Lecture #11: Hierarchical Models

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020

IACS

VE RI TAS

Outline

1. Review of Statistical Modeling
2. Motivation for Hierarchical Models
3. Hierarchical Models and Empirical Bayes

Review of Statistical Modeling

What We Can Do So Far: Models

1. **(Likelihood Models with Observed Variables)** When we have observed data Y_{Obs}, we can model Y_{Obs} as a random variable $Y_{\text{Obs}} \sim p(Y|\theta)$ with a known distribution p.

    - if Y_{Obs} is a label, we can model it as a *Categorical* or *Bernoulli* variable
    - if Y_{Obs} is a count, we can model it as a *Binomial*, *Multinomial* or *Poisson*
    - if Y_{Obs} is continuous, we can model it as a *Gaussian*, *Exponential*, *Dirichlet* etc

2. **(Likelihood Models with Latent Variables)** When we also have unobserved data Z_{Latent}, we can model Z_{Latent} and Y_{Obs} jointly $p(Y_{\text{Obs}}, Z_{\text{Latent}}|\theta)$.

3. **(Bayesian Models)** When we are being Bayesian, we *assume* a prior for θ, encoding our knowledge and uncertainty about θ. We model parameters and data jointly $p(Y_{\text{Obs}}, \theta)$ or $p(Y_{\text{Obs}}, Z_{\text{Latent}}, \theta)$.

What We Can Do So Far: Inference

We can make statements about θ by performing:

I. *Maximum Likelihood Estimation:* for likelihood models, we compute a fixed value θ_{MLE} that maximizes the likelihood of the observed data Y.

II. *Bayesian Inference:* for Bayesian models, we compute the posterior distribution $p(\theta|Y)$.

We choose an ***inference algorithm or method*** to perform inference:

I. *Maximum Likelihood Estimation:*
A. For *models with observed variables*, we *analytically* solve an unconstrained or contrained optimization problem to obtain θ_{MLE}.
B. For *latent variable models*, we use ***expectation maximization*** to approximately find θ_{MLE}.

II. *Bayesian Inference:*
A. If the prior and likelihood are ***conjugate***, *analytically* derive the posterior distribution
B. If the posterior distribution does not have a known form, sample from it using a ***sampler***.
C. If the posterior distribution does not have a known form, approximate it using ***variational inference***.

What Can We Not Do?

1. (***Regression Models***) We don't have any models where some observed variables Y_{Obs} depend on other observed variables X_{Obs}, i.e. none of our models have covariates that we condition on.

2. (***Gradient Descent Methods***) Analytically optimizing the log-likelihood is not always possible, and when possible it is extremely annoying to do by hand. *Can we find a way to black-box optimize any objective function?*

3. (***Hamiltonian Monte Carlo***) MCMC samplers can be extremely inefficient in high-dimensions where the samplers struggle to find area of high mass in the target distribution. *Can we build a proposal distribution with an indicator of where the target distribution mass is located?*

4. (***Black-box Variational Inference etc***) Variational inference sounds like a great idea but maximizing the ELBO using coordinate ascent is an artisenal process that requires a massive amount of derivations per model. *Can we find an algorithm to compute/estimate the gradient of the ELBO that is model independent? Can we perform black-box variational inference?*

What Happens After Inference?

