+(b+n-y)} = \frac{a+y}{a+b+n}$$

$$= \frac{a+b}{a+b+n}\left(\frac{a}{a+b}\right) + \frac{n}{a+b+n}\left(\frac{y}{n}\right)$$

$$= \frac{w}{w+n}\cdot\theta_0 + \frac{n}{w+n}\,\bar{y}$$

Here, we denote $w = a + b$. The above formula captures nicely the combined impacts of data, via the term $\bar{y} = y/n$, and prior knowledge, via the prior mean θ_0. The posterior mean represents a weighted average of the sample mean $\bar{y}$ and our prior guess θ_0.

We may view w as a parameter that represents our confidence in this prior guess. Note that the prior distribution may be expressed as beta($w\theta_0$, $w(1 - \theta_0)$). The posterior distribution is beta($w\theta_0 + y$, $w(1 - \theta_0) + n - y$).

- If we fix w, let us see the impact of data size. If sample size n tends to infinity, then posterior mean tends to the sample mean $\bar{y}$; the prior belief plays a vanishing role no matter how confident we are about it. However, when sample size n is small, the prior belief can be influential, and can be captured by the role of w.
- Let us fix n (to be relatively small). As $w \to 0$, as our confidence in the prior vanishes, the posterior mean converges to the data-driven sample mean $\bar{y}$. If $w \to \infty$, the opposite happens: the posterior mean tends toward that of the prior belief, θ_0; the observed data hardly matters any more.

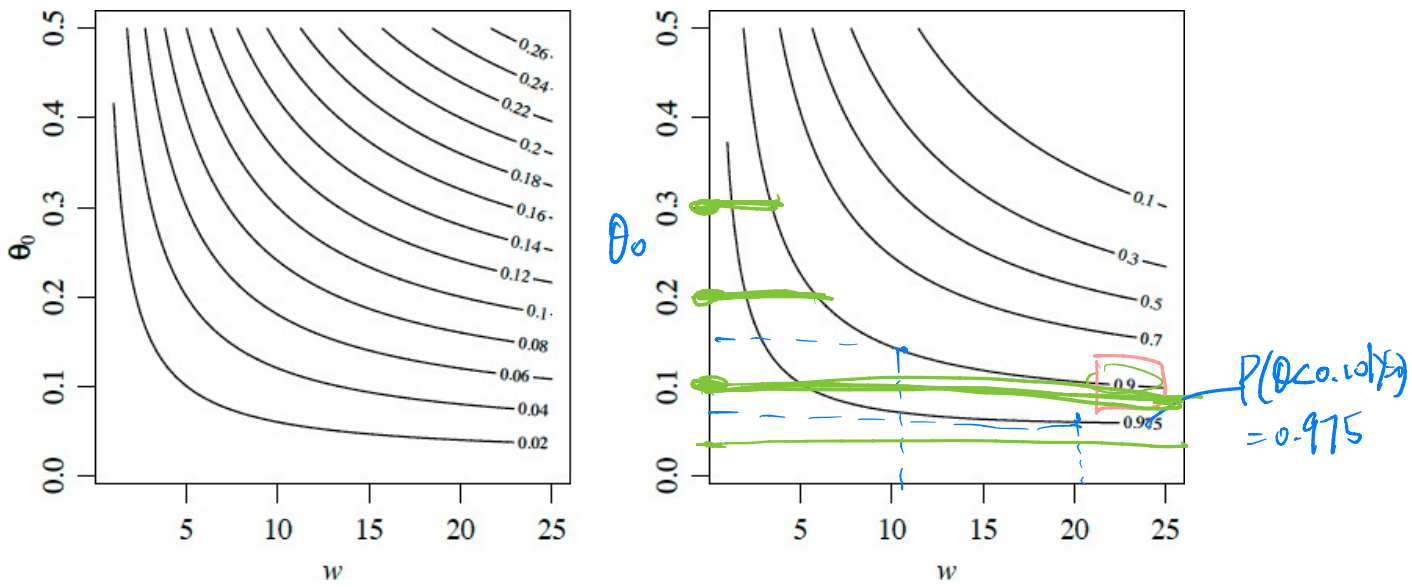


Figure 1.2: Posterior quantities under different beta prior specifications. The left and right hand panels give contours of $E[\theta|Y = 0] = w\theta_0/(w + 20)$ and $Pr[\theta < 0.10|Y = 0]$, respectively. (Fig. 1.2 of PH).

Fig. 1.2 gives a more detailed picture of the sensitivity of the prior specification. The left panel tells a general story: the prior specification can play a big role in our conclusion "after the fact". The sensitivity analysis allows us to be both honest and more confident in drawing our inference.

The confidence in our inference depends on the specific question that we ask about θ. Suppose that the city officials want to recommend a vaccine to the general public unless they were reasonably sure that the current infection rate was less than 10%. Then we may want to look at the right panel, which gives the contours of the posterior for $Pr[\theta < 0.10|Data]$.

- For chosen $\theta_0 \leq 0.1$, which is the average prevalence in other comparable cities from prior studies, we can be reasonably certain that the current infection rate is below 10% (with posterior probability above 90% for a large range of w).
- A higher degree of certainty, say 97.5%, is only achieved by people who already thought the infection rate was lower than the average of the other cities, e.g., if $\theta_0 < 0.05$.

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, and the 3blue1brown channel.