WSABI Summer Research Lab **Summer 2024**

Lecture 13: Fully Bayesian Models

Instructor: Ryan Brill *Scribe: Jonathan Pipping*

13.1 Example Models

Recall the Bayesian idea: treating parameters as random variables with an unknown **distribution** to be estimated, rather than as unknown fixed constants. We will now see how to use this idea to define fully Bayesian models.

13.1.1 Beta-Binomial Model

We have seen this model before, where we wanted to predict an MLB team's end-of-season win percentage from their mid-season wins and losses. Letting n be the number of games played so far, W being the observed number of wins, and p being our team's latent win probability, we have the model

$$\begin{cases} W \sim \text{Binomial}(n, p) \\ p \sim \text{Beta}(\alpha, \beta) \end{cases}$$

Note that we are modeling the team's latent win probability p using a Beta prior , which we will use to encode the prior information that p is more likely to be near $\frac{1}{2}$ than to be very near 0 or 1 (from our knowledge of baseball).

Then from Bayes' rule, we found that the posterior distribution $p \mid W$ is given by

$$\begin{aligned} p \mid W &\sim \text{Binomial}(n + \alpha + \beta - 2,\ \frac{W + \alpha - 1}{n + \alpha + \beta - 2}) \\ &\sim \text{Binomial}(n + \alpha + \beta - 2,\ \frac{W + \alpha - 1}{(W + \alpha - 1) + (L + \beta - 1)}) \end{aligned}$$

And the Bayes estimate (i.e. the posterior mean) is given by

$$\widehat{p}^{(\text{Bayes})} = \frac{W + \alpha - 1}{(W + \alpha - 1) + (L + \beta - 1)}$$

Recall that we can interpret this estimate as adding $\alpha - 1$ fake wins and $\beta - 1$ fake losses to the observed data.

13.1.2 Normal-Normal Model

We have also seen this model before, where we wanted to predict an MLB player's end-of-seaon batting average from their mid-season batting average. Letting H_i be the number of hits for player i in N_i at-bats, we have observed batting averages $X_i = \frac{H_i}{N_i}$. We formulate the model as

$$\begin{cases} X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \quad \sigma_i^2 = \frac{C}{N_i},\ C \text{ known} \\ \mu_i \sim \mathcal{N}(\mu, \tau^2) \end{cases}$$

When we model player i's latent quality μ_i using a Normal prior , we encodee the prior information that player i is also a baseball player, allowing us to share strength information across players.

Then from Bayes' Rule, we found that the posterior distribution $\mu_i \mid X_i$ is given by

$$\mu_i \mid X_i \sim \mathcal{N}\left(\frac{\left(\frac{X_i}{\sigma_i^2}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)},\ \frac{1}{\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}}\right)$$

And the Bayes estimate (i.e. the posterior mean) is given by

$$\widehat{\mu_i}^{(\text{Bayes})} = \frac{\left(\frac{X_i}{\sigma_i^2}+\frac{\widehat{\mu}}{\widehat{\tau}^2}\right)}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\widehat{\tau}^2}\right)}$$

Using $\widehat{\mu}$ and $\widehat{\tau}^2$ in place of μ and τ^2 since they are unknown.

13.1.3 Bayesian Regression

A Bayesian regression model takes the form

$$\begin{cases}\text{Regression:} \quad y_i \sim \mathcal{N}(x_i^T\beta, \sigma^2) \\ \text{Prior:} \quad \beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda})\end{cases}$$

where λ is a hyperparameter that controls the strength of the prior. A larger λ means a tighter prior (smaller variance $\frac{\sigma^2}{\lambda}$), which leads to more shrinkage of the regression coefficients towards 0.

If we use Bayes' Rule to find the posterior distribution $\mathbb{P}(\beta \mid \boldsymbol{X}, \boldsymbol{y})$, we find that

$$\beta \mid \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}\left(\frac{\left(X^TX+\lambda I\right)^{-1}X^T\boldsymbol{y}}{\lambda},\ \text{Var} = ??\right)$$

The variance is a bit messy, but the posterior mean is equivalent to Ridge Regression with a regularization parameter of λ. We will review this in more detail in the next lecture.

13.1.4 Uncertainty in Bayesian Models

To make decisions in sports (e.g. player valuation, play selection, etc.) we need to not only know a best estimate of the value of the player/play/decision, but also uncertainty quantification (e.g. error bars or prediction intervals) to describe how confident or certain we are in our estimate.

In the Bayesian setting, we are estimating the full posterior distribution of the unknown parameter, which naturally allows us to quantify uncertainty with confidence intervals to make more complete decisions.

We'll now see how to implement a fully Bayesian model in practice.

13.2 Creating NFL Power Ratings

We will follow Mark Glickman and Hal Stern's 1998 paper [GS] to create a fully Bayesian model for NFL power ratings.

