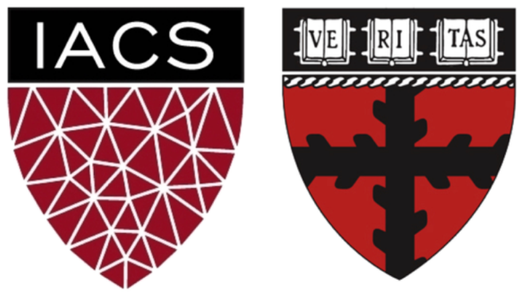
Lecture #3: Bayesian Modeling

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020


IACS
VE RI TAS

Outline

1. Review of the Method of Maximum Likelihood
2. Models for Real Data
3. The Beta-Binomial Model
4. Bayesian Modeling

Review of the Method of Maximum Likelihood

The Method of Maximum Likelihood

1. **(Model)** Assume observations from N number of ***independently and identically distributed*** outcomes, $Y_1, \ldots, Y_N$, with $Y_n \sim p(Y|\theta)$ and where θ is the set of parameters of the distribution $p(Y|\theta)$. The likelihood function is

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$

2. **(Inference)** We estimate θ using the ***maximium likelihood estimate***, defined as

$$\theta_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ \mathcal{L}(\theta) = \underset{\theta}{\text{argmax}} \prod_{n=1}^{N} p(y_n|\theta)$$

3. **(Inference Method)**

- *Unconstrained Optimization:* if θ doesn't need to satisfy any special property, then it's as simple as setting the ***gradient*** of the likelihood equal to zero and solving! **(Except it's not that simple!)**

- *Constrained Optimization*: if θ needs to satisfy special properties, then it's as simple as setting the ***Lagrangian*** of the likelihood and the constraint equal to zero and solving! **(Except it's not that simple!)**

Evaluating the MLE

If we have the true parameters θ, we can compute the ***Mean Squared Error***:

$$\mathrm{MSE}_\theta = \mathbb{E}_{Y^\theta}\left[(\theta_{\mathrm{MLE}}\left(Y^\theta\right) - \theta)^2\right]$$

If we don't have the true parameters θ, we can use θ_{MLE} ***predict*** or ***simulate*** data and compare it with the observed data, i.e. sample

$$Y^{\theta_{\mathrm{MLE}}} \sim p(Y|\theta_{\mathrm{MLE}}),$$

compare $Y^{\theta_{\mathrm{MLE}}}$ and Y^θ.

Properties of The Maximum Likelihood Estimator

Why Choose MLE? Asymptotically, i.e. given an infinite number of samples, the MLE is

1. *Consistent:* θ_{MLE} approaches the true parameters θ.
2. *Unbiased:* The average θ_{MLE}, taken over many different samples of the data, is θ.
3. *Minimum Variance:* The MLE has the lowest variance of all unbiased estimators.

Why Not Choose MLE? When the sample size is "small", the MLE can be

1. *Overfitted:* The MLE can is sensitive to outliers in the data.
2. *Biased:* The average θ_{MLE}, taken over many different data samples, is not θ.
3. *Imprecise:* The MLE can have high variance.

What Other Estimators are There?

1. Method of Moments
2. Minimum Variance Unbiased Estimator
3. Regularized MLE

Note: the computation of (1) and (2) can be much more complex than MLE.

Is Bias Always Bad?

Recall that we can use the properties of expectations to decompose the mean squared error of the maximum likelihood estimator:

$$\text{MSE}_{\theta_{\text{MLE}}} = \underbrace{\mathbb{E}\left[(\theta_{\text{MLE}} - \mathbb{E}[\theta_{\text{MLE}}])^2\right]}_{\text{variance}} + \underbrace{(\mathbb{E}[\theta_{\text{MLE}}] - \theta)^2}_{\text{bias}}$$

Although we want an unbiased estimator, the above decomposition says that if the variance of the estimator is high our espected error will nonetheless be high.

The ***variance-bias trade-off*** refers to the phenomenon that, in many cases, when estimators have low bias they have high corresponding variance (and vice versa), and hence high MSE.