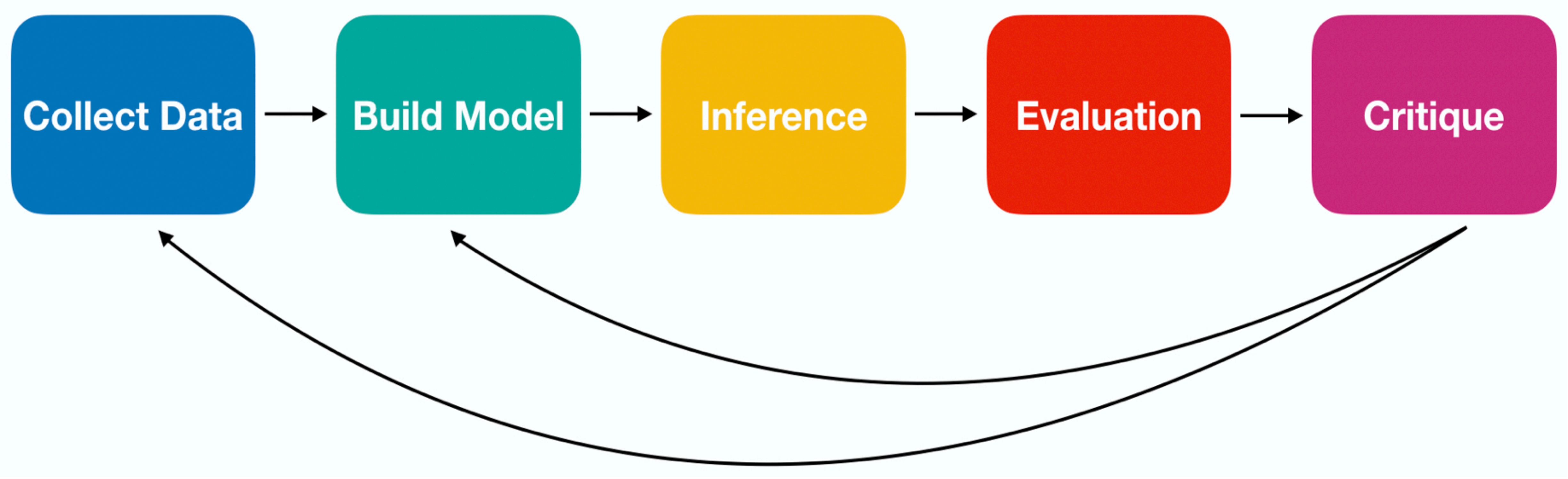
1. **(Predictive Evaluation)** In practice, we do not know the true model θ_{True}! Thus, θ_{MLE} and $p(\theta|Y)$ cannot be evaluated by comparison to θ_{True}.

   - ***Maximimum Likelihood Estimation:*** we compute θ_{MLE} on multiple bootstrap samples of the data; for each θ_{MLE} we sample $Y \sim p(Y|\theta_{\text{MLE}})$. We compare these samples with observed data Y_{Obs}.
   - ***Bayesian Inference:*** we sample θ's from the posterior, for each $\theta \sim p(\theta | Y^{\text{OBS}})$ sample $Y \sim p(Y|\theta)$. We compare these ***posterior predictive samples*** with the observed data Y_{Obs}.

2. **(Uncertainty Evaluation)** Before making decisions with real-life consequence based on your model, you should check the precision of your estimate or uncertainty of you model.

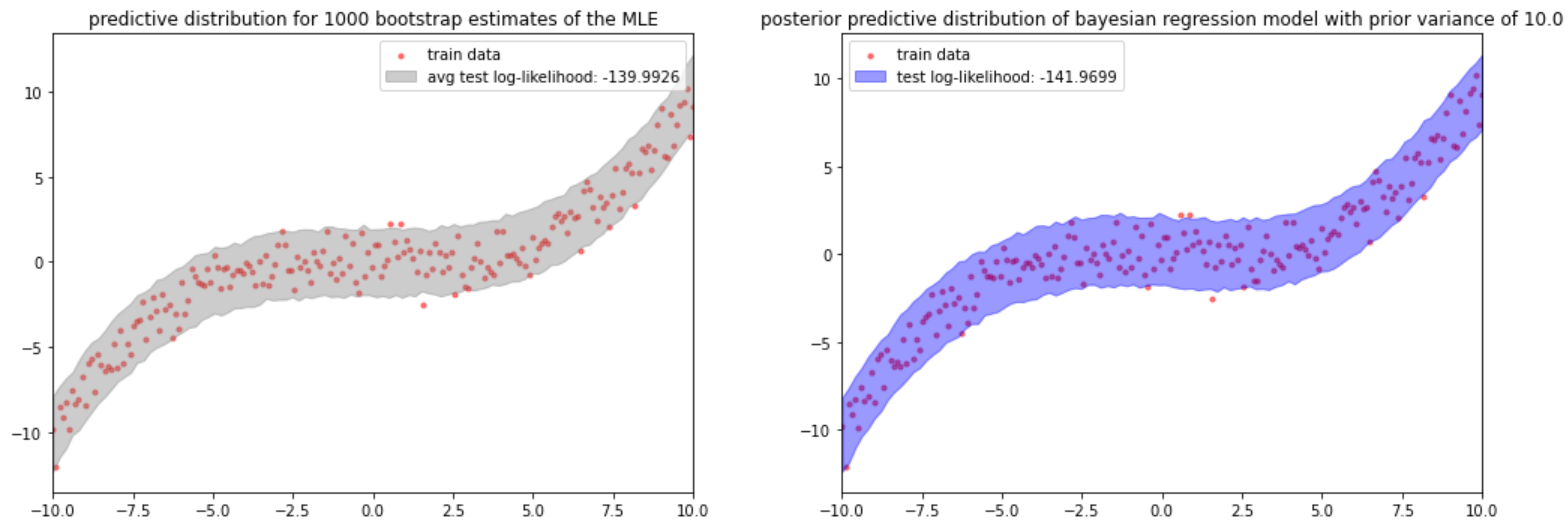
   - ***Maximimum Likelihood Estimation:*** repeat the MLE computation on many bootstrap samples of Y_{Obs}. Compute the confidence interval of θ and the predictive interval for Y. These intervals indicate *precision*.
   - ***Bayesian Inference:*** Compute credible intervals for the posterior $p(\theta|Y)$ and the predictive intervals of the posterior predictive. These intervals indicate *model uncertainty*.