13.2.1 The Data

We begin with a dataset of NFL scores from 2018-2023, containing information about the teams playing, the season/week of the game, and the final score. We display the first few rows of the dataset in Figure 13.1 below.

game_id	home_team	away_team	season_type	week	total_home_score	total_away_score	season	pts_H_minus_A
2018_01_ATL_PHI	PHI	ATL	REG	1	18	12	2018	6
2018_01_BUF_BAL	BAL	BUF	REG	1	47	3	2018	44
2018_01_CHI_GB	GB	CHI	REG	1	24	23	2018	1
2018_01_CIN_IND	IND	CIN	REG	1	23	34	2018	-11
2018_01_DAL_CAR	CAR	DAL	REG	1	16	8	2018	8
2018_01_HOU_NE	NE	HOU	REG	1	27	20	2018	7
2018_01_JAX_NYG	NYG	JAX	REG	1	15	20	2018	-5
2018_01_KC_LAC	LAC	KC	REG	1	28	38	2018	-10
2018_01_LA_OAK	LV	LA	REG	1	13	33	2018	-20
2018_01_NYJ_DET	DET	NYJ	REG	1	17	48	2018	-31
2018_01_PIT_CLE	CLE	PIT	REG	1	21	21	2018	0
2018_01_SEA_DEN	DEN	SEA	REG	1	27	24	2018	3

Figure 13.1: First few rows of the NFL games dataset.

To make writing the model easier, we add a few variables to the dataset:

- $S(i)$: the season the i^{th} game took place, indexed at 1 for the 2018-2019 season
- $H(i), A(i)$, indexes for the home and away teams (respectively) in the i^{th} game
- y_i the home-minus-away score differential $(y_{H_i} - y_{A_i})$ in the i^{th} game

We're now ready to write out our model.

13.2.2 The Model

We model the score differential y_i as follows:

$$y_i \sim \mathcal{N}(\beta_0 + \beta_{H(i),S(i)} - \beta_{A(i),S(i)}, \sigma^2_{\text{games}})$$

where $\beta_{H(i),S(i)}$ represents the strength of the home team in season $S(i)$, and $\beta_{A(i),S(i)}$ represents the strength of the away team in season $S(i)$.

Since this is a Bayesian model, all of our parameters are random variables with distributions. We define their priors next:

Autoregressive Prior

We ideally want to carry over infromation about a team's strength from one season to the next, which we do by using an autoregressive prior on all teams.

$$\beta_{j,S} \sim \mathcal{N}(\gamma \cdot \beta_{j,S-1}, \sigma^2_{\text{season}})$$

for all teams j and seasons $S > 1$.

This introduces more parameters to provide priors for, but the added connection between seasons is important for our model to work well. We now define the remaining priors:

Remaining Priors

- $\beta_0 \sim \mathcal{N}(0, 5^2)$
- $\beta_{j,1} \sim \mathcal{N}(0, \sigma^2_{\text{teams}})$
- $\sigma^2_{\text{teams}} \sim N_+(0, 5^2)$
- $\sigma^2_{\text{games}} \sim N_+(0, 5^2)$
- $\gamma \sim \text{Unif}[0, 1]$
- $\sigma^2_{\text{seasons}} \sim N_+(0, 5^2)$

13.2.3 Introduction to MCMC

In a general Bayesian statistical model, the parameters don't all have Gaussian priors or likelihoods, or it's quite large and complicated. As a result, we can't write on paper a closed-form analytical solution for the posterior distribution. How do we get around this?

Instead, we *approximate* the posterior distribution using **Markov Chain Monte Carlo** sampling methods like Gibbs sampling, Hamiltonian Monte Carlo, and No U-Turn Sampling. We won't cover all of these here, but we can use a probabilistic programming language like Stan, Jags, or NumPyro to do this for us!

13.2.4 Fitting the Model Using Stan

We'll use Stan to fit our model. We define this model in a separate file called `glickman-stern.stan` as written below.

```
data {
    int<lower = 1> N_games; // number of games
    int<lower = 1> N_teams; // number of teams
    int<lower = 2> N_seasons; // number of seasons

    real y[N_games]; // outcome vector (point differential)
    int<lower = 1, upper = N_teams> H[N_games]; // vector of home
       team indices
    int<lower = 1, upper = N_teams> A[N_games]; // vector of away
       team indices
    int<lower = 1, upper = N_seasons> S[N_games]; // vector of season
       indices
}

parameters {
    real beta_0; // intercept (home field advantage)
    real betas[N_teams, N_seasons]; // team strength coefficients for
       each team-season
    real<lower = 0> sigma_games; // game-level variance in point
       differential
    real<lower = 0> sigma_teams; // variance across teams before the
       first season
    real<lower = 0> sigma_seasons; // a team's variance across seasons
    real<lower = 0, upper = 1> gamma; // autoregressive parameter
}

model {
    // game-level model
    for(i in 1:N_games) {
        y[i] ~ normal(beta_0 + betas[H[i], S[i]] - betas[A[i], S[i]],
           sigma_games);
```

```
        }

        // team-level priors
        for(j in 1:N_teams) {
            // initial season prior across teams
            betas[j, 1] ~ normal(0, sigma_teams);
            for (s in 2:N_seasons) {
                // auto-regressive model across seasons
                betas[j, s] ~ normal(gamma * betas[j, s-1],
                   sigma_seasons);
            }
        }

    // priors
    sigma_games ~ normal(0, 5);
    sigma_teams ~ normal(0, 5);
    sigma_seasons ~ normal(0, 5);
    gamma ~ uniform(0, 1);
}
```

