Pitch Sequences in Baseball: Analysis Using a Probabilistic Topic Model

Keisuke YOSHIHARA[a*] Kei TAKAHASHI[a,b]

[a]*Center for Mathematics and Data Science, Gunma University, Gunma, Japan*

[b]*School of Statistical Thinking, The Institute of Statistical Mathematics, Tokyo, Japan*

November 11, 2020

Abstract

In baseball, the pitch sequence is one of the most important factors in winning or losing a game; a single pitch may even define the game outcome. Therefore, it is useful for both batteries and hitters to have a deep understanding of the pitch sequence trend based on a variety of factors such as pitcher/hitter characteristics, game situations, and hitting results. However, while statistical techniques and machine learning methods have been increasingly applied to baseball data in recent years, pitch sequencing remains an under-studied area in baseball theory. This study investigates the relationship between pitch sequences and pitcher/hitter characteristics, game situations, and hitting results. A probabilistic topic model is applied to pitch-by-pitch data of all regular season games of Nippon Professional Baseball (NPB) in the period 2016–2018. The obtained results show that the model effectively extracts the pitch sequence trend. The model successfully identifies the pitch sequences likely to be used against sluggers, the effective pitch sequences to strike out a hitter, and the pitch sequence trend of an individual pitcher and in a specific match-up.

Keywords: Probabilistic topic model, Baseball data, Pitch sequence

1 Introduction

A pitch sequence is defined as a series of pitches based on the concept that initial pitches influence the behavior of the hitter at bat during later pitches. It is one of the most significant factors in baseball games, often deciding the winners and losers. Because a single pitch may sometimes define the outcome of a game, the catchers or batteries must decide the pitch sequence (i.e., the next ball to be pitched) by considering a variety of factors such as the current inning, counts, base runners, and hitter characteristics. When facing a pitch, it is important for hitters to consider the pitch sequences employed by the opposing battery. Therefore, understanding the pitch sequence trend based on the aforementioned factors helps batteries and hitters to decide and forecast pitch sequences, respectively. Although there are no correct or incorrect pitch sequences, batteries can reduce their risk of losing by managing pitch sequences. If they can prevent runners from moving up a base by making an extra-base hit a one-base hit and a one-base hit a mishit, the probability of losing a point and missing the game can be decreased. Furthermore, in baseball, hitters generally have a strong disadvantage over pitchers as a hitter is often classified as first rate if they bat .300. Thus, anticipating the pitch sequence employed by the opposing batteries is invaluable for hitters facing a pitch. Although any hitter could make a miss, there is a

*Corresponding author: yoshi@gunma-u.ac.jp

Electronic copy available at: https://ssrn.com/abstract=3728430

significant difference between facing a pitch with and without an understanding of the pitch sequence trend.

The simple and general approach to analyze pitch sequences is to create a scatter plot using the coordinates or locations of pitches and investigate their relationships with other factors such as counts and hitting results. However, it is difficult to individually investigate every plate appearance because the possible combinations of factors may be infinite. Alternatively, if one stratifies the samples and creates scatter plots using the coordinates of pitches in several plate appearances simultaneously, the information displayable at a given time is restricted. As an example, the pitch distribution with respect to ball-strike count is shown in Figure 1. It can be observed that it is fairly difficult to add information such as pitch speed and sequence to Figure 1.

In this study, pitch sequence data were accessed as text data, and a probabilistic topic model was applied to investigate the relationship between pitch sequences and pitcher/hitter characteristics, game situations, and hitting results. The pitch-by-pitch data of all regular season games of Nippon Professional Baseball (NPB) for the period 2016–2018 were utilized. This approach can obtain the pitch sequencing pattern that is common to a particular hitting result and that has higher probability of setting down the hitters, which can be helpful to pitch sequencing strategies of batteries. It can obtain the patterns of pitch sequences based on hitter characteristics and game situations. This can enable hitters to better forecast pitch sequences. The results show that the probabilistic topic model provides clarity about the pitch sequence trend. It also demonstrates the existence of a pitch sequencing pattern that is common to a specific hitter's characteristics, game situations, and hitting results.

In recent years, statistical techniques and machine learning methods have been increasingly applied to baseball data. However, these methods remain limited despite the growing availability of rich data with the introduction of sensor technologies such as PITCHf/x. Koseler and Stephan (2017) have provided a strong systematic review of the applications of machine learning methods to baseball data. They demonstrated that the approaches can be categorized into three classes: regression, binary classification, and multi-class classification. They also found that two algorithms are heavily used in existing research: support vector machines for classification problems and k-nearest neighbors for both classification and regression problems. To the best of our knowledge, the present study is the first to apply a probabilistic topic model to baseball data. It addresses the need for further development of baseball data analysis and contributes to the study of pitch sequences, which have not been sufficiently analyzed yet. Although various studies on pitch sequences have been conducted (e.g., Gray (2002a), Gray (2002b), Glaser (2010), Roegele (2014), Bonney (2015), Healey and Zhao (2017), Martin (2019)), they have largely been studied from the perspective of the battery (e.g., for obtaining the optimal pitch sequence to set hitters down). However, as mentioned earlier, it is also important for hitters to understand the pitch sequence trend. The present work aims to provide insights into pitch sequences from the perspective of the hitter, in addition to those from the perspective of the battery (e.g., the type of pitch sequences utilized for a specific type of hitter). Healey and Zhao (2017) have conducted a study that adopts a similar approach in terms of investigating the relationship between pitch sequences and other variables. They defined pitch sequences as a pitch-by-pitch correlation between location, velocity, and movement and investigated the relationship between pitch sequences and strikeout rates by estimating the regression model utilizing PITCHf/x data. However, they focused exclusively on the relationship between pitch sequences and strikeout rates and ignored other influential variables such as pitcher/hitter characteristics. In the present work, the probabilistic topic model enables a comprehensive investigation of the relationship between pitch sequences and several other variables such as pitcher/hitter characteristics, game situations, and hitting results.