The Modeling Process



Interpreting the Data Log-Likelihood

In [5]:
```
fig, ax = plt.subplots(1, 2, figsize=(15, 5))
prior_var = 10.
ax, log_likelihood_bayes = mle_vs_bayesian(x_2, y_2, ax, prior_var=prior_var)
plt.tight_layout()
plt.show()
```



Evaluating and Quantifying Uncertainty

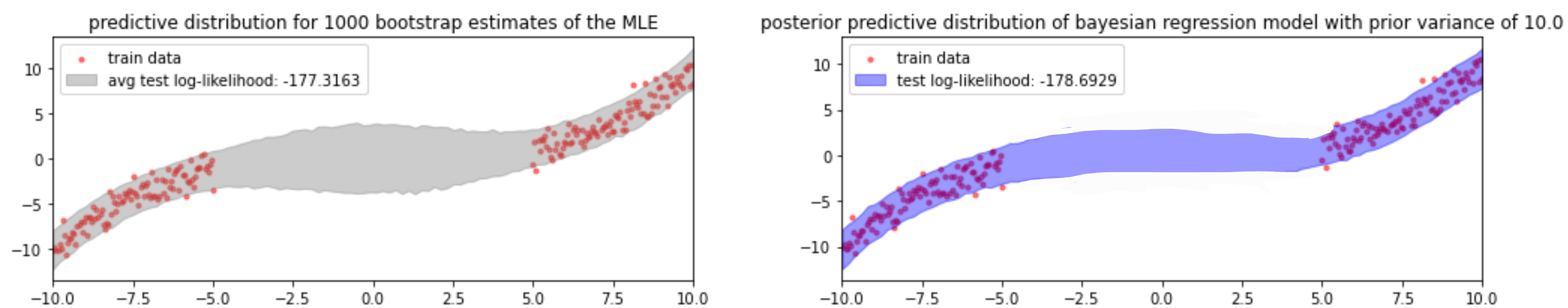
How do we know our model uncertainties (confidence intervals, credible intervals, predcitive intervals, posterior predictive intervals) are any good? What information about the data/model do we want our uncertainties to capture?

Epistemic Uncertainty: uncertainty due to small number of samples across all scenarios. This can be reduced by more samples!

Aleatoric Uncertainty: uncertainty due to the underlying randomness of the data generation process. This cannot be reduced no matter what.

Can we use log-likelihood of as a metric for the quality of predictive uncertainty?

