A Hidden Markov Model for an MLB Pitcher's Earned Runs

Ryan Brill. April 2021.

Contents

1 Overview: A 2-state Hidden Markov Model with Truncated-Normal Emissions

Some days, a pitcher is "hot", and other days, a pitcher is "cold." Even the legendary Clayton Kershaw, who is such a good pitcher that he won a league MVP in 2014 on top of the Cy Young award, has cold days, as evidenced by his various postseason meltdowns. But, we know Kershaw is a great pitcher, because his cold days are rare, and his hot days are associated with an extremely low ERA. This suggests the following criteria to judge how good a pitcher really is: How hot are his hot days? How cold are his cold days? And, how often is he hot? We shall explore these questions.

It is natural to evaluate a pitcher's performance in a given game by earned runs (ER). Therefore, to evaluate a pitcher's performance over the course of a season, we shall use his game-by-game earned runs as our observed data. Moreover, we shall model a pitcher's "coldness" or "hotness" in a given game as a latent variable that is responsible for his number of earned runs. Therefore, we shall use a Hidden Markov Model to model a pitcher's game-by-game earned runs.

Specifically, let $\mathcal{P}$ be the set of starting pitchers in 2019. Fix a pitcher $p \in \mathcal{P}$. The p^{th} pitcher's 2019 pitching performance is modelled by an HMM with 2 latent states, $\mathcal{S} = \{1, 2\} = \{\text{Hot}, \text{Cold}\}$. For his n^{th} game of the season, the p^{th} pitcher's hidden state is denoted by X_n, and his number of earned runs y_n is modelled by

$$y_n|X_n = \begin{cases} \mathcal{TN}(\mu_1, \sigma) & \text{if } X_n = 1 \\ \mathcal{TN}(\mu_2, \sigma) & \text{if } X_n = 2 \end{cases}$$

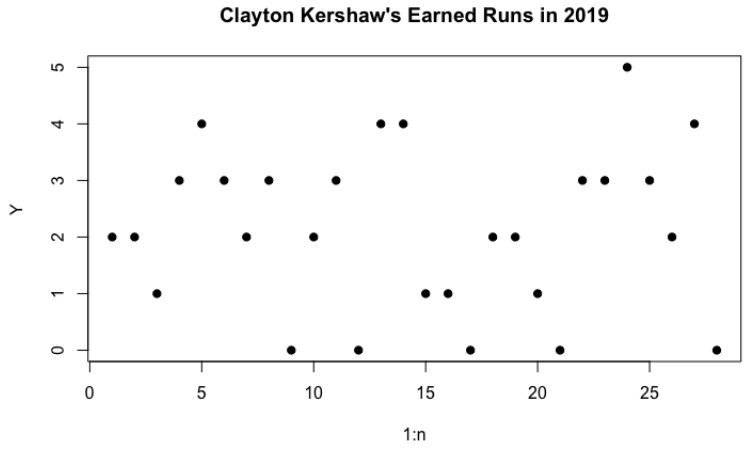
where $\mathcal{TN}$ is the normal distribution truncated at 0, and σ is a constant, different for each pitcher, which we assume to be known in order to simplify our model. We use a truncated normal because ER is a nonnegative statistic. So, the observed data for the p^{th} pitcher consists of $\mathbf{y} = \{y_n\}_{n=1}^{N_p}$, where N_p is the number of games he pitched in 2019, and y_n is his ER for the n^{th} game. For each pitcher, we use $\sigma = \text{sd}(\mathbf{y})/2$ because this value seemed to work well on examples I ran. I scraped the data from `baseballmusings.com`.

Now, as μ_1, μ_2, and $\mathbf{X} = \{X_n\}$ are unknown, it is our goal to estimate these parameters from the observed data $\mathbf{y}$. As good Bayesians, we use a Gibbs Sampler to obtain full posterior samples from $\mu_1|\mathbf{y}$, $\mu_2|\mathbf{y}$, and $\mathbf{X}|\mathbf{y}$. Note that prior to running the Gibbs Sampler on any pitching data, I tested it using synthetic data to check that it works, by comparing the computed hidden state posterior probabilities to the true states.