Electronic copy available at: https://ssrn.com/abstract=3728430

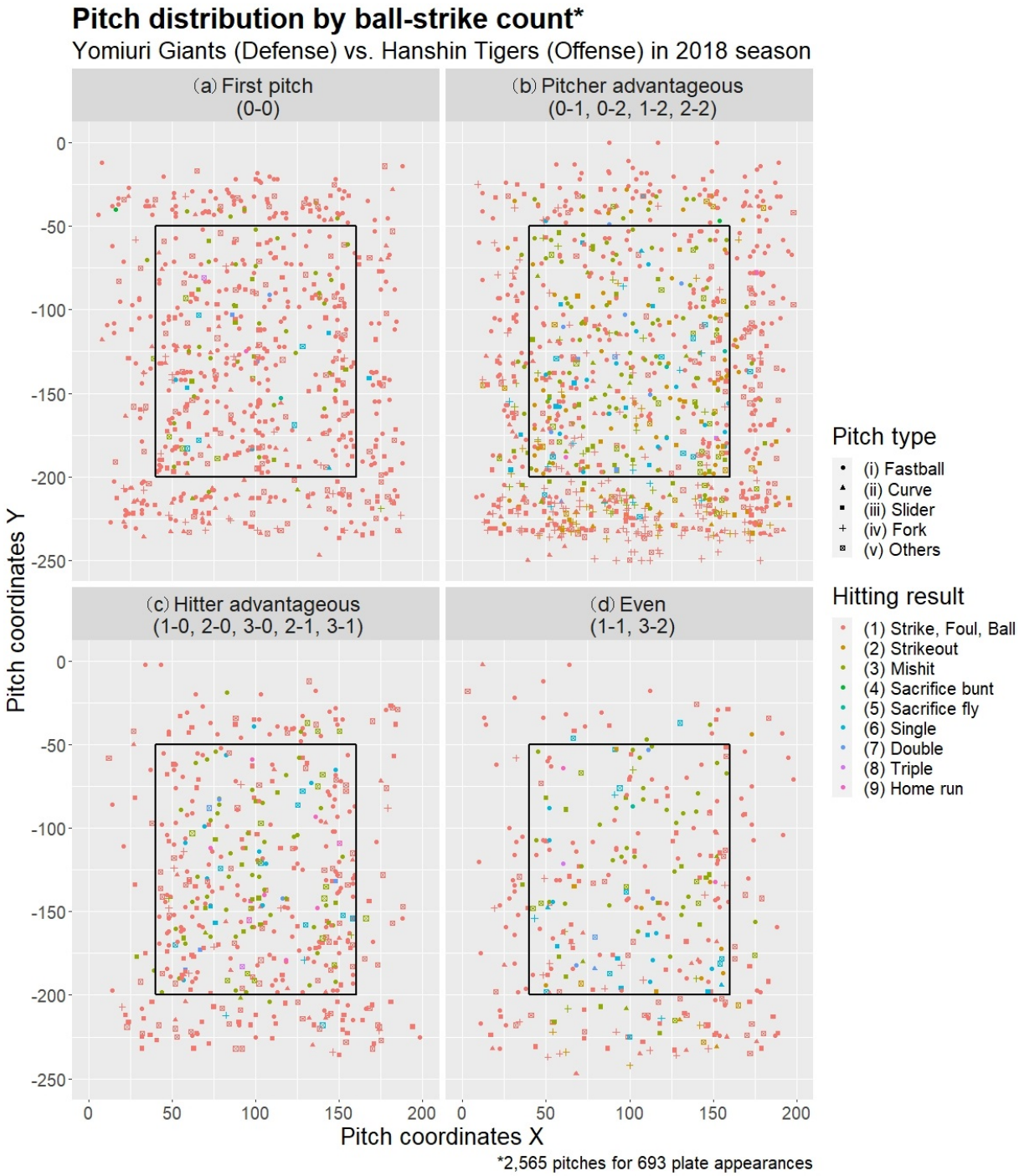

Figure 1 Pitch distribution

The remainder of this paper is organized as follows. A brief introduction to the probabilistic topic model and an explanation of its application to the obtained data are presented in Section 2. The model, data, and estimation method are detailed in Sections 3, 4, and 5, respectively. The results of the application of the model are provided in Section 6. Finally, the conclusions are presented in Section 7.

Electronic copy available at: https://ssrn.com/abstract=3728430

1. For document $d = 1, \cdots, M$
   Draw a topic distribution $\boldsymbol{\theta}_d \sim Dirichlet(\boldsymbol{\alpha})$
2. For topic $k = 1, \cdots, K$
   Draw a word distribution $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta})$
3. For document $d = 1, \cdots, M$
   For word $i = 1, \cdots, n_d$
      Draw a topic $z_{d,i} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$
      Draw a word $w_{d,i} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{d,i}})$

Figure 2 Generative process of LDA

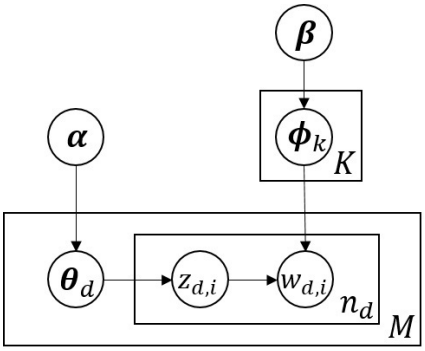

Figure 3 Graphical model of LDA

2 Probabilistic Topic Model

2.1 Brief Introduction

The probabilistic topic model is a type of statistical model used to analyze text data in natural language processing.[1] It is frequently used as a text-mining tool to discover the hidden semantic structures in documents. This model is essentially a generative model for documents; a document is represented as a set of words, and its generative process is stochastically modeled. The probabilistic topic model also enables dimension reduction in the sense that the mass of information in a document is summarized into a few topics that indicate the document's latent semantics. This allows functionalities such as extracting the most significant topics from a large set of documents, estimating the subject of each document, and classifying documents based on the estimated topics. By virtue of its general versatility, the probabilistic topic model is widely applied in various fields such as recommendation systems, customer segmentation, social network analysis, image processing, bioinformatics, and music information processing.