Limitations of MLE: Overfitting Under Scarcity of Data

Suppose that we have three observations from a Bernoulli distribution, $Ber(\theta)$: $\{H, H, H\}$. From what we've see before, the MLE of θ is

$$\theta_{\text{MLE}} = \frac{3}{3} = 1.$$

Is this a good estimate of the bias of the coin? What can we do to make this estimate better?

Limitations of MLE: Overfitting Under Scarcity of Data

Suppose that we have two YouTube videos with 4/5 likes and 3,500/5,000 likes respectively. We can model the probability that a viewer will like each video as two Bernoulli distribtutions, $Ber(\theta_1)$, $Ber(\theta_2)$, where θ_i is the "inherent" likeability of each video.

Again, we can compute the MLE of the Bernoulli parameters:

$$\theta_1 = 4/5 = 0.8, \;\; \theta_2 = 3,500/5,000 = 0.75.$$

It is fair to say that the second video is more likeable base on our estimates?

Regularization

We can prevent MLE from overfitting to the observations when training data is scarce by constraining it from unreasonable values.

Recall that in order to prevent the MLE solution for linear regression from learn unrealistically large slopes and intercepts, we add ℓ_2-regularization on the model parameters during training, essentially forcing the parameters to stay as small (close to zero) as they can be while still capturing the data.

Similarly, if want the MLE of the parameter θ of a Bernoulli distribution to avoid extreme values (1 and 0), we need to 'anchor' our estimation of θ to some 'reasonable' value:

$$\theta_{\text{MLE Reg.}} = \frac{\#\text{ positive outcome} + \alpha}{\#\text{ total trials} + \beta}$$

The fraction α/β expresses your notion of what is a reasonable looking probability.

But is regularization a principled way to perform inference (i.e. will it ruin the nice properties of MLE)? How do you choose the hyperparameters α, β in a principled manner?

Models for Real Data

Video Ranking

We can model the outcome, Y, of a user rating for a specific YouTube video as a Bernoulli distribution,

$$Y \sim Ber(\theta)$$

where θ is the probability that a user will like the video.

Given $Y_1, \ldots, Y_N$ identically and independently distributed outcomes with $Y_n \sim Ber(\theta)$. Denote the total number of likes by L. Then L can be modeled with a ***binomial*** distribution,

$$L \sim Bin(N, \theta).$$

Model Critique: What are the assumptions made in this model? Are they realistic?

Kidney Cancer Rates

Given a dataset with N number of US counties and the incidents of kidney cancer in each county, we can model the observed incidents of cancer, C_n, of the n-th county with a Poisson distribution,

