STAT 237
Cross-Validation

May 9-13, 2022

Colin Reimer Dawson

Outline

Outline

Estimating Marginal Likelihood Via Sampling

Marginal Likelihood as Expected Value

- Normalized marginal likelihoods can be intractable to compute analytically
- However, as with many things, they are **expected values of functions of the parameters** of a model

$$
\begin{aligned}
p(\mathbf{y} \mid m) &= \int p(\mathbf{y}, \theta \mid m)\ d\theta \\
&= \int p(\mathbf{y} \mid \theta, m) p(\theta \mid m)\ d\theta \\
&= \mathbb{E}\left[p(\mathbf{y} \mid \theta, m)\right]
\end{aligned}
$$

where the expected value in this case is with respect to the **prior distribution** of θ (note that $\mathbf{y}$ is constant, because we are **using the data** to calculate likelihoods for each parameter value)

Marginal Likelihood as Expected Value

$$p(\mathbf{y} \mid m) = \mathbb{E}\left[p(\mathbf{y} \mid \theta, m)\right]$$

- We can **approximate** this expected value using samples, $\theta^{(1)}, \ldots \theta^{(T)}$ sampled from from the **prior**, $p(\theta \mid m)$ (as with prior predictive checks):

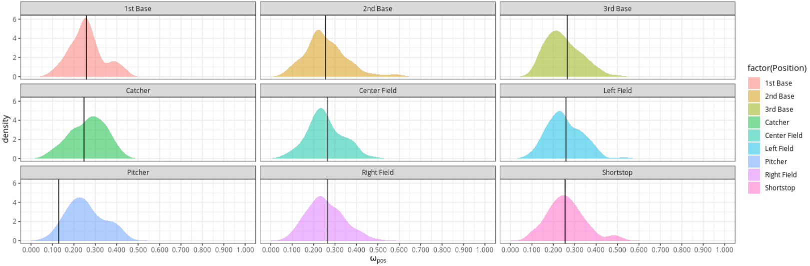
$$\mathbb{E}\left[p(\mathbf{y} \mid \theta, m)\right] \approx \frac{1}{T} \sum_{t=1}^{T} g(\theta^{(t)})$$

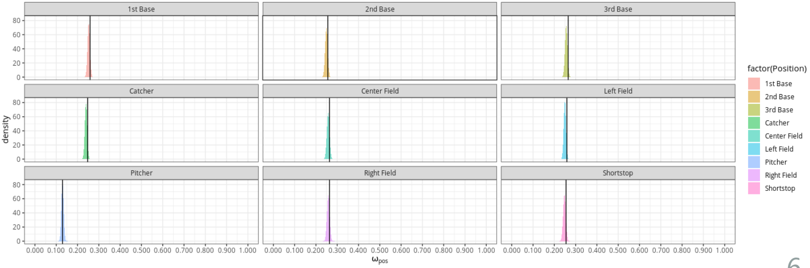
where, in this case,

$$g(\theta^{(t)}) = p(\mathbf{y} \mid \theta^{(t)}, m)$$

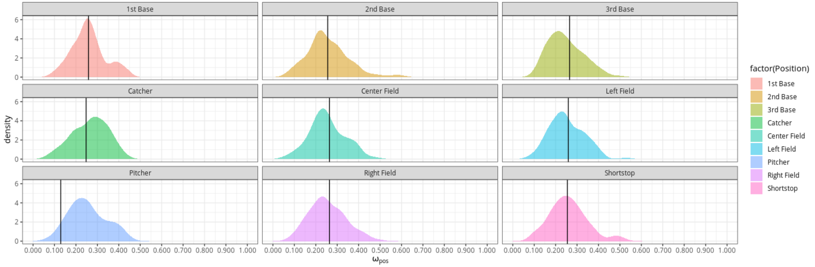
Inefficiency of Prior Sampling

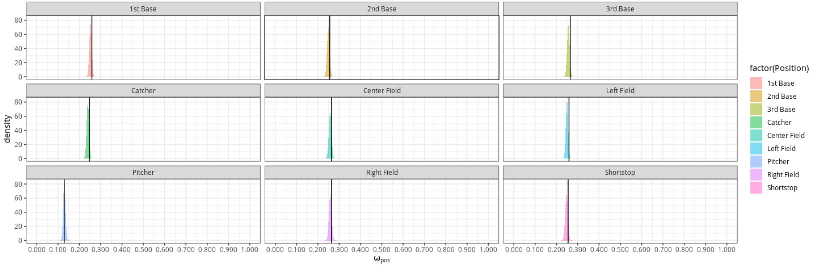
Compare the priors on ω_{pos} from our batting average model (top) to the corresponding posteriors (bottom)





