Application of VBGM for pitch type classification: analysis of TrackMan's pitch tracking data

Kazuhiro Umemura¹ · Toshimasa Yanai² · Yasushi Nagata¹

Received: 26 August 2019 / Accepted: 6 April 2020 / Published online: 27 April 2020
© The Author(s) 2020

Abstract

In the game of baseball, each pitcher throws various types of pitches, such as cutter, curve ball, slider, and splitter. Although the type of a given pitch may be inferred by audience and/or obtained from the TrackMan data, the actual pitch type (i.e., the pitch type declared by the pitcher) may not be known. Classification of pitch types is a challenging task, as pitched baseballs may have different kinematic characteristics across pitchers even if the self-declared pitch types are the same. In addition, there is a possibility that the kinematic characteristics of pitched baseballs are identical even if the self-declared pitch types are different. In this study, we aimed to classify TrackMan data of pitched baseballs into pitch types by applying the Variational Bayesian Gaussian Mixture Models technique. We also aimed to analyze the kinematic characteristics of the classified pitch types and indices related to batting performance while pitching each pitch type. The results showed that the pitch types could not be accurately classified solely by kinematic characteristics, but with consideration of the characteristics of the fastball the accuracy improves substantially. This study could provide a basis for the development of a more accurate automatic pitch type classification system.

Keywords Clustering · Machine learning · Sports statistics · Baseball · TrackMan

✉ Kazuhiro Umemura
284re-kazuhiro.u@akane.waseda.jp

Toshimasa Yanai
tyanai@waseda.jp

Yasushi Nagata
ynagata@waseda.jp

¹ Faculty of Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan

² Faculty of Sport Sciences, Waseda University, Tokorozawa, Saitama 359-1192, Japan

1 Introduction

Major League Baseball (MLB) witnessed the introduction of Pitch f/x for the first time in 2006. This is a device that instantly identifies the trajectory of a pitch or hit ball (Fast 2010). The evolved version of Pitch f/x, StatCast, was introduced in 2015 in all baseball stadiums (Sandomir 2015). The introduction of these devices has influenced the recent “Flyball Revolution” of the MLB and the extreme defensive shift. In Nippon Professional Baseball (NPB), Tohoku Rakuten Golden Eagles introduced TrackMan as part of StatCast for the first time in 2014, and in the 2018 season, 11 teams excluding Hiroshima Toyo Carp set TrackMan in the headquarters.

In the game of baseball, a pitcher strives to put batters out by throwing various pitch types. Although the information on the pitch type is obtained from the TrackMan data, the actual name of the pitch type (self-declared by the pitcher (Nagami et al. 2016)) is not known. Therefore, pitched baseballs may have different kinematic characteristics across pitchers even if their self-declared pitch types are the same. In addition, there is a possibility that the kinematic characteristics of pitched baseballs are identical even if the self-declared pitch types are different. Accurate classification of pitch types is a challenging task, but its outcomes could provide pitchers and coaches with objective information to illuminate the directions for improving existing pitch types, mastering new pitch types and developing strategies. For example, it will be possible to analyze the characteristics of pitch types and how these are similar to those of other pitchers, which will expand the range of tactics. To investigate such a possibility, we aim to objectively evaluate pitch types using TrackMan data.

According to Nathan (2008), the trajectory of a pitched baseball is determined by three forces: gravity, drag, and lift due to the Magnus effect. The angular velocity (spin rate), speed, and the direction of the spin axis of the pitched baseball greatly influence the drag and lift (Jinji and Sakurai 2007; Bahill and Baldwin 2007). The spin axis is often expressed by the azimuth and elevation angles in the polar coordinate system. Details are provided in the studies of Jinji and Sakurai (2007) and Nagami et al. (2011). Several studies (e.g., Nagami et al. 2016; Nagami et al. 2015; Whiteside et al. 2016) have analyzed the speed, spin rate, and spin axis to characterize each pitch type. In these studies, each pitcher declared the pitch types in advance and the obtained kinematic data of the pitched baseballs were analyzed to characterize the pitch types. Nagata and Minami (2017) quantitatively analyzed Pitch f/x data of the fastballs thrown by selected pitchers with high whiff rates and found that the horizontal break distance of the pitched baseball is an important factor.