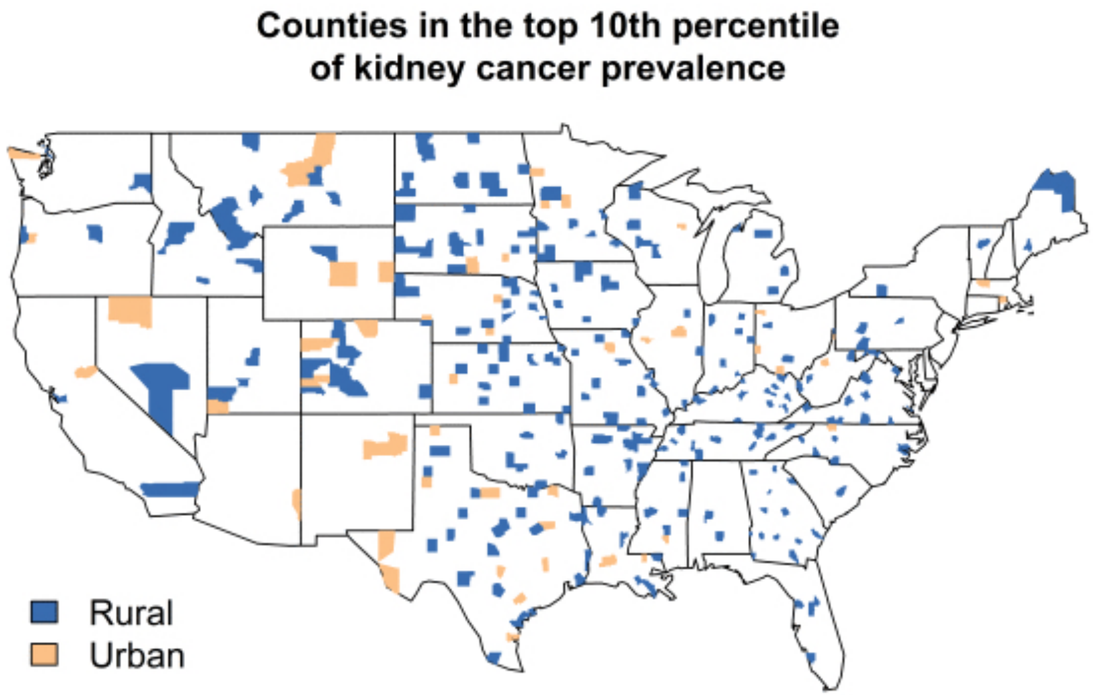
$$C_n \sim Poi(T_n \theta_n)$$

where T_n is the total population of the county and θ_n is the "true" cancer rate of the county.

Model Critique: What are the assumptions made in this model? Are they realistic?

Kidney Cancer Rates

The following is a visualization of the counties with the highest rates of kidney cancer (Gelman 1998). Is there any noticeable spatial pattern in these maps? Recall that the MLE of the rate of the Poisson distribution for each county is $\frac{C_n}{T_n}$.



Birth Weights

In our lab, we work with IVF clinics to build models for prediction and data analysis. One of our current tasks is to model the birth weights of the infants born in the clinic.

Naively, given observed birth weights $Y_1, \ldots, Y_N$, we can model each outcome Y_n with a normal distribution,

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

where μ is the average birth weight for this population and σ^2 is the population variance.

Model Critique: What are the assumptions made in this model? Are they realistic?

The Beta-Binomial Model

The Coin Toss Model: Revisited

Suppose that we have three observations from a Bernoulli distribution, $Ber(\theta)$: $\{H, H, H\}$. The MLE of θ is

$$\theta_{\text{MLE}} = \frac{3}{3} = 1.$$

This is a clear case of the MLE overfitting to the observed data.

Last time, you'd suggested tying the estimate to some fixed reasonable number, for example,

$$\theta_{\text{MLE Reg}} = \frac{H+1}{N+2},$$

and in general,

$$\theta_{\text{MLE Reg}} = \frac{H+\alpha}{N+\beta}.$$

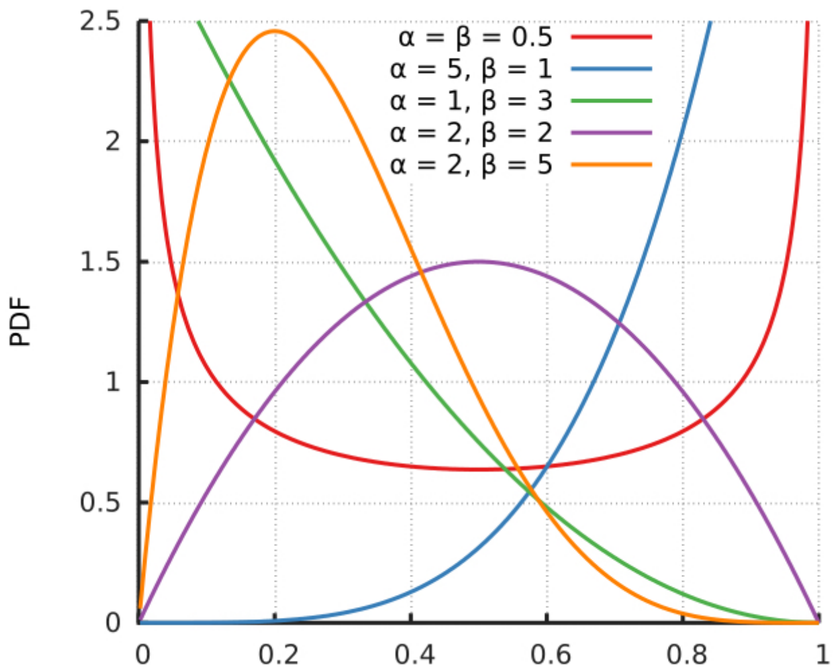
The terms α and β are called ***regularization terms*** .