Note that we restrict ourselves to the 2019 MLB season, the most recent full-length season, because we do not want to take inter-seasonal effects into account. We also restrict ourselves to starting pitchers with at least 20 starts in 2019, because we want to have a sufficient amount of data to feed our model. This yields a dataset of $|\mathcal{P}| = 113$ pitchers.

2 Case Study: Modeling Clayton Kershaw's Earned Runs in 2019

In 2019, Clayton Kershaw was the starting pitcher for $N = 28$ games. Below is a plot of his earned runs in each of these games.

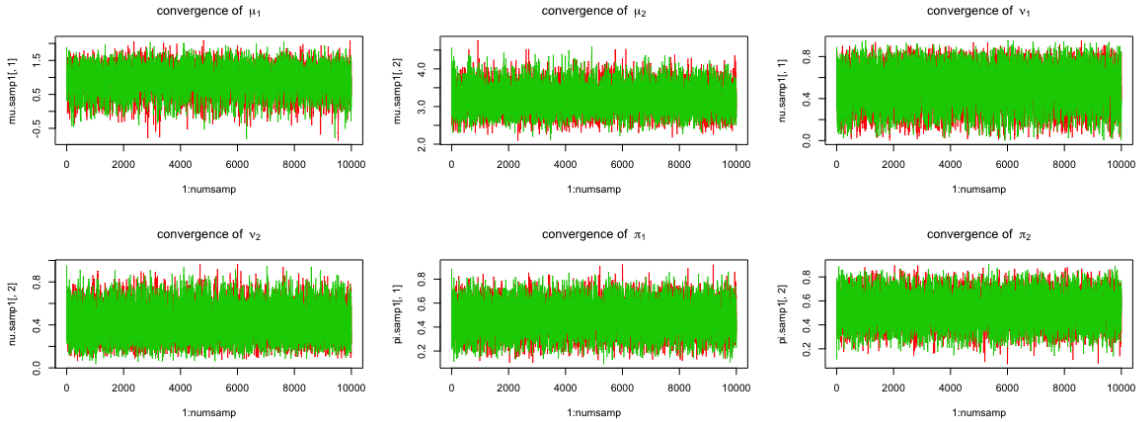


In Kershaw's worst game, he had 5 earned runs, which isn't that bad for a pitcher's worst game. In many of his games, Kershaw had 0 or 1 earned runs. So, we expect Kershaw's hot-state mean μ_1 to be somewhere in $(0, 1)$, and we expect his cold-state mean μ_2 to be somewhere in $(3, 5)$.

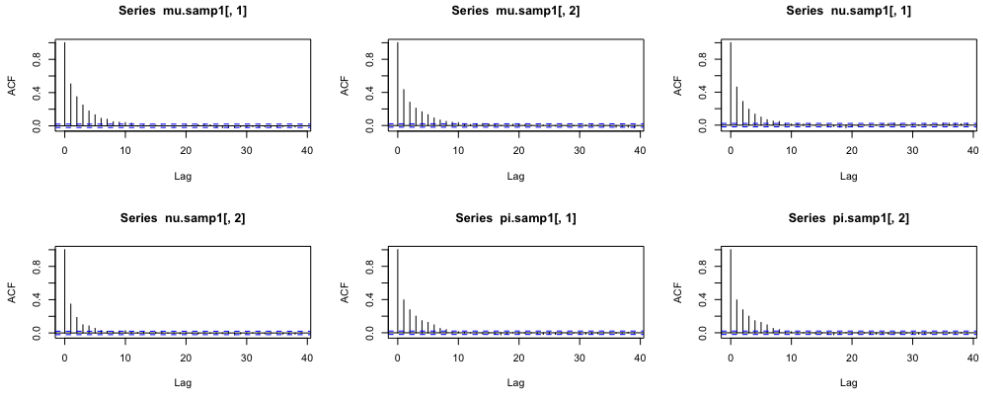
To obtain posterior samples of $\mu_1|\mathbf{y}$, $\mu_2|\mathbf{y}$, $\nu|\mathbf{y}$, and $\mathbf{X}|\mathbf{y}$, we run a Gibbs Sampler for 10000 iterations, with Kershaw's earned runs $\mathbf{y}$ as his observable data and $\sigma = sd(\mathbf{y})/2 = 0.72$ taken to be a known constant. To check convergence of the chain, we run the chain with 2 sets of initial parameters:

$$\nu = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}, \ \pi = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \ \mu = \begin{pmatrix} 1.5 \\ 3 \end{pmatrix} \qquad \text{and} \qquad \nu = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \ \pi = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix}, \ \mu = \begin{pmatrix} 0.4 \\ 6 \end{pmatrix}.$$

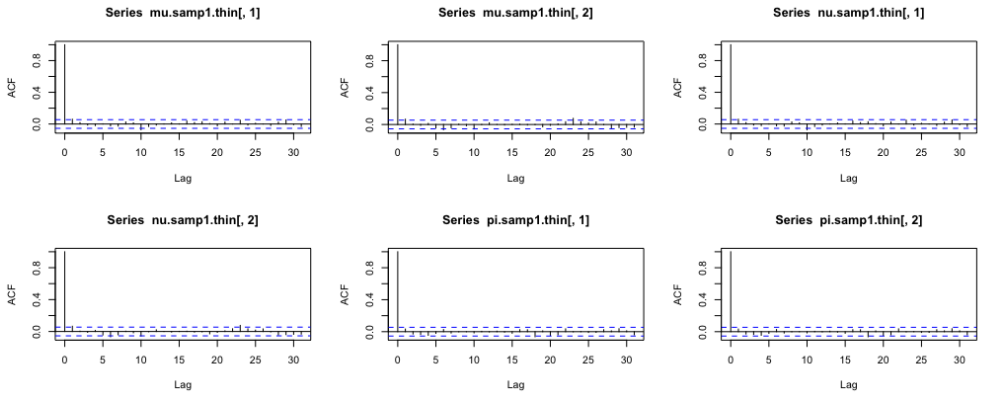
As indicated in the plot below, we see that the chain indeed converges from different starting values.



After throwing out the first 1000 samples as burn-in, we check the auto-correlation plots for ν, π, and μ, for both chains. To save space, we only plot the ACF for the first chain.

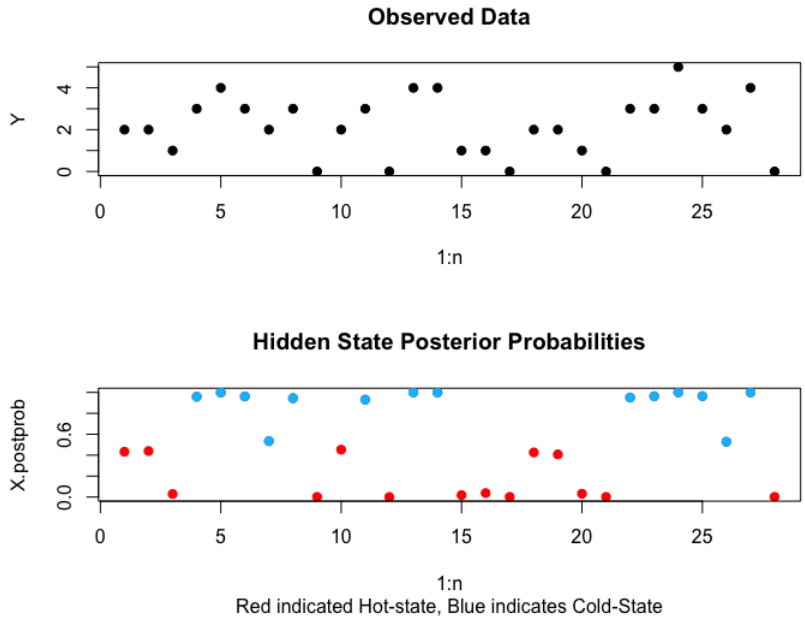


These plots indicate that we should thin the chains by keeping, say, every 7^{th} sample. After thinning, we check auto-correlation again just to be safe, and indeed the remaining auto-correlation is negligible.