In this study, we quantitatively classified TrackMan data of pitched baseballs into pitch types with a two-step approach: pre-classify pitchers into multiple classes of similar fastball characteristics and then apply a technique called the Variational Bayesian Gaussian Mixture Models (VBGMM) to classify pitched baseballs into pitch types for each pre-classified class. Moreover, we analyzed the kinematic characteristics of the classified pitch types and indices related to batting performance while pitching each pitch type. Furthermore, this study could provide a basis for the development of a more accurate automatic pitch type classification system.

2 TrackMan overview

TrackMan is a system developed by TRACKMAN of Denmark, and it serves to track pitching and hitting using a Doppler radar. It is widely used for tracking golf shots, and the parent company Interactive Sports Games is said to have applied the technology to baseball (Fast 2010). In NPB, the Tohoku Rakuten Golden Eagles used the system for the first time in 2014. Subsequently, in 2018, 11 teams excluding the Hiroshima Toyo Carp set up the system in baseball stadiums. The main data that can be obtained are as follows.

- Data on the game situation.
Date and time, pitcher, batter, count, and result, with a total of 35 variables.
- Data on pitching.
Speed, release position, and break distance, with a total of 29 variables.
- Data on hit ball.
Exit velocity, launch angle, distance, and hang time, with a total of 11 variables.

There are two ways of determining the pitch types, the “AutoPitchType” and the “TaggedPitchType”, according to TrackMan. In the former, the pitch type is automatically assigned. However, when comparing the two pitch types, the degree of coincidence of pitch types thrown by a special delivery style, such as the submarine style, was low. In the latter, the operator in the stadium assigns a pitch type to each pitch, and thereby there is a possibility of input errors.

In this study, data acquired by TrackMan in the NPB seasons of 2017 and 2018 seasons are used. For the 2017 season, 140,196 balls (469 games, 7 stadiums) are targeted, and 223,358 balls (738 games, 11 stadiums) are targeted in the 2018 season. In addition, we use the variable names of TrackMan data in this study.

3 Flow of analysis

The main analysis flow of this study is as follows:

1. Pre-classification of pitchers.
2. Classification into pitch types for each class.
3. Analysis of classified pitch types.
4. Application to next season data.

The outline of each analysis flow is described below.

3.1 Pre-classification of pitchers

A problem arises when one tries to classify an entire set of TrackMan data of various pitchers into pitch types with a single criterion. This is because even with the same pitch types, the spin axis of the ball may be different between right- and left-handed pitchers or between overhand and side arm delivery styles, inducing a substantial difference in the trajectory of the pitched baseballs. At times, a changeup thrown by a pitcher may be similar to a fastball of another pitcher. Therefore, prior to the classification of pitch types, the pitchers need to be classified into similar types.

The spin axis of a pitched fastball is determined primarily by the angle of the forearm and the direction of the palm at release (Nagami et al. 2011; Jinji et al. 2011), such that the differences in delivery style affect the characteristics of the fastball. Moreover, it is expected that the difference in delivery style also affects the difference in the characteristics of pitch types. Therefore, in this study, we defined that similar types of pitchers were synonymous with similar characteristics of the fastballs, and thus we classified pitchers based on the characteristics of the fastball prior to the classification of the pitch types.

Specifically, after stratifying the right- and left-handed pitchers, we classified every pitcher based on the four variables of the fastball, namely RelSpeed (RS; speed), SpinRate (SR; spin rate), InducedVerticalBreak (IVB; vertical break distance), and HorzBreak (HB; horizontal break distance), via the Ward hierarchical clustering method (Ward 1963). The direction of the spin axis also affects the trajectory of the pitched baseballs (Jinji and Sakurai 2007; Bahill and Baldwin 2007), but the variable on the spin axis obtained by TrackMan (SpinAxis) is the angle obtained by projecting the elevation angle onto the x - z plane. Therefore, in this study, the vertical and horizontal break distances which were closely related to the direction of the spin axis were added to the classification criteria. Further, Nathan (2018) published a formula for estimating the azimuth and elevation angle of the spin axis from StatCast and TrackMan data.