In [6]:
```
fig, ax = plt.subplots(1, 2, figsize=(15, 3))
ax = mle_vs_bayesian(x_1, y_1, ax, prior_var=prior_var)
plt.tight_layout()
plt.show()
```



Motivation for Hierarchical Models

A Binomial Model for Movie Rankings

We model the number of likes Y_n received by the n-th movie as a binomial variable $Y_n|\theta_n \sim Bin(R_n, \theta_n)$, where R_n is the number of times the n-th movies was rated and θ_n is the "likeability" of the movie.

In [11]:

```
#Print results of ranking
print('Top 10 Movies')
print('***************************')
for movie, likes, total_ratings, likable in top_movies:
    print (movie, ':', likable, '({}/{})'.format(likes, total_ratings))
```

```
Top 10 Movies
****************************
French Twist (Gazon maudit) (1995) : 1.0 (2.0/2.0)
Exotica (1994) : 1.0 (2.0/2.0)
Three Colors: Red (1994) : 1.0 (12.0/12.0)
Three Colors: White (1994) : 1.0 (8.0/8.0)
Shawshank Redemption, The (1994) : 1.0 (39.0/39.0)
Brother Minister: The Assassination of Malcolm X (1994) : 1.0 (1.0/1.0)
Carlito's Way (1993) : 1.0 (4.0/4.0)
Robert A. Heinlein's The Puppet Masters (1994) : 1.0 (2.0/2.0)
Horseman on the Roof, The (Hussard sur le toit, Le) (1995) : 1.0 (2.0/2.0)
Wallace & Gromit: The Best of Aardman Animation (1996) : 1.0 (6.0/6.0)
```

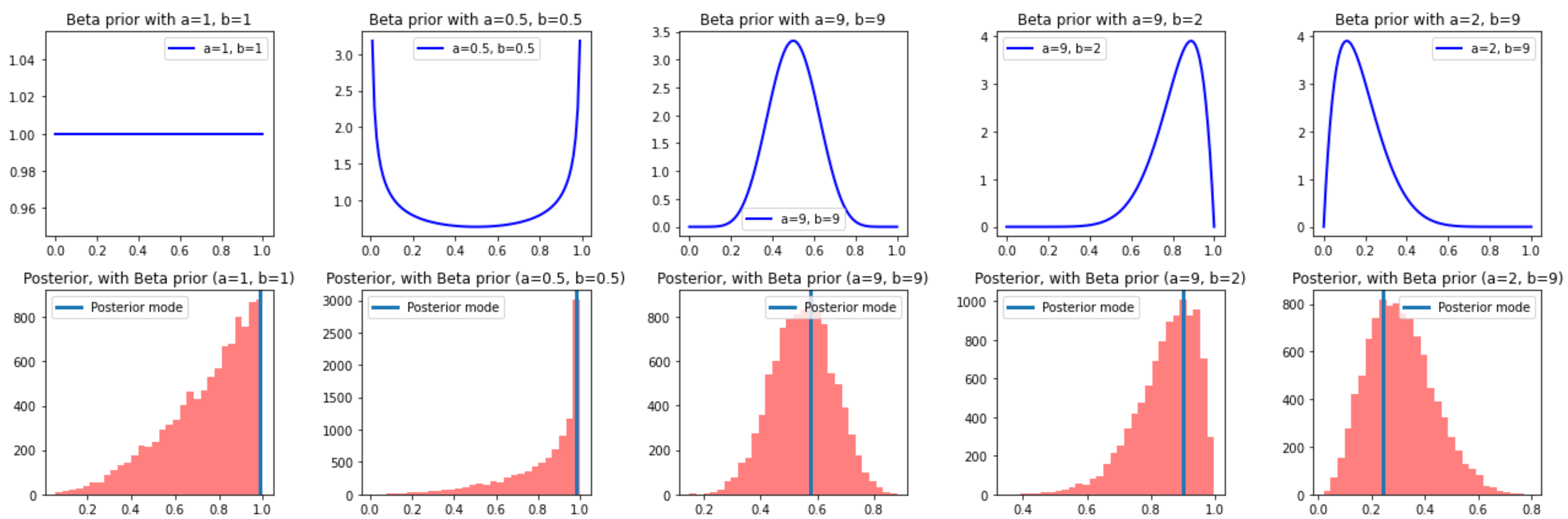
A Beta-Binomial Model for Movie Rankings

We model the number of likes Y_n received by the n-th movie as a binomial variable $Y_n|\theta_n \sim Bin(R_n, \theta_n)$, we model our prior beliefs and uncertainty about θ using a beta distribution $\theta_n \sim Beta(\alpha, \beta)$.