Question: What is the effect of α, β on our estimate? What values should we choose for α, β?

Incoporating Prior Beliefs

Our choise of the regularization terms, α, β, depends on our prior beliefs about the coin. The way we chose to incorporate these beliefs doesn't indicate any uncertainty.

Alternatively, we can incorporate our prior belief about θ as a distribution, this is called the ***prior distribution***. Since θ is a number between 0 and 1, a beta distribution is an appropriate choice.



The Beta-Binomial Model

A model that involves both a likelihood for the data and prior on the parameters in the likelihood is called a ***Bayesian model***.

Bayesian Model for Coin Flip

$$
\begin{aligned}
Y &\sim Bin(N, \theta) && \text{(Likelihood)} \\
\theta &\sim Beta(\alpha, \beta), && \text{(Prior)}
\end{aligned}
$$

where α, β are called ***hyperparameters*** of the model.

Now, computing the MLE no longer makes sense (since the MLE only considers the likelihood). Luckily, Bayes' Rule allows us to derive a distribution that considers both the prior and the likelihood:

$$
p(\theta|Y) = \frac{\overbrace{p(Y|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(Y)}_{\text{marginal data likelihood}}} = \frac{p(Y, \theta)}{\int p(Y, \theta) d\theta}
$$

The distribution $p(\theta|Y)$ is called the ***posterior***.

Posterior for the Beta-Binomial Model

In our case, the posterior is given by

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{\overbrace{\binom{N}{Y}\theta^Y(1-\theta)^{N-Y}}^{\text{binomial pdf}}\overbrace{\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{beta pdf}}}{p(Y)}$$

We can rewrite the posterior as

$$p(\theta|Y) = const * \theta^{(Y+\alpha)-1}(1-\theta)^{(N-Y+\beta)-1}$$

where $const = \frac{\binom{N}{Y}}{B(\alpha,\beta)p(Y)}$ must be the normalizing constant for $p(\theta|Y)$.

We recognize the posterior as a beta distribution, $Beta(Y+\alpha,\, N-Y+\beta)$! Can you see why?

Interpreting the Posterior: Bayesian Update

Rather than a point estimate for θ, we now have a posterior distribution, $p(\theta|Y)$, over θ. What does the posterior tell us about θ?