The most well-known probabilistic topic model is latent Dirichlet allocation (LDA). LDA was first proposed by Blei et al. (2003) as a probabilistic generative model for documents. Since then, a large number of extended models based on it have been proposed. The bag-of-words (BOW) assumption, i.e., ignoring the sequential order of words, is crucial to LDA, and this assumption is generally made in the extended models. The LDA model is based on the principle that a document is a mixture over topics (i.e., a single document can be composed of multiple topics) and a topic is a mixture over words (i.e., each word has some probability of belonging to a topic). The generative process and graphical model of LDA are illustrated in Figure 2 and 3, respectively. To generate a new document d, a probability distribution over topics (topic distribution) $\boldsymbol{\theta}_d$ and a probability distribution over words (word distribution) $\boldsymbol{\phi}_k$ for each topic k are chosen. Then, for each word i in document d, a topic $z_{d,i}$ is randomly drawn from the topic distribution $\boldsymbol{\theta}_d$, and a word $w_{d,i}$ is drawn based on the word distribution $\boldsymbol{\phi}_{z_{d,i}}$. The goal of LDA is to estimate the topics $\boldsymbol{z}$, topic distribution $\boldsymbol{\theta}$, and word distribution $\boldsymbol{\phi}$. Several estimation algorithms, including variational inference methods and Markov chain Monte Carlo (MCMC) methods, have been proposed. The collapsed Gibbs sampling algorithm proposed by Griffiths and Steyvers (2004) is known to be fairly efficient for estimating the LDA, among others. As per their approach, the sampling equation

[1] Blei (2012) presents a comprehensive survey of probabilistic topic models.

Electronic copy available at: https://ssrn.com/abstract=3728430

Table 1 Plate appearance as a set of pitches

#Pitch	Type	Speed (km/h)	Result	B/S/O
1	Fast	143	Ball	0/0/0
2	Slider	122	Called strike	1/0/0
3	Slider	123	Swinging strike	1/1/0
4	Fast	143	Foul	1/2/0
5	Slider	125	Ball	1/2/0
6	Cutter	121	Grounded to second	2/2/0

for $z_{d,i}$ is written as follows:

$$p(z_{d,i} = k | w_{d,i} = v, \boldsymbol{w}^{\backslash d,i}, \boldsymbol{z}^{\backslash d,i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{n_{k,v}^{\backslash d,i} + \beta_v}{n_{k,\cdot}^{\backslash d,i} + \sum_{v'} \beta_{v'}} \cdot \frac{n_{d,k}^{\backslash d,i} + \alpha_k}{n_{d}^{\backslash d,i} + \sum_{k'} \alpha_{k'}} = \phi_{k,v} \cdot \theta_{d,k},$$

where $n_{d,k}^{\backslash d,i} = \sum_{i=1}^{n_d} \delta(z_{d,i} = k) - \delta(z_{d,i} = k)$ is the number of times topic k is assigned to some word in document d (excluding the current instance i) and $n_{k,v}^{\backslash d,i} = \sum_{d=1}^{M} \sum_{i=1}^{n_d} \delta(w_{d,i} = v, z_{d,i} = k) - \delta(w_{d,i} = v, z_{d,i} = k)$ is the number of times word v is assigned to topic k (excluding the current instance i).

2.2 Application to Baseball Data

The pitch sequence data, presented in Table 1, can be viewed as text data. A plate appearance is represented as a set of pitches. By regarding a plate appearance as a *document* and a pitch as a *word*, the probabilistic topic model can be applied to baseball data, and the process by which the pitch sequence in a plate appearance is generated can be stochastically modelled. For example, a pitch can be defined as a combination of pitch type, speed, location, and ball-strike count. Moreover, the baseball data have an additional feature in the context of the application of the probabilistic topic model: the existence of document-level metadata. Each plate appearance has various auxiliary information such as pitcher/hitter characteristics, game situations (such as the inning and the number of outs), and hitting result (such as strikeouts and home runs). In this study, these types of document-level metadata are used alongside the document-word information to estimate the probabilistic topic model and investigate the relationship between pitch sequences and diverse information on each plate appearance.

3 Structural Topic Model

The structural topic model (STM) proposed by Roberts et al. (2013) is employed in this study. STM combines and extends three existing probabilistic topic models: the correlated topic model (CTM) proposed by Blei and Lafferty (2007), Dirichlet-multinomial regression (DMR) proposed by Mimno and McCallum (2008), and sparse additive generative model (SAGE) proposed by Eisenstein et al. (2011). In STM, as is the case with LDA, a document is defined as a mixture over topics, and a topic is defined as a mixture over words under the BOW assumption. The distinct difference between STM and LDA is that the former allows us to incorporate document-level metadata, such as author information of academic papers and scores of movie reviews, into topic estimation. This yields more accurate estimation and superior qualitative interpretability of estimated topics. In particular, the topical prevalence and topical content can be functions of the document's metadata. Topical prevalence refers to the extent of a document that

Electronic copy available at: https://ssrn.com/abstract=3728430

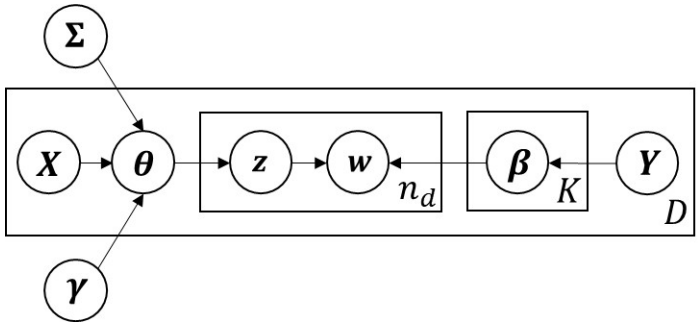

Figure 4 Graphical model of STM

is associated with a topic, and topical content refers to the words used within a topic. Essentially, it can be said that STM allows the topic and word distributions to depend on the document's metadata. The graphical model of STM is illustrated in Figure 4. The metadata that explain topical prevalence and topical content are referred to as prevalence covariates ($\boldsymbol{X}$ in Figure 4) and content covariates ($\boldsymbol{Y}$ in Figure 4), respectively. STM allows the inclusion of topical prevalence covariates only, topical content covariates only, both, or neither. In the case of no covariates, the STM reduces to CTM. The generative process of STM is summarized as follows:

1. For document $d = 1, \cdots, D$, draw the document-specific topic distribution from a logistic-normal generalized linear model based on a vector of prevalence covariates $\boldsymbol{x}_d$:

$$\boldsymbol{\theta}_d \sim \text{LogisticNormal}(\boldsymbol{\Gamma}'\boldsymbol{x}_d', \boldsymbol{\Sigma}) \text{ for } d = 1, \cdots, D,$$

   where $\boldsymbol{\Gamma}$ is the matrix of coefficients for $\boldsymbol{x}_d$ and $\boldsymbol{\Sigma}$ is the covariance matrix.