In [13]:

```
print('{}: {} ({}/{})'.format(movie_name, likability, likes, total_ratings))
fig, ax = plt.subplots(2, n, figsize=(18, 6))
ax = plot_priors_with_posteriors(ax, beta_shapes, likes, total_ratings, samples)
plt.tight_layout()
plt.show()
```

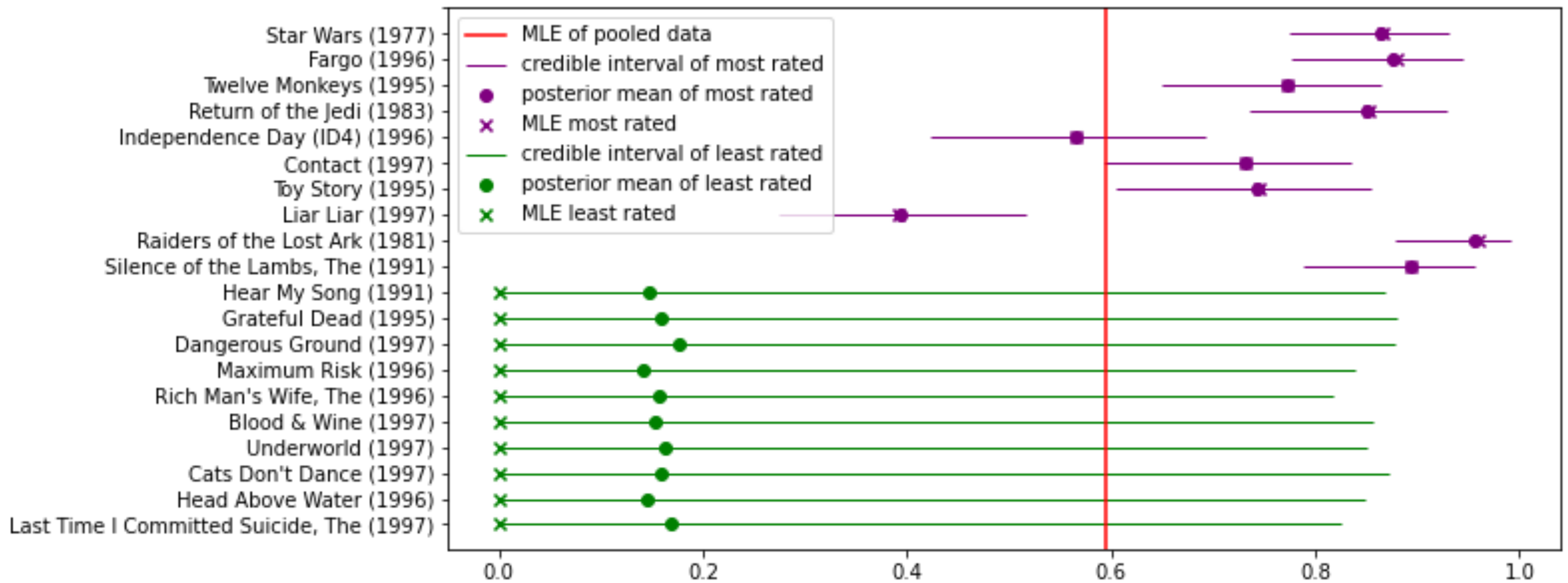
```
French Twist (Gazon maudit) (1995): 1.0 (2.0/2.0)
```



Credible Intervals for Movies with the Most and the Least Number of Ratings

In [15]:

```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax = plot_credible_intervals(a, b, ax)
plt.show()
```



Emprirical Bayes (ML-II) For the Beta-Binomial Model

Since the prior has a significant impact on the posterior when the number of ratings is small, we want to choose a prior that is appropriate for the data.