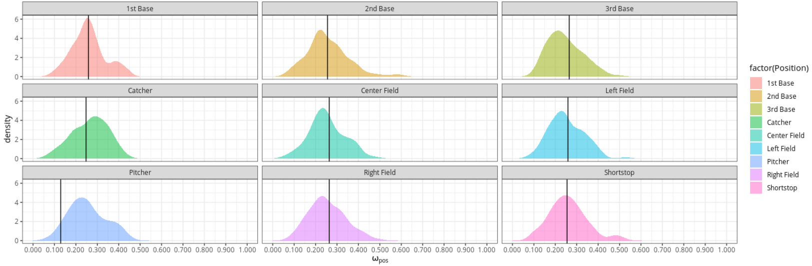
Inefficiency of Prior Sampling

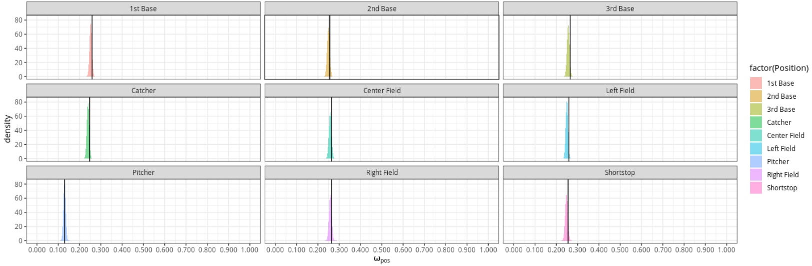




Notice how much more concentrated the posterior is compared to the prior

Inefficiency of Prior Sampling





What does that tell us about the likelihood, $p(\mathbf{y} \mid \omega)$?

Inefficiency of Prior Sampling

- With a decent amount of data about each position, only a relatively small region of the parameter space is remotely consistent with the data

Inefficiency of Prior Sampling

- With a decent amount of data about each position, only a relatively small region of the parameter space is remotely consistent with the data
- In other words, the likelihood drops off by orders of magnitude away from the parameter values near the maximum likelihood settings

Inefficiency of Prior Sampling

- With a decent amount of data about each position, only a relatively small region of the parameter space is remotely consistent with the data
- In other words, the likelihood drops off by orders of magnitude away from the parameter values near the maximum likelihood settings
- This is why the posterior has the vast majority of its mass there

Inefficiency of Prior Sampling

- With a decent amount of data about each position, only a relatively small region of the parameter space is remotely consistent with the data
- In other words, the likelihood drops off by orders of magnitude away from the parameter values near the maximum likelihood settings
- This is why the posterior has the vast majority of its mass there
- What does this mean for our sampling-based estimate of the **marginal** likelihood?

Inefficiency of Prior Sampling

- With a decent amount of data about each position, only a relatively small region of the parameter space is remotely consistent with the data
- In other words, the likelihood drops off by orders of magnitude away from the parameter values near the maximum likelihood settings
- This is why the posterior has the vast majority of its mass there
- What does this mean for our sampling-based estimate of the **marginal** likelihood?
- As a prior-weighted average of the likelihood, the marginal likelihood is heavily influenced by the proportion of samples that wind up in this "high likelihood" region, with the other samples contributing essentially zeroes to the average

Inefficiency of Prior Sampling

- As a prior-weighted average of the likelihood, the marginal likelihood is heavily influenced by the proportion of samples that wind up in this "high likelihood" region, with the other samples contributing essentially zeroes to the average

Inefficiency of Prior Sampling

- As a prior-weighted average of the likelihood, the marginal likelihood is heavily influenced by the proportion of samples that wind up in this "high likelihood" region, with the other samples contributing essentially zeroes to the average
- This presents two problems; one theoretical and one computational:

Inefficiency of Prior Sampling

- As a prior-weighted average of the likelihood, the marginal likelihood is heavily influenced by the proportion of samples that wind up in this "high likelihood" region, with the other samples contributing essentially zeroes to the average
- This presents two problems; one theoretical and one computational:
    1. The **actual** marginal likelihood is heavily dependent on how much probability mass the prior, in a way that the posterior isn't (as much)

Inefficiency of Prior Sampling

- As a prior-weighted average of the likelihood, the marginal likelihood is heavily influenced by the proportion of samples that wind up in this "high likelihood" region, with the other samples contributing essentially zeroes to the average
- This presents two problems; one theoretical and one computational:
  1. The **actual** marginal likelihood is heavily dependent on how much probability mass the prior, in a way that the posterior isn't (as much)
  2. The **estimate** of the marginal likelihood obtained by sampling from the prior is more and more unstable with more parameters, as the "high likelihood" region occupies a smaller and smaller share of the parameter space

Two Solutions

1. **Bridge Sampling**: A tailored sampling procedure for more robust estimates of marginal likelihoods