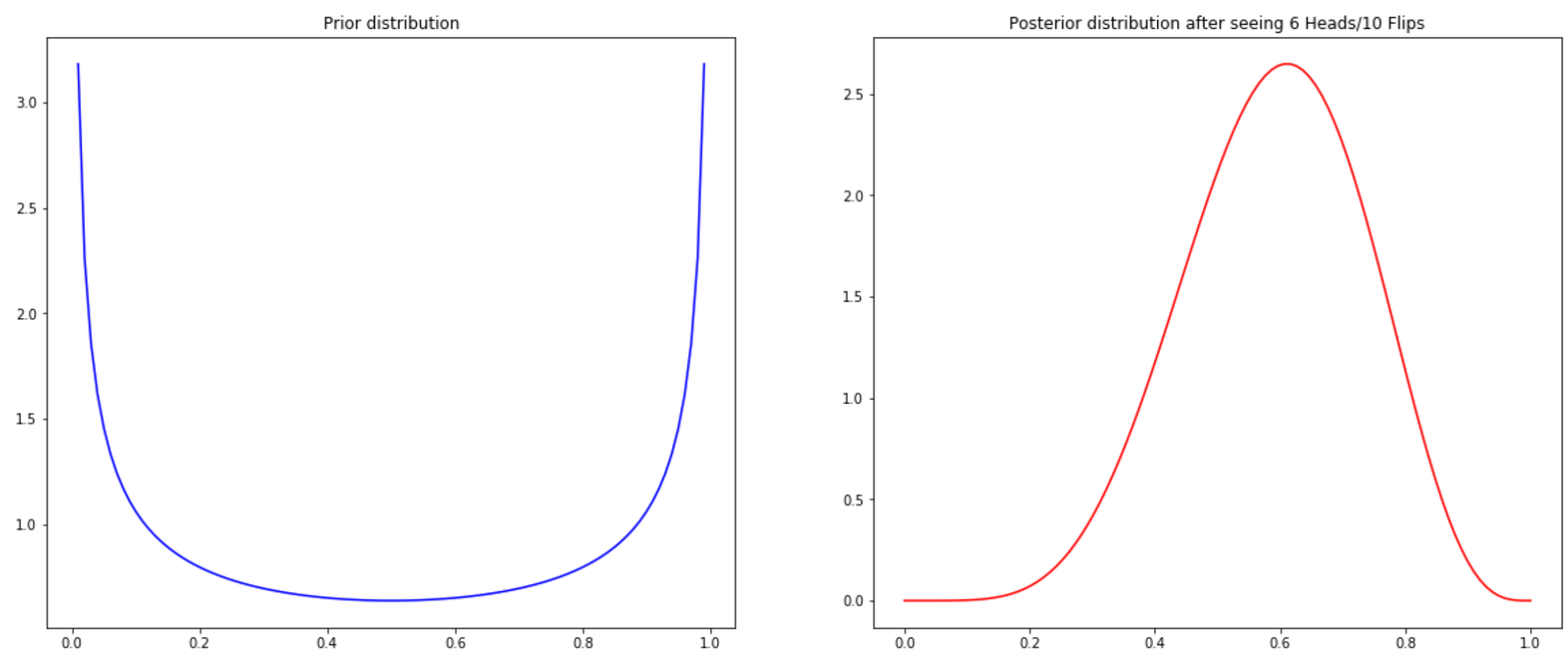
Since the prior distribution $p(\theta)$ encoded our beliefs about θ along with our uncertainty, it is natural to interpret the posterior as yet another **belief** about θ.

Since the posterior includes the likelihood, this belief has been **updated by the data**.

Simulation: Bayesian Update for the Coin Flip

What is the effect of the choice of the prior on the posterior? What is the effect of the number of observations, N, on the posterior?

In [123]:
```
fig, ax = plt.subplots(1, 2, figsize=(20, 8))
ax = plot_prior_posterior(ax, prior, posterior, H, N)
plt.show()
```



Making Predictions

If the posteriors we infer represent beliefs, how do we evaluate these beliefs?

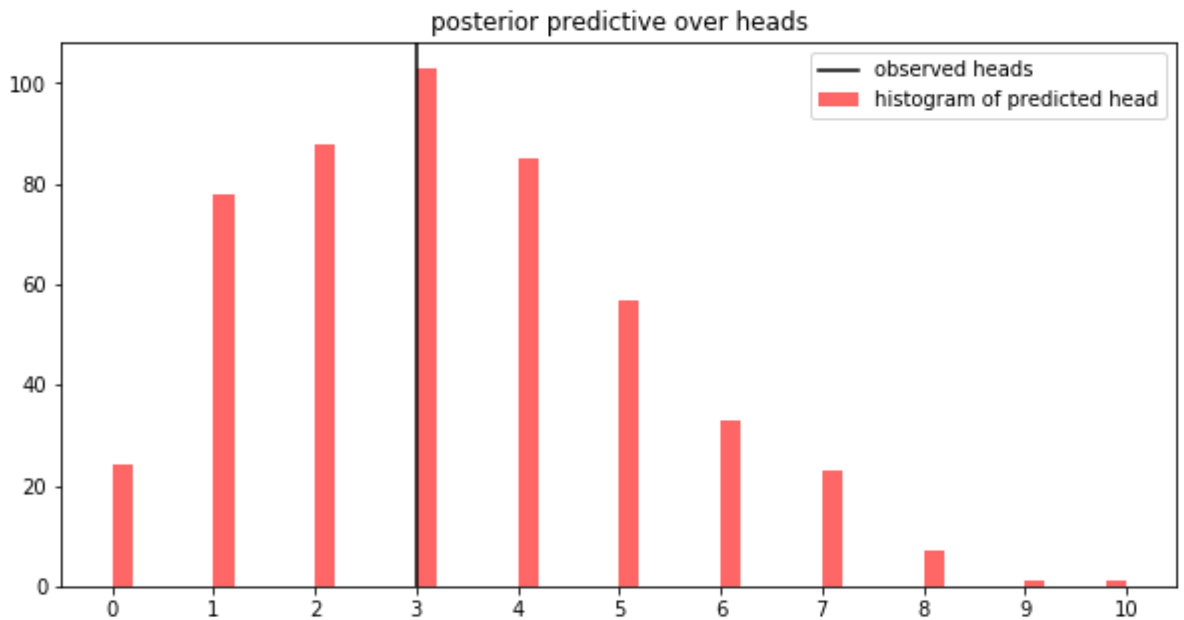
1. In the case that we the true parameter θ^{true}, we can check to see if the posterior assigns high likelihood to θ^{true}, and the certainty the posterior has about θ^{true}.