2. Form the document-specific word distribution $\beta_{d,k,v}$ for document d, topic k, and word v using the baseline word distribution m_v, topic-specific deviation $\kappa_{k,v}^{topic}$, content covariate-specific deviation $\kappa_{y_d,v}^{cov}$, and interaction between them $\kappa_{y_d,k,v}^{int}$ as follows:

$$\beta_{d,k,v} = \frac{\exp\left(m_v + \kappa_{k,v}^{topic} + \kappa_{y_d,v}^{cov} + \kappa_{y_d,k,v}^{int}\right)}{\sum_v \exp\left(m_v + \kappa_{k,v}^{topic} + \kappa_{y_d,v}^{cov} + \kappa_{y_d,k,v}^{int}\right)} \text{ for } d = 1, \cdots, D, k = 1, \cdots, K, v = 1, \cdots V$$

3. For each word $i = 1, \cdots, n_d$ in document $d = 1, \cdots, D$,

   (a) Draw a topic assignment from the document-specific topic distribution:

$$z_{d,i} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$$

   (b) Draw a word, given the assigned topic as

$$w_{d,i} \sim \text{Multinomial}(\boldsymbol{\beta}_{d,z_{d,i}})$$

4 Data

Pitch-by-pitch data provided by Data Stadium Inc. are utilized in this study. The data cover every pitch thrown during the 2016–2018 NPB regular seasons and include a total of 799,507 pitches and 196,789

Electronic copy available at: https://ssrn.com/abstract=3728430

plate appearances. The data were preprocessed to eliminate the plate appearances corresponding to the following conditions from the analysis: (1) a hitting result is other than a strikeout, mishit, single hit, two-base hit, three-base hit, or home run; (2) a baserunner situation is changed during the plate appearance. Following the elimination of plate appearances with missing data, the data for the analysis included a total of 523,905 pitches and 140,877 plate appearances.

As mentioned earlier, each plate appearance is considered a document and each pitch a word. Here, a pitch is defined by a combination of pitch speed, type, location, and ball/strike call. The unit of pitch speed is km/h. There are ten pitch types: fastball (FA), forkball (FO), slider (SL), tailing fastball (SH), curve (CV), cutter (CT), changeup (CH), sinker (SI), palm ball (PA), and knuckleball (KN). The pitch location is defined by a combination of the hitter's right- or left-handedness (R or L); inside (In), middle (Mid), or outside (Out); and high (Hi), middle (Mid), or low (Low). For example, "*148FAROutHiS*" indicates 148 km/h (*148*), fastball (*FA*), high and outside for a righted-handed hitter (*ROutHi*), and a strike (*S*).

The hitter and pitcher characteristics, game situations, and hitting results are included as document-level metadata. The hitter characteristics include their handedness, batting order, plate appearances in the game, and dummy variables for a high-average and home-run hitter. The high-average hitter dummy takes a value of 1 if the hitter clears the .300 mark at least once in the 2016–2018 seasons and 0 otherwise. The home-run hitter dummy takes a value of 1 if the hitter reaches the 30-homer mark at least once in the 2016–2018 seasons and 0 otherwise. The pitcher characteristics include their handedness, role (starter or relief), number of pitches in the game, and a dummy variable for winningest pitcher. The winningest pitcher dummy takes a value of 1 if the pitcher leads the club in wins for the 2016–2018 seasons and 0 otherwise. The game situations include an inning and a combination of the number of outs and baserunner situations. The number of outs is grouped into two categories: (i) none or one out, and (ii) two outs. The baserunner situations are grouped into three categories: (a) base empty, (b) runner on first, and (c) scoring position. The hitting results are grouped into six categories: (1) called strikeout, (2) swinging strikeout, (3) weak mishit, (4) strong mishit, (5) single hit, and (6) extra-base hit. In addition, the dummy variables for bunt, hit and run, and double play are included.

5 Model Estimation

The model is estimated by a semi-collapsed variational expectation-maximization algorithm using the R package "stm". The STM assumes a fixed user-specified number of topics; there is no specific number of topics that is objectively appropriate for a given corpus (Grimmer and Stewart (2013)). In the present study, the model estimation is attempted multiple times while varying the number of topics, and the final number of topics is set based on two criteria: "*Semantic coherence*" and "*Exclusivity*". Semantic coherence is a criterion based on the principle that it is more effective to extract a large number of topics composed of words that are semantically close to each other, and the effectiveness is maximized when the most probable words in a given topic frequently co-occur. Exclusivity is a criterion based on the principle that it is more effective to extract a small number of topics in which the most common words are similar to each other, and the effectiveness is maximized when the most probable words for a given topic are exclusive. Figure 5 depicts the relationship between the number of topics and the two criteria mentioned earlier. The dashed line indicates the median for each criterion. Based on the trade-off between the two criteria and the fact that setting an excessively large (or small) number of topics increases the difficulty of interpreting the estimated topics, the number of topics is set at 10 in this study.