Two Solutions

1. **Bridge Sampling**: A tailored sampling procedure for more robust estimates of marginal likelihoods
    - Addresses the computational volatility, requiring a separate sampling algorithm

Two Solutions

1. **Bridge Sampling**: A tailored sampling procedure for more robust estimates of marginal likelihoods
    - Addresses the computational volatility, requiring a separate sampling algorithm
    - Does nothing to alleviate the heavy dependence of marginal likelihood on the prior

Two Solutions

1. **Bridge Sampling**: A tailored sampling procedure for more robust estimates of marginal likelihoods
    - Addresses the computational volatility, requiring a separate sampling algorithm
    - Does nothing to alleviate the heavy dependence of marginal likelihood on the prior
2. **Alternatives to Marginal Likelihood**

Two Solutions

1. **Bridge Sampling**: A tailored sampling procedure for more robust estimates of marginal likelihoods
   - Addresses the computational volatility, requiring a separate sampling algorithm
   - Does nothing to alleviate the heavy dependence of marginal likelihood on the prior
2. **Alternatives to Marginal Likelihood**
   - May be more practical, despite the conceptual elegance of marginal likelihood as an "automatic" Occam's Razor

Predictive Log Likelihood

- A general statistical principle, spanning both Bayesian and frequentist approaches is that **a good model should make good predictions on new data**

Predictive Log Likelihood

- A general statistical principle, spanning both Bayesian and frequentist approaches is that **a good model should make good predictions on new data**
- For a Bayesian model, prediction is **probabilistic**: we can't really score "correct" or "incorrect"

Predictive Log Likelihood

- A general statistical principle, spanning both Bayesian and frequentist approaches is that **a good model should make good predictions on new data**
- For a Bayesian model, prediction is **probabilistic**: we can't really score "correct" or "incorrect"
- What we can do, however, is examine $p(\mathbf{y}_{\text{unseen}} \mid \mathbf{y}_{\text{seen}}, m)$

Predictive Log Likelihood

- A general statistical principle, spanning both Bayesian and frequentist approaches is that **a good model should make good predictions on new data**
- For a Bayesian model, prediction is **probabilistic**: we can't really score "correct" or "incorrect"
- What we can do, however, is examine $p(\mathbf{y}_{\text{unseen}} \mid \mathbf{y}_{\text{seen}}, m)$
- Measures how much probability/density the model places on the data it hasn't seen, after learning what it can from the data it has seen

(Bayesian) K-fold Cross Validation

A. For each model, m, under consideration

1. Divide dataset into K subsets ("folds") with (approximately) equal cases per fold
2. For $k = 1, \ldots, K$:
   (a) **Designate fold k the "validation set"**, and the others the **training set**
   (b) **Estimate the Posterior** distribution of the parameters given only the training set
   (c) **Compute the predictive marginal likelihood**:

$$\begin{aligned} & p(\mathbf{y}_{\text{validation(k)}} \mid \mathbf{y}_{\text{training(k)}}, m) \\ & = \int p(\mathbf{y}_{\text{validation(k)}} \mid \theta, \mathbf{y}_{\text{training(k)}}) p(\theta \mid \mathbf{y}_{\text{training(k)}}) \; d\theta \\ & = \mathbb{E}\left[p(\mathbf{y}_{\text{validation(k)}} \mid \theta) \right] \end{aligned}$$

   where the expectation this time is with respect to the **posterior** for θ given $\mathbf{y}_{\text{training(k)}}$
3. **Return the average log predictive likelihood** across folds

Practical Challenge and a Practical Solution

- **Problem:** This requires doing MCMC K times, which is cumbersome

Practical Challenge and a Practical Solution

- **Problem:** This requires doing MCMC K times, which is cumbersome
- **Solution:** Use an **approximation** to the true CV marginal likelihood

Practical Challenge and a Practical Solution

- **Problem:** This requires doing MCMC K times, which is cumbersome
- **Solution:** Use an **approximation** to the true CV marginal likelihood
- Current state-of-the-art: **Pareto Smoothed Importance Sampling** to approximate **leave-one-out cross validation**

Practical Challenge and a Practical Solution

- **Problem:** This requires doing MCMC K times, which is cumbersome
- **Solution:** Use an **approximation** to the true CV marginal likelihood
- Current state-of-the-art: **Pareto Smoothed Importance Sampling** to approximate **leave-one-out cross validation**
- Abbreviated PSIS-LOO

Practical Challenge and a Practical Solution

- **Problem:** This requires doing MCMC K times, which is cumbersome
- **Solution:** Use an **approximation** to the true CV marginal likelihood
- Current state-of-the-art: **Pareto Smoothed Importance Sampling** to approximate **leave-one-out cross validation**
- Abbreviated PSIS-LOO
- Available via RStan with the `loo` R package