Idea: choose the hyperparameters α, β for the beta prior such that the expected likelihood of the data, over $\theta_n \sim Beta(\alpha, \beta)$, is maximized:

$$p(Y_1, \ldots, Y_N | \alpha, \beta) = \prod_{n=1}^{N} \int_0^1 Bin(Y_n | R_n, \theta_n) Beta(\theta_n | \alpha, \beta) d\theta_n = \prod_{n=1}^{N} \binom{R_n}{Y_n} \frac{B(\alpha + Y_n, \beta + R_n - Y_n)}{B(\alpha, \beta)}$$

where B is the beta function. The marginal likelihood of the data $p(Y_1, \ldots, Y_N | a, b)$ is called ***evidence***.

This method of choosing the hyperparameters of the prior based on the data is called ***empirical Bayes*** or ***type-II maximum likelihood***.

Question: doens't this violate the principle of choosing the prior independent of the data?

Method of Moments for Empirical Bayes (ML-II)

Since each marginal $p(Y_n|\alpha, \beta)$ is a Beta-Binomial distribution, we know its first two moments:

$$
\begin{aligned}
\mathbb{E}\left[Y_n\right] &= R_n \frac{\alpha}{\alpha+\beta} \\
\mathrm{Var}\left[Y_n\right] &= \frac{R_n \alpha\beta}{(\alpha+\beta)^2} \frac{\alpha+\beta+R_n}{\alpha+\beta+1}
\end{aligned}
$$

Now we can make the simplyfying approximations that the Y_n's are iid data from the *same* binomial, i.e. all have the same moments as above. Then we can use empirical moments to approximate the theoretical moments and solve for α, β:

$$
\begin{aligned}
\widehat{\mathbb{E}}\left[\frac{Y_n}{R_n}\right] &= \frac{\alpha}{\alpha+\beta} \\
\widehat{\mathrm{Var}}\left[\frac{Y_n}{R_n}\right] &= \frac{\alpha\beta}{\overline{R}_n(\alpha+\beta)^2} \frac{\alpha+\beta+\overline{R}_n}{\alpha+\beta+1}
\end{aligned}
$$

where $\widehat{\mathbb{E}}$ is sample mean and $\widehat{\mathrm{Var}}$ is sample variance and $\overline{R}_n$ is the average total number of ratings.