Electronic copy available at: https://ssrn.com/abstract=3728430

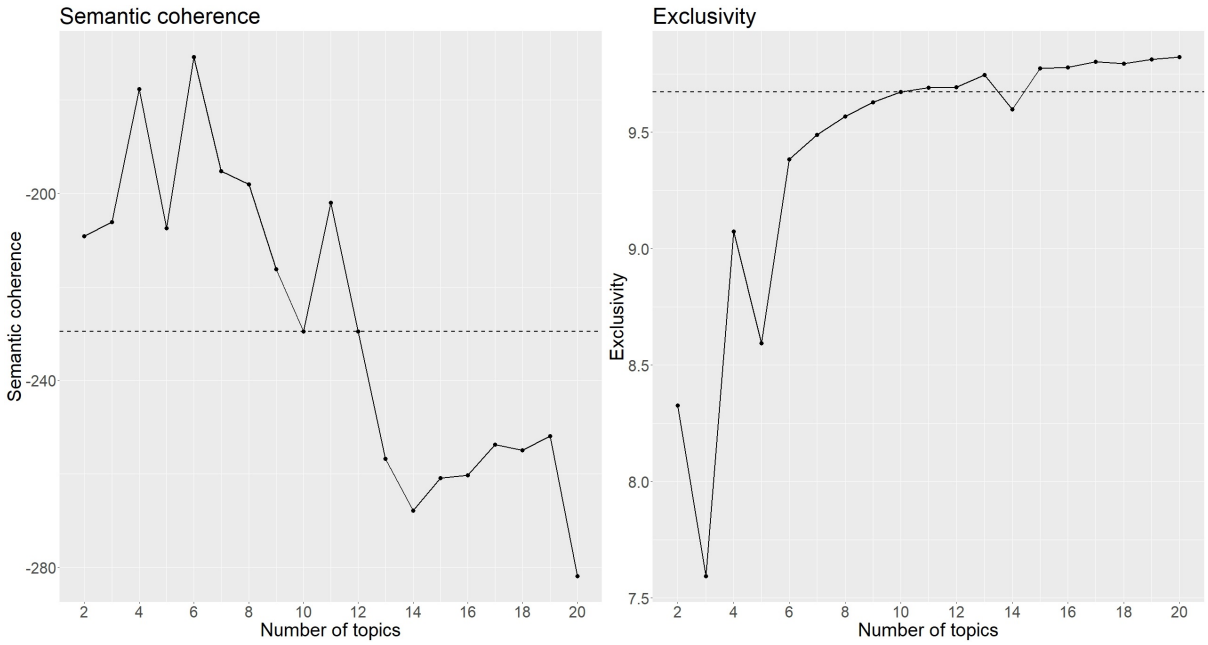

Figure 5 Number of topics

6 Results

6.1 Word Distribution for Each Topic

The word distribution for each topic is shown in Figure 6. Because a topic in the probabilistic topic model is a set of words that are likely to co-occur, the topic can be interpreted as a pattern of pitch sequences or an intention of batteries. For example, Topic 3 consists of fastballs on the inside and outside. Therefore, it can be interpreted as a pitch sequence that the battery intends to jam a hitter or to get a hitter concerned inside in order to get him out outside.

6.2 Relationship between Metadata and Topic Proportions

First, the marginal effects of high-average and home-run hitters on topic proportions are shown in Figure 7. For high-average hitters, compared to other hitters, the batteries tend to heavily utilize Topics 4 and 6 or the pitch sequences intended to keep the ball off the barrel of the bat. Because high-average hitters generally have good contact with a pitch and are difficult to strike out, such pitch-sequencing strategies might be effective for them. For home-run hitters, compared to other hitters, the batteries tend to heavily utilize Topic 7 or the pitch sequence intended to make hitters swing at a bad pitch. Because home-run hitters can easily hit a home run if the ball catches much of the plate, it might be dangerous for pitchers to put a ball over the plate without careful consideration. It can be observed that the batteries intend to make the hitter swing by throwing a low and bad breaking ball for home-run hitters.

Second, the marginal effects of the number of outs on topic proportions for the case of a runner on first are shown on the left side of Figure 8. In the case of none or one out, compared to the case of two outs, the pitch sequences intended to induce a double play are strongly used. In particular, against right-handed hitters, the batteries tend to heavily utilize Topics 3, 4, and 6 or the pitch sequences intended

Electronic copy available at: https://ssrn.com/abstract=3728430

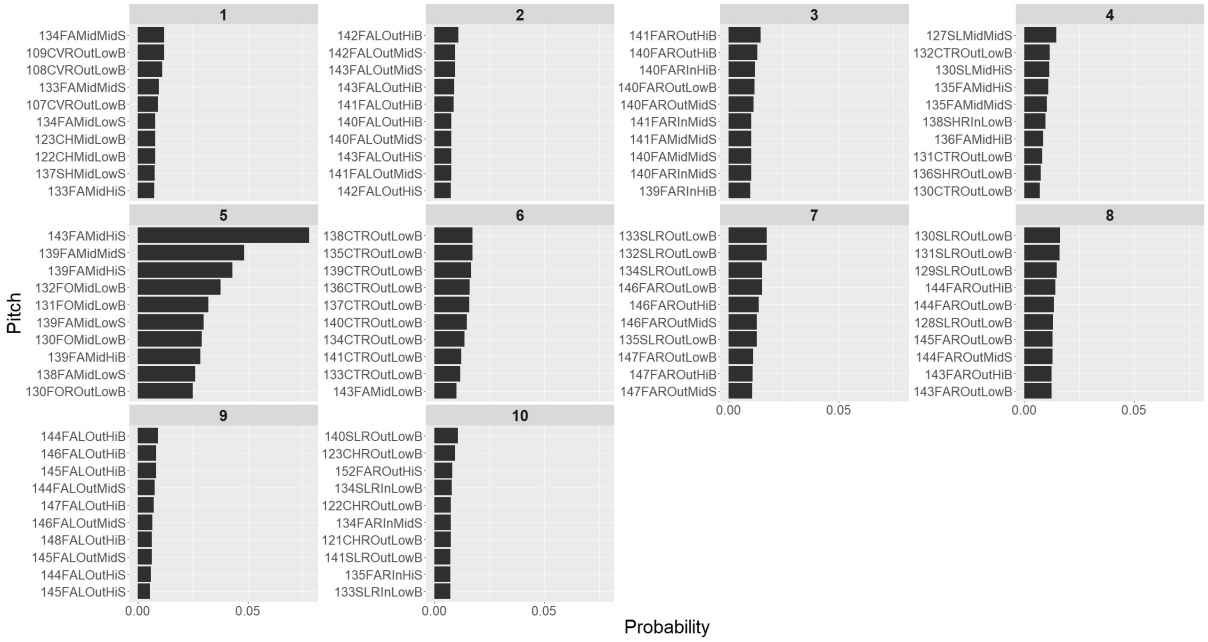

Figure 6 Word distribution for each topic

to jam a hitter or to keep the ball off the barrel of the bat. In contrast, against left-handed hitters, the batteries tend to heavily utilize Topic 2 or the pitch sequence intended to prevent the hitter from pulling a ball by keeping the ball outside. From the right side of Figure 8, it can be observed that such pitch sequences are generally used in the plate appearances resulting in double plays compared to other plate appearances.

Finally, the pitch sequence trend based on hitting results is shown in Figure 9. The left side of the figure shows the comparison of plate appearances resulting in swinging strikeouts and weak mishits. The right figure shows the comparison of plate appearances resulting in single and extra-base hits. In the plate appearances resulting in swinging strikeouts, Topics 7 and 9 or the pitch sequences intended to trick hitters into swinging at a bad pitch are heavily used, compared to those resulting in weak mishits. Because the number of swinging strikeouts is approximately 3.4 times greater than that of called strikes in our data, such pitch sequences may be effective in cases where the batteries want strikeouts. Moreover, Topic 3 is strongly used in plate appearances resulting in extra-base hits compared to those resulting in single hits. In other words, the pitch sequence that paints the strike zone's edges is likely to get extra-base hits. Because the probability of achieving long hits increases if the ball catches much of the plate, the pitch sequences intended to keep the ball off the barrel of the bat (Topic 6) or trick hitters into swinging at a bad pitch (Topic 7) may be effective in inhibiting hitters from making a long hit.