Stan uses Hamiltonian Monte Carlo (HMC) to approximate the posterior distribution of each parameter. Stan returns the approximate posterior distribution via posterior samples from this distribution. For example, the posterior of the β_j's are m draws $\beta_j^{(1)}, \ldots, \beta_j^{(m)}$ from the approximate posterior distribution.

To fit the model, we run the following code in R:

```
library(rstan)

# load stan model
model = stan_model(file = "glickman-stern.stan")
model

# create list of data compliant with stan model
data_train = list(
    N_games = nrow(nfl_data),
    N_teams = nrow(team_indices),
    N_seasons = length(unique(nfl_data$S)),
    y = nfl_data$y,
    H = nfl_data$H,
    A = nfl_data$A,
    S = nfl_data$S
)

# train the model
fit = sampling(
    model, data = data_train, iter = 1500, chains = 1, seed = 12345
)
fit
```

13.2.5 Results

From the model's output, we calculate the posterior median and form a 50% credible intervals using the 25% and 75% quantiles for non-strength parameters, which we display in Figure 13.2.

	param	post_lower	post_med	post_upper
	<chr>	<dbl>	<dbl>	<dbl>
1	beta_0	0.968	1.53	2.08
2	sigma_games	12.0	12.4	12.9
3	sigma_teams	3.70	4.94	6.75
4	sigma_seasons	3.02	3.70	4.49
5	gamma	0.493	0.662	0.814

Figure 13.2: Posterior median and 95% credible intervals for non-strength parameters.

We also display the posterior means and 50% credible intervals for our team strength parameters from 2018-2023 in Figure 13.3.

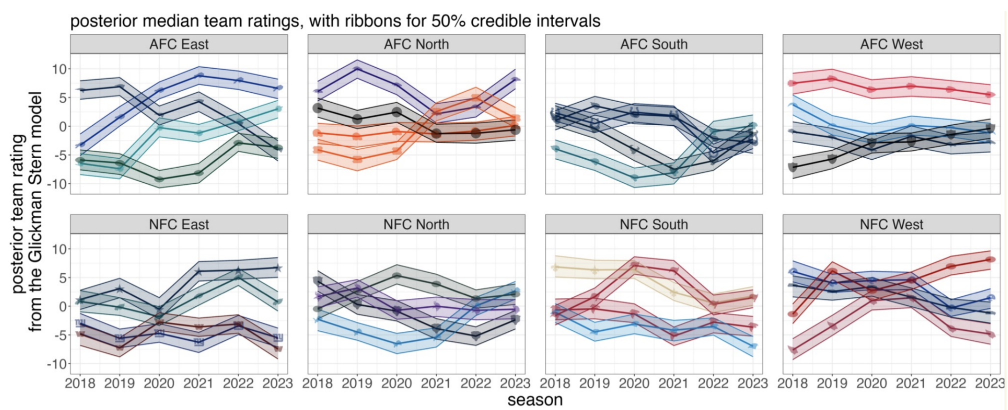

Figure 13.3: Posterior means and 50% credible intervals for NFL team strength parameters.

Papers Using Fully Bayesian Models

There are many great sports papers that use fully Bayesian models because they're interpretable, quantify uncertainty, capture various sources of uncertainty (variance) at once, and use shrinkage to improve estimation. Some examples include evaluating pitch framing in baseball [DW], evaluating fielding [JSW], evaluating home advantage in football [BL], and understanding randomness across North American sports [LMB]. All of these papers are included in the References section.

References

[BL] Benz, L. S., Bliss, T. J., & Lopez, M. J., *A comprehensive survey of the home advantage in American football*, arXiv preprint, 2024.

[DW] Deshpande, S. K., & Wyner, A., *A hierarchical Bayesian model of pitch framing*, Journal of Quantitative Analysis in Sports, 2017.

[GS] Glickman, M. E., & Stern, H. S., *A State-Space Model for National Football League Scores*, Journal of the American Statistical Association, 1998.

[JSW] Jensen, S. T., Shirley, K. E., & Wyner, A. J., *Bayesball: A Bayesian hierarchical model for evaluating fielding in major league baseball*, Annals of Applied Statistics, 2009.

[LMB] Lopez, M. J., Matthews, G. J., & Baumer, B. S., *How often does the best team win? A unified approach to understanding randomness in North American sport*, arXiv preprint, 2017.

[RB] Brill, R., Deshpande, S., & Wyner, A.J., *A Bayesian analysis of the time through the order penalty in baseball*, Wharton Sports Analytics & Business Initiative, 2023.