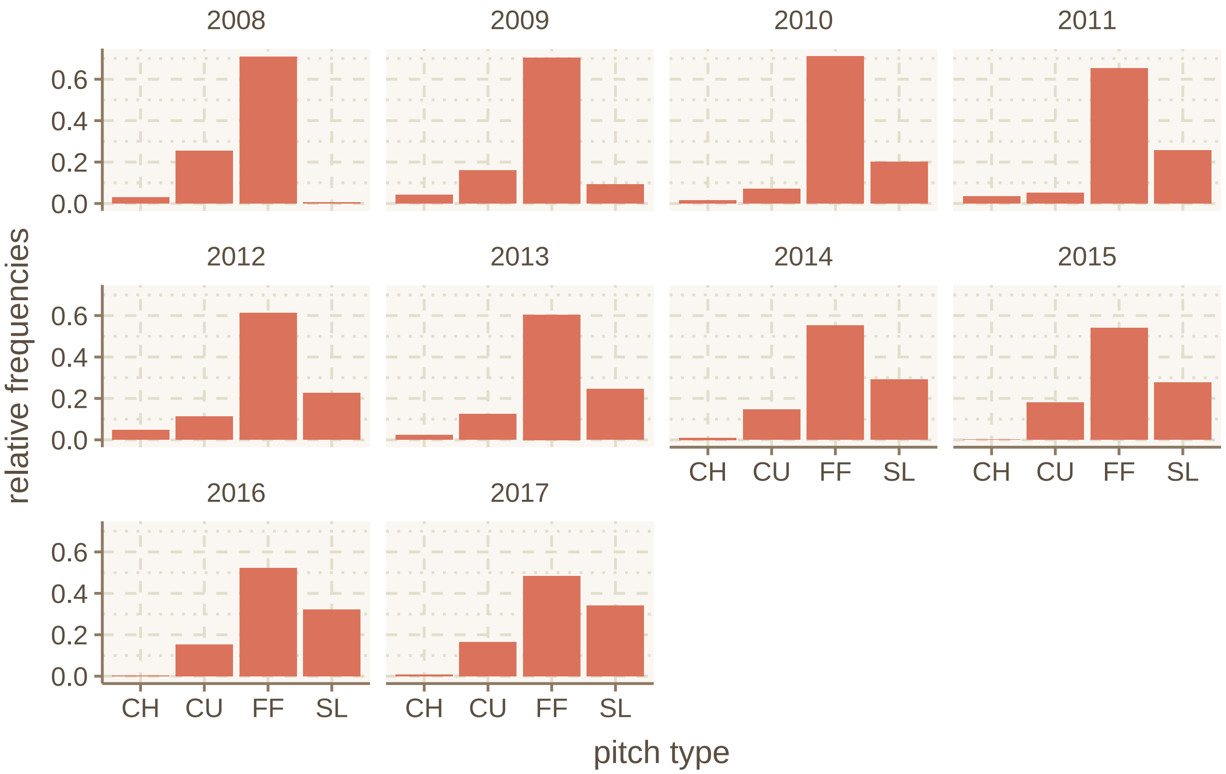
Hierarchical Kershaw model using Polya-gamma data augmentation

1 Introduction

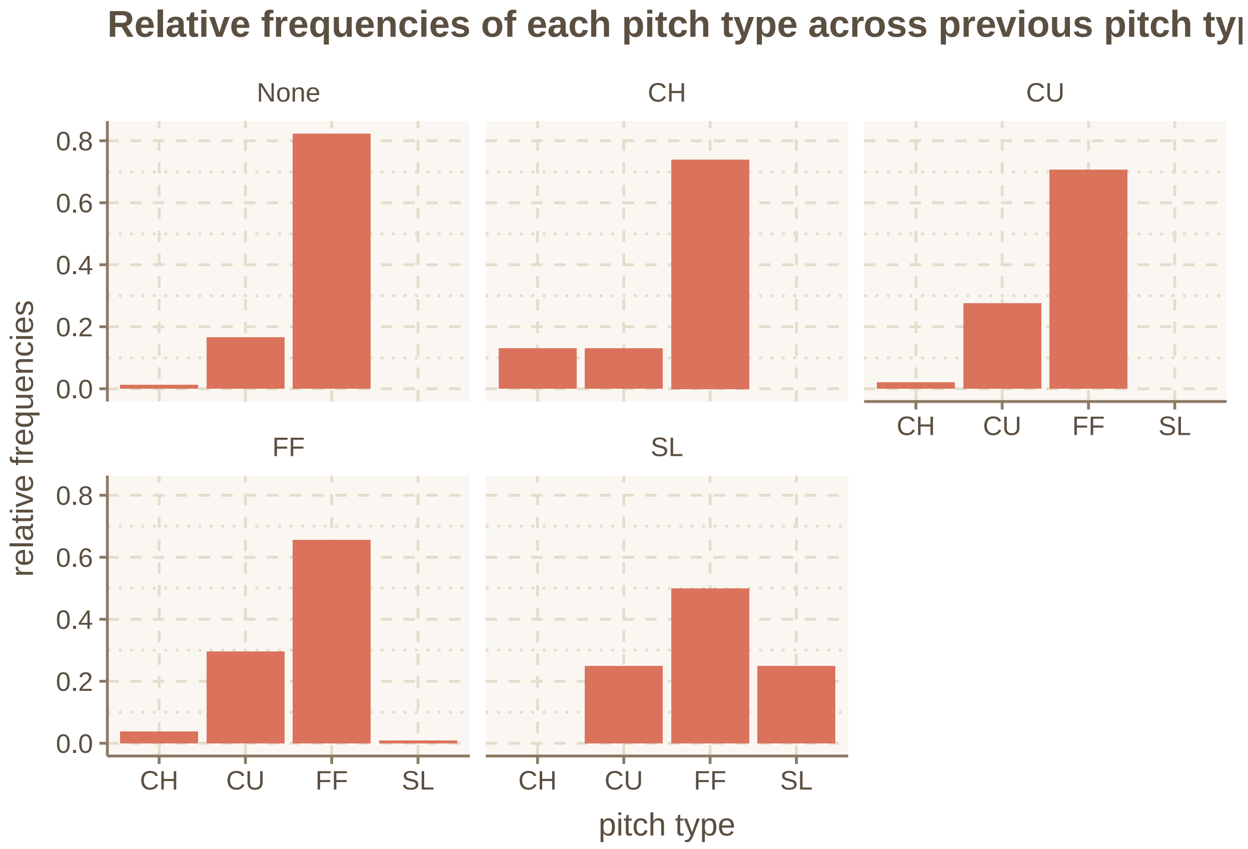
The goal of this document is to outline the logic used to create the pitch prediction model for Clayton Kershaw. In this document, we cover an exploratory data analysis, describe the model fit to the data, and conclude the results of that model and a discussion of those results.