Processing math: 100%

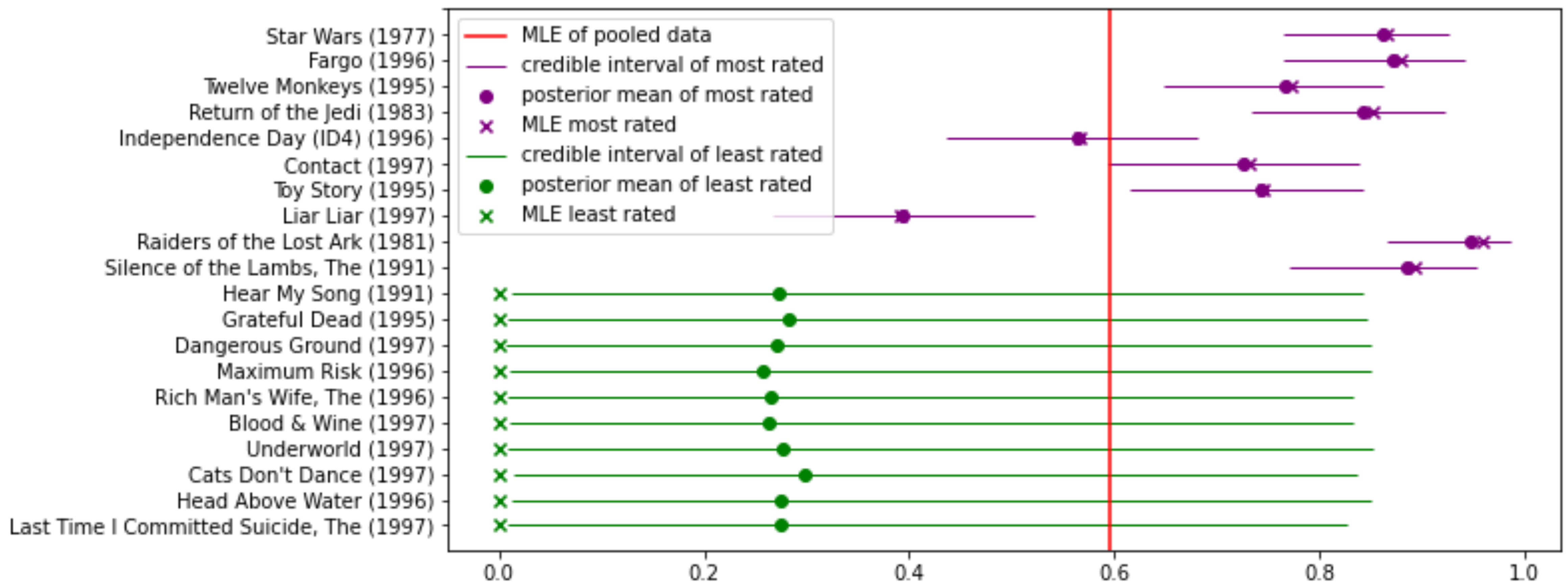
Empirical Bayes and Shrinkage

Computing the hyperparameters α, β of the beta prior on θ from the data, allows the ratings rich movies to influence the prior of ratings poor movies, since all movies contribute to the empirical Bayes estimate.

As a result, the estimates from ratings poor movies tend to ***shrink*** towards the population mean more so than ratings rich movies.

In [17]:

```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax = plot_credible_intervals(alpha_eb, beta_eb, ax)
plt.show()
```



Hierarchical Models and Empirical Bayes

A Hierarchical Model for Movie Rankings

We model the number of likes Y_n received by the n-th movie as a binomial variable $Y_n|\theta_n \sim Bin(R_n, \theta_n)$, we model our prior beliefs and uncertainty about θ using a beta distribution $\theta_n \sim Beta(\alpha, \beta)$; finally, we model our uncertainty about α, β using uniform distributions $\alpha, \beta \sim U(0.5, 100)$:

$$\begin{aligned} \alpha, \beta &\sim U(0.5, 100) \\ \theta_n|\alpha, \beta &\sim Beta(\alpha, \beta) \\ Y_n|\theta_n &\sim Bin(R_n, \theta_n) \end{aligned}$$

This is an example of a ***hierarchical model*** -- a model with multiple layers of unknown variables.

There are overlaps between hierarchical models and latent variable models. Generally, we want the hiearchy in a ***hierarchical model*** to express scientifically meaningful conditional relationships. In ***latent variable models*** we want the latent variable to represent unknown aspects of the data rather than unknown parameters of our model.

Point Estimate Approximations of Inference in Hierachical Models (MAP-II)

The posterior of the hierarchical model for movie ratings is $p(\alpha, \beta, \theta_1, \ldots, \theta_N | Y_1, \ldots, Y_N)$, but since we know how θ_n is conditioned on α, β, it is often easier to marginalize out θ_n and work with $p(\alpha, \beta | Y_1, \ldots, Y_N)$.

The central idea is that by infering the posterior $p(\alpha, \beta | Y_1, \ldots, Y_N)$, α and β are influenced by the entire data set and thus ratings poor movies can ***borrow statistical strength*** from ratings rich movies through the way θ_n depends on α, β.

However, performing full Bayesian inference on hierarchical models can be difficult. Thus, we can make a point estimate approximation of $p(\alpha, \beta | Y_1, \ldots, Y_N)$:

$$\alpha^*, \beta^* = \mathrm{argmax}_{\alpha,\beta} p(\alpha, \beta | Y_1, \ldots, Y_N).$$

When we perform the usual Bayesian inference on $p(Y_n | \theta_n) p(\theta_n | \alpha^*, \beta^*)$, this is called the ***type-II MAP method***.

But when α, β are uniform random varibles, the above becomes:

$$\alpha^*, \beta^* = \mathrm{argmax}_{\alpha,\beta} p(Y_1, \ldots, Y_N | \alpha, \beta),$$

When we perform the usual Bayesian inference on $p(Y_n | \theta_n) p(\theta_n | \alpha^*, \beta^*)$, this is just our empirical Bayes or type-II MLE method!