3.2 Classification into pitch types for each class

After having classified pitchers into multiple classes, we used VBGMM (Bishop 2006; Corduneanu and Bishop 2001) to classify TrackMan data of all pitchers in each pre-classified group into pitch types. The name assigned to pitch types in the procedure is to be described later. In addition to comparing the classification result and TaggedPitchType, we used the Random Forest method (Breiman 2001) to check whether the pitched baseballs of unknown type were classified correctly. We also verified the necessity of classifying pitchers in advance and how well the classified pitch types corresponded to the actual pitch types.

The actual pitch types were known only by the pitchers and catchers. Therefore, the evaluation of classification performance of this study was conducted based on the results of expert hearings.

3.3 Analysis of classified pitch types

To analyze the characteristics of each classified pitch type for each pitcher class, the average values of the speed and spin rate were calculated, along with the indices relating to the batting performance when the pitch type was thrown, such as the strike swinging rate and strike called rate, exit velocity, and the launch angle at the time of strike. The relevance to the batting performance of each pitch type was analyzed focusing particularly on two indices; expected weighted On-Base Average (xwOBA) and xwOBA on contact (xwOBAcon). Details of the indices are provided in Sect. 7.2. In addition, we calculated “Pitch Type Share”, which indicated the extent to which pitchers executed each pitch type, and clarified the peculiarity of the pitch type. Furthermore, we calculated xwOBAcon based on the presence or absence of the combination of pitch types and considered them.

3.4 Application to next season data

In this study, we used the data of the 2017 season to classify pitchers and the pitch types. Based on this classification result, we predicted the pitcher class in the 2018 season using Support Vector Machines (Cortes and Vapnik 1995; Vapnik and Lerner 1963) and predicted the pitch types using the Random Forest method. By doing this, classification results of the 2017 season were utilized to predict pitching data of the 2018 season.

4 Variational Bayesian Gaussian mixture models

4.1 Outline of VBGMM

In this study, we use VBGMM to classify the pitch types. Clustering based on VBGMM is a method that estimates the mixed Gaussian model by variational Bayes when the number of clusters is unknown, and classifies it into a class with the maximum responsibility when each item is provided. The contents of Sects. 4.1.1 and 4.1.2 are based on Bishop (2006).

4.1.1 Mixture of Gaussians and its Bayes model

For each observation $\mathbf{x}_n$, we have a corresponding latent variable $\mathbf{z}_n$ comprising a 1-of- K binary vector with elements z_{nk} for $k = 1, \dots, K$. Note that when the k th element of the latent variable $\mathbf{z}_n$ takes 1, the corresponding $\mathbf{x}_n$ is generated from the k th mixture element of the Gaussian mixture models.

Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be the observed data set and $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ be the latent variables. Then the conditional distribution of $\mathbf{Z}$ for a given the mixing coefficients $\boldsymbol{\pi}$, and the conditional distribution of $\mathbf{X}$ for a given set of latent variables and component parameters can be defined as Eqs. (1) and (2), respectively.

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}}, \quad (1)$$

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^N \prod_{k=1}^K N(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}, \quad (2)$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_k\}$, which are the mean vectors and covariance matrices of component parameters, respectively.

Next, we introduce priors over the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\pi}$. We choose a Dirichlet distribution over the mixing coefficients $\boldsymbol{\pi}$ as shown in (3).

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\alpha_0) = C(\alpha_0) \prod_{k=1}^K \pi_k^{\alpha_0-1}, \quad (3)$$

where $C(\alpha_0)$ is the normalization constant for the Dirichlet distribution defined by (4–6).

$$C(\alpha_0) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)}, \quad (4)$$

$$\alpha = \sum_{k=1}^K \alpha_k, \quad (5)$$

$$\alpha_k = \alpha_0 + N_k, \quad (6)$$

where N_k is a statistic evaluated with respect to the responsibilities. See Bishop (2006) for the concrete definition.

We introduce an independent Gaussian–Wishart prior governing the mean and precision of each Gaussian component, given by (7).