6.3 Trend of an Individual Pitcher's Pitch Sequence

This subsection details the pitch sequence trend of individual pitcher Takahiro Norimoto of the Tohoku Rakuten Golden Eagles. He is the pitcher who faced the highest number of hitters in the 2016–2018 seasons. It can be seen that his pitch sequence essentially consists of Topics 7, 8, and 9 through three seasons, as shown in Figure 10. In other words, he is likely to employ pitch sequences intended to trick hitters into swinging at a bad pitch by throwing a fastball above the hands and a low slider out of the

Electronic copy available at: https://ssrn.com/abstract=3728430

strike zone. Thus, he is considered the type of pitcher to aggressively strike out a hitter. Next, his pitch sequence against a specific hitter, Kenta Imamiya of Fukuoka SoftBank Hawks, is considered. He is the hitter most faced by Norimoto throughout the 2016–2018 seasons. As shown on the left side of Figure 11, Norimoto additionally used Topic 10 against Imamiya in the 2018 season, although his pitch sequence consisted of Topics 7 and 8 in the 2016–2017 seasons. It can be observed that he makes a combination of fast and slow pitches by adding a changeup to the preceding pitch sequences consisting of a fastball and a slider. Moreover, as shown on the right side of Figure 11, the sum of fractions of strong mishits, single hits, and extra-base hits increased from the 2016 to 2017 seasons. Given this fact, the change in Norimoto's pitch sequence in the 2018 season is considered to be the response to the fact that his fastball and slider get to be crushed by Imamiya year by year. The right side of Figure 11 also shows that the number of strikeouts significantly increased from the 2017 to 2018 seasons, while the sum of fractions of single and extra-base hits increase every year. This fact might indicate that Imamiya had a greater opportunity to swing ahead of a pitch owing to the addition of a changeup while he gradually adjusted to Norimoto's fastball and slider.

7 Conclusion and Future Work

This study attempted to extract the pitch sequence trend based on pitcher/hitter characteristics, game situations, and hitting results by applying the probabilistic topic model to pitch-by-pitch data on all regular season games of NPB for the period 2016–2018. It is found that there exists a pitch sequence pattern that is common to a specific hitter's characteristics, game situations, and hitting results. Using the probabilistic topic model, it is possible to (1) analyze a large amount of pitching data in a single framework at a time, and (2) deeply investigate the correlations between pitch sequencing and diverse game-related factors. The findings of this study have practical implications for both batteries and hitters. Although there are no correct or incorrect pitch sequences, as mentioned in section 1, a battery would be able to increase the probability of setting hitters down by exploring the correlation between pitch sequencing and hitting results. In addition, the model can effectively extract the pitch sequence trend of a specific pitcher and in a specific match-up, as shown in subsection 6.3. This finding is beneficial to hitters. For example, a hitter would be able to study the pitch-sequencing strategy of a pitcher who is scheduled to face in the next game by exploring his pitch-sequencing patterns, in particular, in the past match-ups.

In this work, each plate appearance was defined as a simple set of pitches under the BOW assumption. However, in reality, the order of pitches is meaningful in pitch sequences. For example, it may be effective to pitch a slider low and outside after pitching a fastball high and inside. In future work, the order of pitches should be considered. Moreover, the metadata utilized herein are only part of the information considered by catchers when developing pitch-sequencing strategies. The catchers usually develop these strategies by observing a more diverse set of elements including the hitter's physical attributes, power, strengths and weaknesses, and hitting stance. Adding such information to the metadata in the model may help to more deeply understand the pitch sequences.