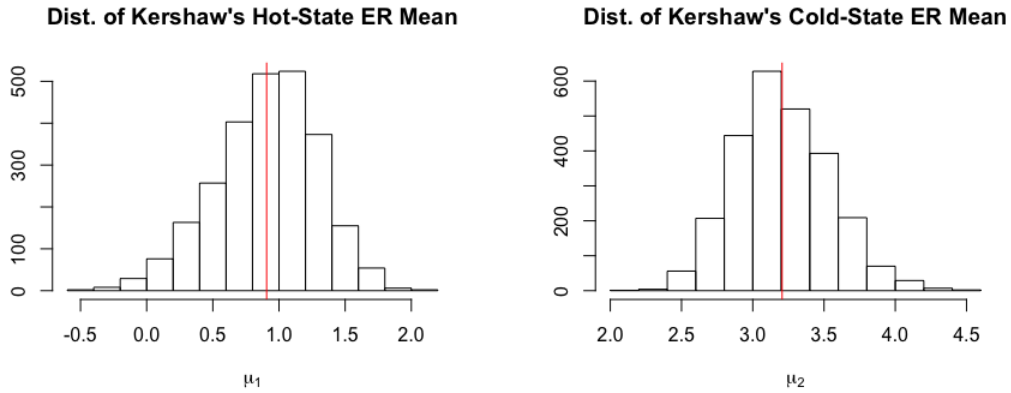
Then, we combine the two chains, and end up with 2570 samples of $\mu, \mathbf{X}, \nu$, and π. From these samples of the hidden-state path $\mathbf{X}$, we calculate the posterior probability that Kershaw was in the cold-state in game n as the empirical proportion of sampled paths that had Kershaw in the cold-state during game n.

We classify Kershaw as cold if his posterior probability of being cold is greater than 0.5; otherwise, we classify him as hot. This yields the following plot, indicating Kershaw's posterior probability of being cold during each game in 2019.



We see that in the games in which Kershaw is hot, his ER is in $[0, 2]$, and in the games in which Kershaw is cold, his ER is in $[2, 5]$. This provides a sanity check that the Gibbs sampler for $\mathbf{X}$ works. Also, the plot classifies Kershaw as hot for 50% of his games in 2019. Because Kershaw is such a good pitcher, I expect him to be hot for more than 50% of his games. However, his cold-state is still pretty good, and 2019 wasn't his best season (by his lofty standards), so the plot is still reasonable.

An important question remains: Just how good is Kershaw when he's hot, and how bad is he when he's cold? We address these questions by examining posterior samples of μ_1 and μ_2.