2. When we do not know θ^{true}, we can simulate data Y^{θ} using samples of θ from the posterior. We compare the distribution of simulated data, or ***posterior predictive***, to the observed data.

Simulation: Posterior Predictive for the Coin Flip

In [102]:
```
posterior_samples = np.random.beta(posterior_alpha, posterior_beta, size=500)
posterior_pred = [np.random.binomial(N, theta_sample, 1)[0] for theta_sample in
posterior_samples]

fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax = plot_posterior_predictive(ax, posterior_pred, H)
plt.show()
```



$\theta \sim p(\theta|Data)$

$y \sim p(y|\theta)$

Bayesian Modeling - A Summary

The Bayesian Modeling Process

In order to make statements about Y, the outcome, and θ, parameters of the distribution generating the data, we form the joint distribution over both variables and use the various marginals/conditional distributions to reason about Y and θ.

1. we form the ***joint distribution*** over both variables $p(Y, \theta) = p(Y|\theta)p(\theta)$.

2. we can condition on the observed outcome to make inferences about θ,
$$p(\theta|Y) = \frac{p(Y, \theta)}{p(Y)}$$
where $p(\theta|Y)$ is called the ***posterior distribution*** and $p(Y)$ is called the ***evidence***.

3. before any data is observed, we can simulate data by using our prior
$$p(Y^*) = \int_\Theta p(Y^*, \theta)d\theta = \int_\Theta p(Y^*|\theta)p(\theta)d\theta$$
where Y^* represents new data and $p(Y^*)$ is called the ***prior predictive***.
4. after observing data, we can simulate new data simliar to the observed data by using our posterior
$$p(Y^*|Y) = \int_\Theta p(Y^*, \theta|Y)d\theta = \int_\Theta p(Y^*|\theta)p(\theta|Y)d\theta$$
where Y^* represents new data and $p(Y^*|Y)$ is called the ***posterior predictive***.

Evaluating Bayesian Models

As we have seen in the Beta-Binomial model, we can simulate the posterior (and prior) predictive rather than compute them analytically. That is, you don't need to know the pdf of $p(Y^*|Y)$.

The posterior predictive can be represented by **samples** of predictions:

1. we sample values of θ_n from the posterior, $p(\theta|Y)$.
2. we sample an outcome Y_n from $p(Y|\theta_n)$ for each posterior sample θ_n.

The set Y_n we obtain emprirically represents the posterior predictve distribution $p(Y^*|Y)$.

Where do Priors Come From?

Hopefully you've noticed a key property of the priors we chose:

> *All the priors combined with the likelihood to form a distribution we recognize! Specifically, the posterior distribution is of the same type as the prior!*

These priors are called ***conjugate priors*** for the corresponding likelihoods. This is a purely mathematical property.

Question: is it right to choose priors that are mathematically convenient? What is a good way to choose a prior? What if we "choose wrong"?