Electronic copy available at: https://ssrn.com/abstract=3728430

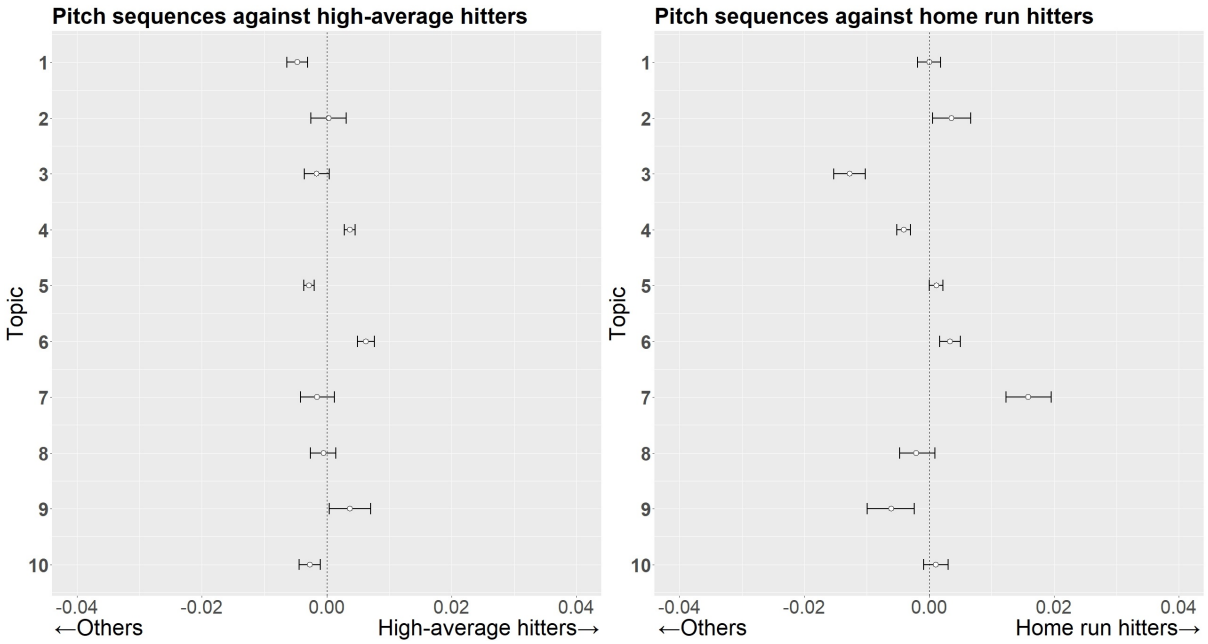

Figure 7 Trend of pitch sequences against sluggers

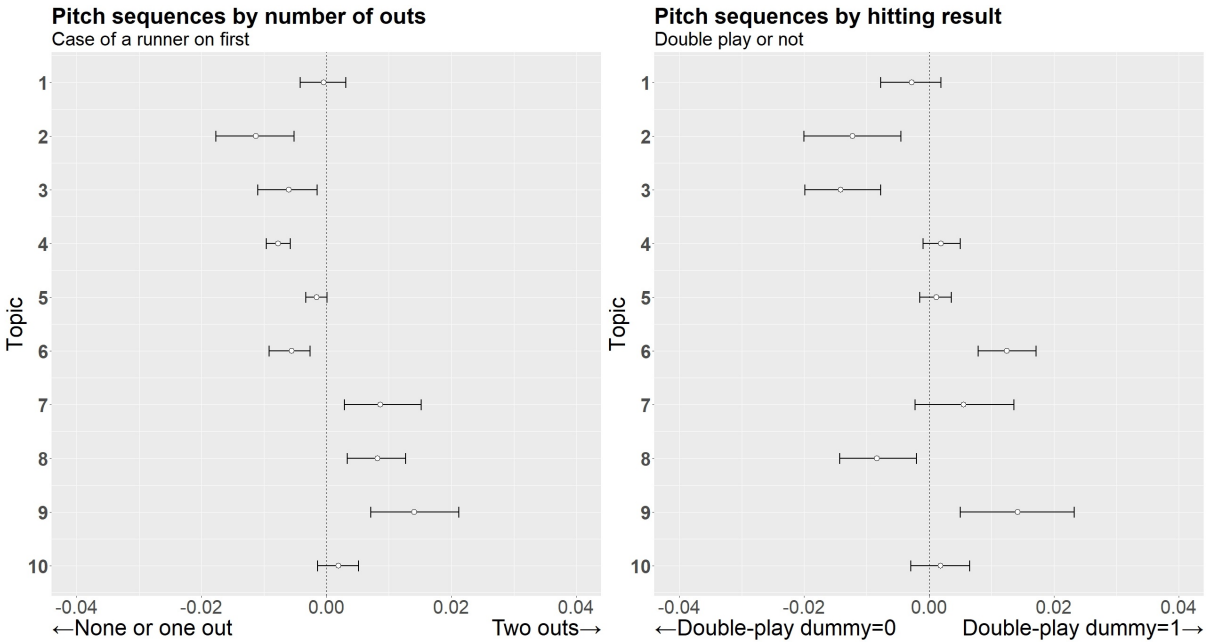

Figure 8 Trend of pitch sequences by number of outs

Electronic copy available at: https://ssrn.com/abstract=3728430

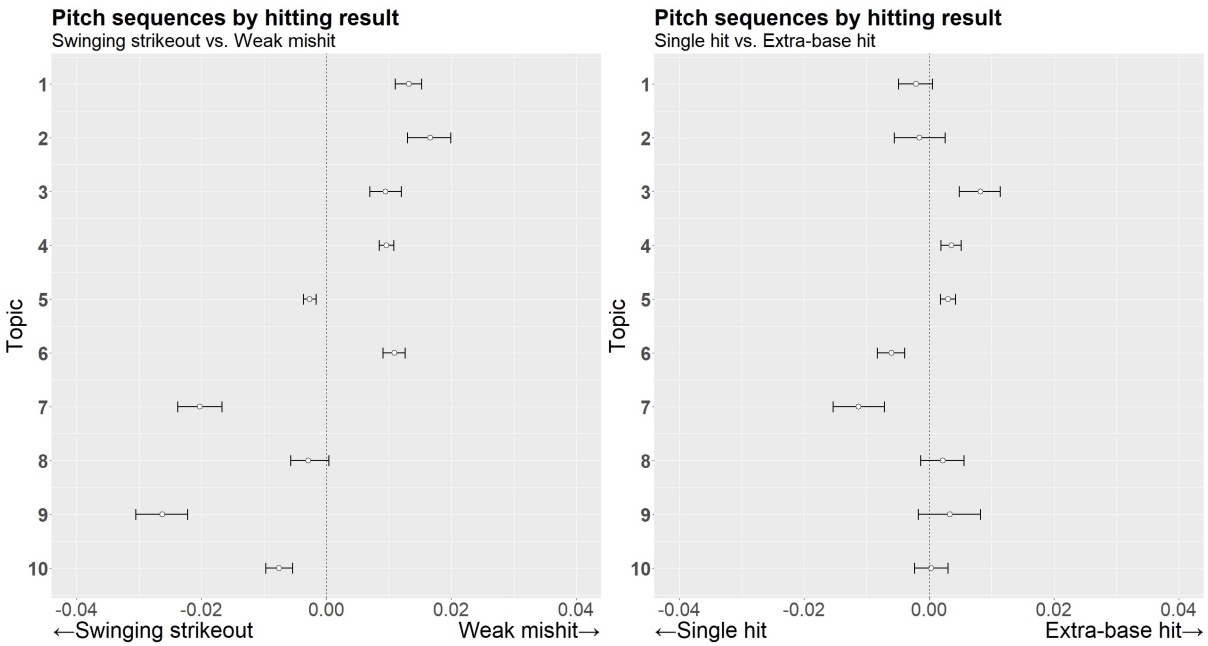

Figure 9 Trend of pitch sequences by hitting results

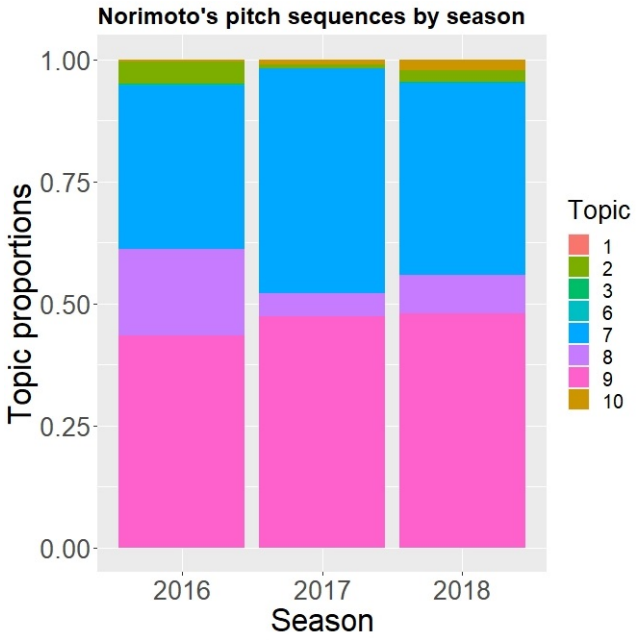

Figure 10 Pitch sequences of Takahiro Norimoto by season

Electronic copy available at: https://ssrn.com/abstract=3728430

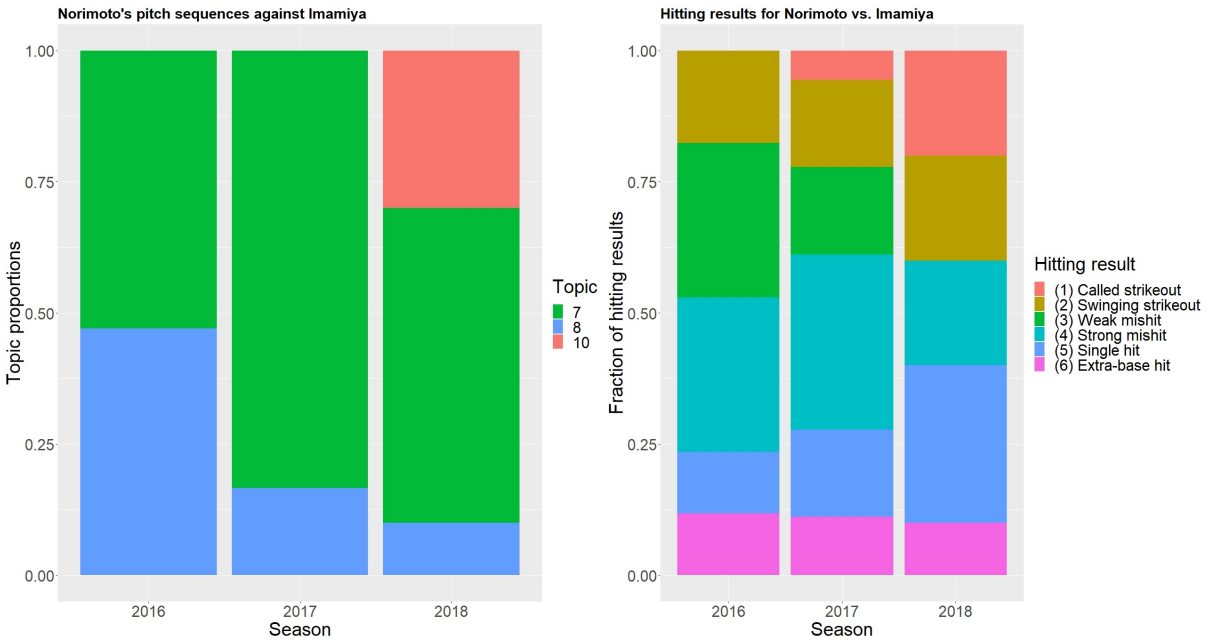

Figure 11 Trend of pitch sequences: Takahiro Norimoto vs. Kenta Imamiya

Electronic copy available at: https://ssrn.com/abstract=3728430

References

[1] Blei, D. M.: “Probabilistic Topic Models,” *Communications of the ACM*, Vol. 55, No. 4, pp. 77-84 (2012)

[2] Blei, D. M. and Lafferty, J. D.: ”A Correlated Topic Model of Science,” *The Annals of Applied Statistics*, Vol. 1, No. 1, pp. 17-35 (2007)

[3] Blei, D. M., Ng, A. Y. and Jordan, M. I.: “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022 (2003)

[4] Bonney, P.: ”Defining the Pitch Sequencing Question,” https://tht.fangraphs.com/defining-the-pitch-sequencing-question/ (2015)

[5] Eisenstein, J., Ahmed, A. and Xing, E. P.: ”Sparse Additive Generative Models of Text,” *Proceedings of the 28th International Conference on Machine Learning*, pp. 1041-1048 (2011)

[6] Glaser, C.: ”The Influence of Batters’ Expectations on Pitch Perception,” https://tht.fangraphs.com/tht-live/the-influence-of-batters-expectations-on-pitch-perception/ (2015)

[7] Gray, R.: ”Behavior of College Baseball Players in a Virtual Batting Task,” *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 28, No. 5, pp. 1131-1148 (2002)