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^K N(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) W(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0), \end{aligned} \quad (7)$$

where $\mathbf{m}_0$, $\mathbf{W}_0$, β_0 , and ν_0 are hyperparameters, $\mathbf{m}_0$ is a D -dimensional vector, $\mathbf{W}_0$ is a symmetric positive semidefinite matrix, and β_0 and ν_0 are scalars that take positive values. In this study, as a special case of (7), we use Gauss–Wishart distribution, where $\mathbf{m}_0$ is the zero vector and $\mathbf{W}_0$ is an identity matrix.

4.1.2 Variational distribution

To estimate the Bayesian model, we can use stochastic techniques such as Markov chain Monte Carlo (MCMC). However, according to Bishop (2006), sampling methods, such as MCMC, are computationally demanding and often limit their use to small-scale problems. Therefore, in this paper, we introduce variational Bayes which is suitable for analyzing large-scale data.

To formulate a variational treatment of Gaussian mixture models, we use the joint distribution of all the random variables, which is given by (8).

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}). \quad (8)$$

Note that only the variables $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ are observed. Here, we assume that the latent variables and parameters are independent of each other. That is,

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \quad (9)$$

Under these assumptions, we derive an algorithm for estimating the Gaussian mixture models using the variational Bayes method. An algorithm for finding the optimal solution can be easily implemented by repeating the variational E and M steps in a similar manner to the EM algorithm of maximum likelihood estimation. See Bishop (2006) and Corduneanu and Bishop (2001) for estimation algorithms.

4.2 Reasons for adopting VBGMM

When deciding the method of classifying the pitch types, we checked that the following two points were satisfied.

- The method can automatically estimate the number of clusters.
- The method is suitable for pitched baseballs data.

Initially, this study adopted the k -means method (MacQueen 1967), which is one of the basic methods of clustering. However, in this method, it is known that the number of clusters k needs to be specified prior to analysis. Since the number of pitch types (clusters) is not determined in advance, we required a method capable of automatically estimating the number of clusters.

Figure 1 shows the break distance of TaggedPitchType of a pitcher. As Guha et al. (1998) highlights, we can see that this distribution resembles cases that are inappropriate for the k -means method. Therefore, Kamishima (2003) recommends that a method based on a probability model using the EM algorithm instead of the k -means method should be considered when there is a difference in the number of objects in the cluster. From the above viewpoint, we adopted VBGMM that uses the variational Bayes method, which extends the EM algorithm so that the number of clusters can be automatically estimated.

Fig. 1 Break distance of TaggedPitchType of Pitcher A

5 Pre-classification of pitchers

5.1 Overview

As mentioned in Sect. 3.1, the pitchers were classified based on the characteristics of the fastball prior to the classification of pitch types. After stratifying the pitchers according to the handedness, we performed cluster analysis for every pitcher based on four variables of the fastball, RelSpeed (speed), SpinRate (spin rate), InducedVertical-Break (vertical break distance), and HorzBreak (horizontal break distance), by using the Ward method. For an explanation of each variable, see TrackMan Overview.

220 right-handed pitchers and 85 left-handed pitchers who threw a fastball in the 2017 season were targeted. We standardized the four variables of the data of up to the 24th fastball thrown by each pitcher for each game and used the average value. We considered the possibility that the characteristic of the fastball changed because of fatigue as the pitched baseballs number increased during a game. Therefore, we use 24 balls, which was the average number of fastballs pitched by one pitcher in three innings. In this study, we classified the right-handed pitcher to one of the eight classes and the

left-handed pitcher to one of the five classes. The number of classes, however, was set by our own judgment. In this paper, only the result in the right-handed pitchers is discussed.

5.2 Calculation of correction value

Fast (2010) pointed out that an error based on the stadium existed in the value of the break distance obtained by the Pitch f/x system, and the same problem was also observed in the TrackMan data. Additionally, even if the stadium was the same, expert interviews indicated that the difference between right- and left-handed pitchers affected the error of the break distance. Therefore, we stratified the pitchers according to the handedness first, and then calculated and corrected the error of break distance (InducedVerticalBreak and HorzBreak) between the ball stadiums using as a reference a certain baseball stadium, via Hayashi's quantification method type I (Hayashi 1951). Specifically, we prepared the multiple regression model shown in (10) and estimated the regression parameter $\beta_0, \beta_{1(j)}, \beta_{2(k)}$ to minimize the residual sum of squares.