2 Exploratory data analysis

In this section, we visualize the distribution of pitch type across a number of interesting predictors.



Relative frequencies of each pitch type across year



3 Model

In this section, we discuss the model and sampler.

3.1 Multinomial model

3.2 Hierarchical multinomial model

The model we fit to these data is described below.

$$\boldsymbol{y}_1, ...\boldsymbol{y}_{n_m}|\boldsymbol{\pi}_{im} \sim \text{multnomial}(1, \boldsymbol{\pi}_{im}), \qquad \pi_{ijm} = \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}_{jm}}}{\sum_{k=1}^{J} e^{\boldsymbol{x}_i \boldsymbol{\beta}_{km}}}$$

$$\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_M|\boldsymbol{\psi} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\psi} = \{\boldsymbol{\theta}, \boldsymbol{\Sigma}\}$$

where i indexes the observation $(i = 1, 2, \ldots, n_m)$, j the level of the response $(j = 1, 2, \ldots, J)$, and m indexes the member of the hierarchy $(m = 1, 2, \ldots, M)$. To draw from the joint posterior distribution of $\{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, we first sample the hyper-parameters with Gibbs draws using the full-conditional distributions derived on page 199 of (Hoff, 2009). To sample the regression coefficients, we make use of the Polya-gamma data augmentation strategy described by (Polson, Scott and Windle, 2013), the details of which are provided in appendix S1. This model requires prior distributions on the hyper-parameters. The following are semi-conjugate priors that allow for a Gibbs sampler.

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$
$$\boldsymbol{\Sigma} \sim \text{inverse-Wishart}(\eta_0, \boldsymbol{S}_0)$$

In general, we chose $\boldsymbol{\mu}_0 = \boldsymbol{0}$, $\boldsymbol{\Lambda}_0 = 100\boldsymbol{I}_p$, $\eta_0 = p$ and $\boldsymbol{S}_0 = \boldsymbol{I}_p$, where p is the number of regression coefficients being estimated, which results in weakly informative priors. These methods result in the following Gibbs sampler.

CLEAN UP NOTATION

```
for j in 1:J
    - sample $\boldsymbol{\theta}|\boldsymbol{\Sigma}, \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$
      where $\boldsymbol{\mu}_n = \boldsymbol{\Lambda}_n \left(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{k=1}^{m}\boldsymbol{\beta}_k\right)$ and $\boldsymbol{\Lambda}_n = \left(\boldsymbol{\Lambda}_0^{-1} + m\boldsymbol{\Sigma}^{-1}\right)^{-1}$
    - sample $\boldsymbol{\Sigma}^{-1}|\boldsymbol{\theta}, \boldsymbol{\beta} \sim \text{Wishart}(\eta_0 + m, \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}})$
      where $\boldsymbol{S}_{\boldsymbol{\theta}} = \sum_{k=1}^{m}(\boldsymbol{\beta}_k - \boldsymbol{\theta})(\boldsymbol{\beta}_k - \boldsymbol{\theta})'$
for m in 1:M
for j in 1:J
    - sample $\beta_j^m|\boldsymbol{\theta}_j, \boldsymbol{\Sigma}_j, \boldsymbol{z}_j \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{V})$
      where $\boldsymbol{V}_j = \left(\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$, $\boldsymbol{m} = \boldsymbol{V}(\boldsymbol{X}'(\boldsymbol{\kappa_j} + \boldsymbol{\Omega_j c_j}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$, $\eta_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta}_j - c_{ij}$, and
      $c_{ij} = \log\left(\sum_{k \neq j}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)\right)$.
```