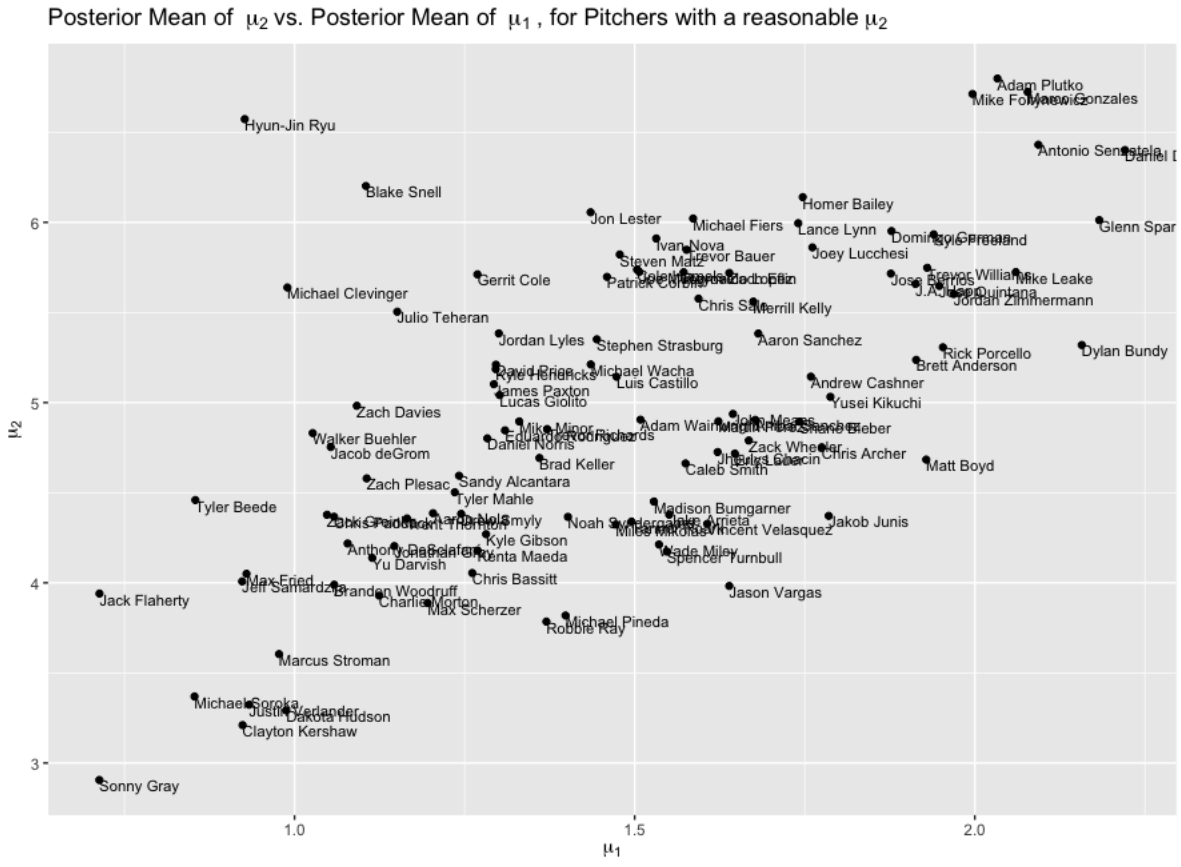
The posterior mean of μ_1 is $\hat{\mu}_1 = 0.9$, and the posterior mean of μ_2 is $\hat{\mu}_2 = 3.2$, which matches the expectations we had prior to running the Gibbs sampler. According to this model, when he's hot, Kershaw will have 0.9 earned runs on average, and when he's cold, 3.2 earned runs on average. Moreover, both μ_1 and

μ_2 have a posterior standard deviation of about 0.35, which is small enough to feel good about using the posterior mean as an estimate of the hot-state and cold-state number of earned runs in a game.

3 Comparing 2019's MLB Starting Pitchers

We run the Gibbs Sampler on all 113 starting pitchers who had at least 20 starts in 2019. Because it would take forever to evaluate the convergence and auto-correlation of all 113 pitchers, for all pitchers we run the Gibbs Sampler for 10000 iterations, use a burn-in of 1000, keep every 7^{th} sample, and use the same initial parameter values as with Kershaw.

Our first comparison of these pitchers consists of plotting $\hat{\mu}_1$ vs. $\hat{\mu}_2$, where $\hat{\mu}_i$ indicates the posterior mean of the sampled μ_i values for a given pitcher. In this plot, we enforce $\hat{\mu}_2 < 7.5$, which includes all but 5 pitchers. This condition allows us to read the names of the plot. Of these remaining 108 pitchers, $\hat{\mu}_1 \in [0, 2.5]$ and $\hat{\mu}_2 \in [2.75, 7.5]$. This makes sense: in order to start 20 games over the course of an MLB season, you need to be a decent pitcher, which on your hot days amounts to having an ERA in $[0, 2.5]$. Having a cold-day ERA of $[2.75, 7.5]$ isn't great, but it's fine as long as these cold days are less common than hot days.



It is easy to jump to the conclusion that pitchers in the bottom left corner of the plot, who have low $\hat{\mu}_1$ and $\hat{\mu}_2$, are the best. Indeed, Clayton Kershaw and 2019 AL Cy Young Winner Justin Verlander are in this lower left corner of the plot. However, 2019 NL Cy Young winner Jacob DeGrom is closer to the middle of the plot, with a much higher cold-state ER mean of $\hat{\mu}_2 \approx 4.5$, so $\hat{\mu}_1$ and $\hat{\mu}_2$ clearly don't tell the full

story of how good a pitcher really is. Indeed, someone with a high cold-state ERA can be better than someone with a lower cold-state ERA, as long as that pitcher is rarely cold! So, in order to rank the 2019 MLB starting pitchers, we need to incorporate their hidden-state paths $\mathbf{X}$ alongside μ_1 and μ_2.