$$y_i = \beta_0 + \sum_j \beta_{1(j)} x_{i1(j)} + \sum_k \beta_{2(k)} x_{i2(k)} + \varepsilon_i, \quad (10)$$

where y_i is the InducedVerticalBreak or HorzBreak of the pitched baseball of the sample No. i , and $x_{i1(j)}$ ($j = 2, 3, \dots$), $x_{i2(k)}$ ($k = 2, 3, \dots$) are dummy variables for pitchers and stadiums, respectively. The estimated regression parameter $\beta_{2(k)}$ can be considered as an error at a specific baseball stadium ($k = 1$), such that the errors between the stadiums InducedVerticalBreak and HorzBreak are calculated and corrected.

5.3 Classification result

Table 1 shows the average values of RelSpeed (RS), SpinRate (SR), InducedVerticalBreak (IVB), and HorzBreak (HB) of fastball of pitchers belonging to each

Table 1 Fastball's characteristics of right-handed pitchers

Class	Pitcher	RS (mph)	SR (rpm)	IVB (in)	HB (in)
1	63	89.9	2090.4	17.0	8.8
2	30	89.9	2310.7	19.6	9.6
3	21	89.4	2109.7	12.5	13.8
4	15	90.3	2398.9	14.4	4.2
5	35	89.2	2177.0	20.5	4.8
6	34	93.5	2262.4	19.2	7.7
7	20	93.5	2361.6	15.3	11.2
8	2	78.6	1979.2	3.6	8.3

RS RelSpeed, SR SpinRate, IVB InducedVerticalBreak, HB HorzBreak, mph mile per hour, rpm revolutions per minute, in inch

class in the right-handed pitcher. As shown in Table 1, the categorization is based on features such as class with high speed and class with large vertical break distance. This is considered as a reasonable classification result in light of expert knowledge. However, as previously mentioned, it should be noted that this classification only focuses on the characteristics of the fastball and does not consider the combination of the pitch types of the pitcher. Hereafter, this classification result is used as a pitcher class.

6 Classification into pitch types for each class

6.1 Overview

In this section, we describe the method for classifying the pitch types of all pitchers for each class based on VBGMM. In this paper, only the result of the Class 1 in right-handed pitchers is discussed.

6.2 Classification procedure

With respect to classification of pitch types based on VBGMM, the following 14 variables were used:

RelSpeed, VertRelAngle, HorzRelAngle, SpinRate, SpinAxis, ZoneSpeed, VertApprAngle, HorzApprAngle, vx0, vy0, vz0, ax0, ay0, az0.

The definitions of these variables are described in TrackMan overview. Variables relating to position, such as different release points depending on the pitcher and the break distance that appeared as a result, were excluded. With respect to the 14 variables, standardized values were used for classification. The flow of the analysis was as follows:

1. Standardize the above variables and execute VBGMM-based classification 100 times.
2. Count the frequency of the number of pitch types (clusters) and determine the most frequent number as the number of pitch types of the class. At this time, clusters of 30 balls or less (Class 8 in the right-handed pitchers corresponds to ten balls or less, because the number of pitchers belonging to this class is low) are excluded.
3. Execute VBGMM-based classification until the classification result that becomes the most frequent cluster number is obtained in ten ways, because the classification result depends on the initial clusters, and assign the name to the pitch type.
4. With respect to the obtained ten types of classification results, learn all data with the name of pitch types by the Random Forest method, and that with the lowest Out-of-Bag (OOB) error rate (Breiman 2001) is adopted as the classification result of that class.

- Learn the data with the pitch type names by the Random Forest method, predict the pitch types of the pitched baseball included in the cluster excluded in step 2, and assign the name to the pitch types.

The method to obtain the name of pitch types in step 3 is as follows. We calculated the ratio occupied by the TaggedPitchType in each pitch type and defined the name of pitch types of the TaggedPitchType with the highest ratio and the class obtained in Sect. 3 with “_ (underscore)” as the “VBGMM Pitch Category”. “R” indicates right-handed pitchers, and “L” indicates left-handed pitchers. Additionally, we defined “VBGMM Pitch Type” by assigning numbers in ascending order from the pitch types with a high number of pitched baseballs in each Pitch Category after the underscore.