[8] Gray, R.: ””Markov at the Bat”: A Model of Cognitive Processing in Baseball Batters,” *Psychological Science*, Vol. 13, No. 6, pp. 542-547 (2002)

[9] Griffiths, T. L. and Steyvers, M.: “Finding Scientific Topics,” *Proceedings of the National Academy of Sciences*, Vol. 101, pp. 5228-5235 (2004)

[10] Grimmer, J. and Stewart, B. M.: “Text as Data: Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts,” *Political Analysis*, Vol. 21, No. 3 (2013)

[11] Healey, G. and Zhao, S.: ”Using PITCHf/x to Model the Dependence of Strikeout Rate on the Predictability of Pitch Sequences,” *Journal of Sports Analytics*, Vol. 3, pp. 93-101 (2017)

[12] Koseler, K. and Stephan, M.: ”Machine Learning Applications in Baseball: A Systematic Literature Review,” *Applied Artificial Intelligence*, Vol. 31, No. 9-10, pp. 745-763 (2017)

[13] Martin, E. P.: ”Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types,” *MIT Sloan Sports Analytics Conference* (2019)

[14] Mimno, D. and McCallum, A.: ”Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression,” *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 411-418 (2008)

[15] Roberts, M. E., Stewart, B. M., Tingley, D. and Airoldi, E. M.: “The Structural Topic Model and Applied Social Science,” *Neural Information Processing Society* (2013)

[16] Roberts, M. E., Stewart, B. M. and Tingley, D.: “stm: R Package for Structural Topic Models,” *Journal of Statistical Software*, Vol. 91, No. 2 (2019)

[17] Roegele, J.: ”The Effects of Pitch Sequencing,” https://tht.fangraphs.com/the-effects-of-pitch-sequencing/ (2014)

Electronic copy available at: https://ssrn.com/abstract=3728430