4 Results

5 Discussion

Appendix S1: Multinomial Logistic Regression Derivations

Consider the multinomial regression model using the multinomial logit (softmax) link function.

$$\boldsymbol{y}_i|\boldsymbol{\pi}_i \sim \text{multinomial}\,(1, \boldsymbol{\pi}_i)$$
$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{\sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}$$

where $\boldsymbol{y}_i$ represents the vector of responses for the multinomial trial on observation i and $\boldsymbol{\pi}_i$ represents the vector of probabilities of success for each level of the multinomial trial, and π_{ij} represents the probability of success for level j on trial i.

To sample the joint posterior distribution of $\boldsymbol{\beta}$, we again make use of the Polya-gamma data augmentation strategy described by (Polson et al., 2013). To do so, we require the likelihood contribution of the regression coefficients associated with one level of the response conditional on the others. (Holmes and Held, 2006) showed that this contribution is given by the following:

$$\ell(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}, \boldsymbol{y}) \propto \prod_{i=1}^{N}\left(\frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}}\right)^{y_{ij}}\left(\frac{1}{1+e^{\eta_{ij}}}\right)^{n_i - y_{ij}} = \prod_{i=1}^{N}\frac{(e^{\eta_{ij}})^{y_{ij}}}{(1+e^{\eta_{ij}})^{n_i}}$$

where $\eta_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta}_j - c_{ij}$ and $c_{ij} = \log\left(\sum_{k\neq j}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)\right)$. Thus, it is clear that conditional on the regression coefficients associated with the other levels of the response, the likelihood contribution of $\boldsymbol{\beta}_j$ has the same form as that of the standard logistic regression model. Therefore, we can replicate the samplers described above, looping over $J-1$ (for identifiability) levels of the response.

If we let $z_{ij} = \frac{1}{\omega_{ij}}(y_{ij} - \frac{n_i}{2})$, then $z_{ij}|\boldsymbol{\beta}, \omega_{ij} \sim N(\eta_{ij}, \frac{1}{\omega_{ij}})$. We now derive the full conditional posterior distribution of $\boldsymbol{\beta}_j$, again assuming a $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ prior on $\boldsymbol{\beta}_j$.

$$\begin{aligned}
p(\boldsymbol{\beta}_j|\boldsymbol{z}, \boldsymbol{\Omega}_j) &\propto p(\boldsymbol{z}|\boldsymbol{\beta}_j, \boldsymbol{\Omega}_j)\cdot p(\boldsymbol{\beta}_j) \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{z}_j - (\boldsymbol{X}\boldsymbol{\beta}_j - \boldsymbol{c}_j)\right)'\boldsymbol{\Omega}_j\left(\boldsymbol{z}_j - (\boldsymbol{X}\boldsymbol{\beta}_j - \boldsymbol{c}_j)\right)\right\}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_j - \boldsymbol{\mu}_0\right)'\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\beta}_j - \boldsymbol{\mu}_0\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2\boldsymbol{\beta}_j'\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{z}_j - 2\boldsymbol{\beta}_j'\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{c}_j + \boldsymbol{\beta}_j'\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{X}\boldsymbol{\beta}_j\right)\right\}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_j'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(-2\boldsymbol{\beta}_j'\boldsymbol{X}'\boldsymbol{\Omega}_j(\boldsymbol{z}_j + \boldsymbol{c}_j) + \boldsymbol{\beta}_j'\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{X}\boldsymbol{\beta}_j\right)\right\}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_j'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(-2\boldsymbol{\beta}_j'\left(\boldsymbol{X}'\boldsymbol{\Omega}_j(\boldsymbol{z}_j + \boldsymbol{c}_j) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \boldsymbol{\beta}_j'\left(\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\beta}_j\right)\right\}
\end{aligned}$$

Consequently, we have the following full conditional posterior distributions:

$$\begin{aligned}\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}, \boldsymbol{z}_j, \boldsymbol{\Omega}_j &\sim \mathcal{N}(\boldsymbol{m}_j, \boldsymbol{V}_j)\\ \omega_{ij}|\boldsymbol{\beta}, \boldsymbol{Z} &\sim \mathrm{PG}(n_i, \eta_{ij})\end{aligned}$$

where $\boldsymbol{V}_j = \left(\boldsymbol{X}'\boldsymbol{\Omega}_j\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$, $\boldsymbol{m} = \boldsymbol{V}(\boldsymbol{X}'(\boldsymbol{\kappa_j} + \boldsymbol{\Omega_j c_j}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$, $\eta_{ij} = \boldsymbol{x}_i'\boldsymbol{\beta}_j - c_{ij}$, and $c_{ij} = \log\left(\sum_{k\neq j}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)\right)$.

References

Hoff, P.D. (2009) *A First Course in Bayesian Statistical Methods*. Springer.

Holmes, C.C. and Held, L. (2006) Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, **1**, 145–168.

Polson, N.G., Scott, J.G. and Windle, J. (2013) Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.