3.1 Ranking 2019's MLB Starting Pitchers

For each pitcher $p \in \mathcal{P}$, let w_p be the probability that he is hot, so $1 - w_p$ is the probability he is cold. We define a pitcher's weighted earned-runs-mean, denoted as $wERM$, by

$$wERM_p := w_p \mu_{1,p} + (1 - w_p)\mu_{2,p}.$$

As discussed previously, we estimate $\mu_{1,p}$ and $\mu_{2,p}$ by taking the posterior mean of all the samples from the Gibbs Sampler. To estimate w_p, we use our posterior samples of the hidden-state paths $\mathbf{X}$. To do so, we first estimate the probability that the p^{th} pitcher was hot in his n^{th} game by the proportion $\gamma_{p,n}$ of sampled hidden-states X_n that are in the hot-state. Then, we estimate w_p by taking the proportion of the p^{th} pitcher's games which have greater than 0.5 posterior probability of being in the hot-state. In other words, $\hat{w}_p = \#\{\gamma_{p,n} > 0.5\}$. Therefore, we estimate the p^{th} pitcher's weighted earned-runs mean by

$$\widehat{wERM}_p = \hat{w}_p \hat{\mu}_{1,p} + (1 - \hat{w}_p)\hat{\mu}_{2,p},$$

and we rank the field of starting pitchers using this metric (a lower $wERM$ corresponds to a higher ranking). The top 10 $wERM$ rankings are shown on the left, and the top 10 ERA rankings are shown on the right.

	Pitcher	w	wERM
	<chr>	<dbl>	<dbl>
1	Hyun-Jin Ryu	0.862	1.71
2	Michael Soroka	0.655	1.72
3	Jacob deGrom	0.812	1.75
4	Justin Verlander	0.647	1.78
5	Sonny Gray	0.484	1.84
6	Michael Clevinger	0.810	1.88
7	Jack Flaherty	0.636	1.89
8	Dakota Hudson	0.594	1.92
9	Gerrit Cole	0.848	1.94
10	Chris Paddack	0.731	1.95

(a) Our Weighted Average Ranking

RK	NAME	POS	GP	GS	QS	ERA
1	Hyun Jin Ryu LAD	SP	29	29	22	2.32
2	Jacob deGrom NYM	SP	32	32	23	2.43
3	Gerrit Cole HOU	SP	33	33	26	2.50
4	Justin Verlander HOU	SP	34	34	26	2.58
5	Mike Soroka ATL	SP	29	29	18	2.68
6	Sonny Gray CIN	SP	31	31	17	2.87
7	Max Scherzer WSH	SP	27	27	17	2.92
8	Zack Greinke ARI	SP	33	33	24	2.93
9	Clayton Kershaw LAD	SP	29	28	22	3.03
10	Charlie Morton TB	SP	33	33	19	3.05

(b) Actual 2019 ERA Rankings, from `espn.com`.

Much of the top 10 is the same, although there are a few key differences. In particular, Mike Clevinger, Jack Flaherty, Dakota Hudson, and Chris Paddack are in the $wERM$ top 10, but not in the ERA top 10. These players have particularly high w values, indicating they are more often hot than cold, so the fact that they are in the $wERM$ top 10 but not the ERA top 10 is likely because their hot ERA's are given so much more weight than their cold ERA's.

4 Ideas for Future Work

- Evaluate a pitcher's streakiness by analyzing his posterior state-transition probabilities $\nu|\mathbf{y}$
- Evaluate plate-by-plate pitching, with Bernoulli emissions in $\{0, 1\} = \{\text{hit}, \text{out}\}$
- Use a Dynamic Linear model, with continuous hidden states, instead of discrete $\{\text{hot}, \text{cold}\}$ states