Fig. 2 Difference in the number of pitch types due to differences in hyperparameters

6.3 Hyperparameter setting in VBGMM

As shown in Sect. 4.1, when setting the prior distribution of parameters in VBGMM, it is necessary to determine hyperparameters (Bishop 2006; Corduneanu and Bishop 2001).

We changed the α_0 (alpha) to 0.001, 1.0, and 10.0; the β_0 (beta) to 0.1, 0.5, and 1.0; and the ν_0 (nu) to 20, 25, and 30, and classified based them on VBGMM 50 times for the pitch types of Class 1 in right-handed pitchers in each combination. Figure 2 summarizes the frequency of the number of pitch types (clusters). As Bishop (2006) stated, the smaller the value of α_0 , the smaller the number of pitch types tended to be. The difference in β_0 and ν_0 values seemed to have no big influence on the result. There was no big difference in either value, but in view of the opinion of experts, the values of α_0 , β_0 , and ν_0 were set to 0.001, 1.0, and 30.0, respectively.

6.4 Classification result

According to the classification procedure described in Sect. 6.2, the pitched baseballs of 63 right-handed pitchers in Class 1 were classified based on VBGMM 100 times, and as a result the most frequent classification contained nine pitch types. Table 2 compares VBGMM Pitch Category and VBGMM Pitch Type attached to the classification result with the lowest OOB error rate and TaggedPitchType obtained from TrackMan data for each pitched baseball. Figures 3 and 4 show the relationship between the break distance and speed for the VBGMM Pitch Type.

As shown in Figs. 3 and 4, even though variables of the break distance were not incorporated in the model, the break distance and speed of each VBGMM Pitch Type exhibited a very close distribution. The break distance between Curveball_R1_1 and Slider_R1_3 was very similar; however they were classified as different pitch types, meaning that the pitch type is not determined by the break distance only. Furthermore, as shown in Table 2, a VBGMM Pitch Type was not occupied by a specific TaggedPitchType. This means that there are cases where it can be regarded as the same pitch type with respect to kinematic characteristics, even though the conventional name of pitch type is under a different name. Moreover, in the VBGMM Pitch Category, the pitch types assumed to be Slider_R1 were classified into three pitch types, which indicates that it was better to judge pitch types differently, although they were conventionally considered as the same pitch type.

In the TaggedPitchType, some pitch types such as Curveball and Slider were distributed over a wide range. Nevertheless, the VBGMM-based classification classified the pitch types more finely than TaggedPitchType. By contrast, the pitch type judged as the Sinker by the TaggedPitchType was classified as Fastball_R1_1 or Splitter_R1_1 in the VBGMM Pitch Type. According to experts, the Sinker in the TaggedPitchType is generally assumed to be a kind of ball called the two-seam fastball, which however could not be clearly determined by the VBGMM-based classification. Nevertheless, as Slider_R1 and Curveball_R1 of the VBGMM Pitch Category were further classified, a difference was considered to exist in the kinematic

Table 2 Comparison of pitch types based on VBGM and TaggedPitchType of Class 1 in right-handed pitchers

VBGMM pitch category	VBGMM pitch type	n	TaggedPitchType								
			ChangeUp (%)	Curveball (%)	Cutter (%)	Fastball (%)	Other (%)	Sinker (%)	Slider (%)	Splitter (%)	Undefined (%)
ChangeUp_R1	ChangeUp_R1	1056	47.3	0.6	0.9	0.5	4.2	0.7	3.7	42.1	0.0
Curveball_R1	Curveball_R1_1	1651	0.1	78.4	0.1	0.1	0.1	0.0	21.2	0.1	0.0
	Curveball_R1_2	537	5.4	70.9	2.2	0.0	0.0	1.5	17.7	2.2	0.0
Cutter_R1	Cutter_R1_1	1045	0.0	0.0	52.2	38.0	0.0	3.5	6.1	0.2	0.0
Fastball_R1	Fastball_R1_1	13,787	0.3	0.0	0.6	87.0	0.1	11.7	0.1	0.2	0.0
Slider_R1	Slider_R1_1	2783	1.8	0.3	20.9	2.5	0.1	1.4	67.8	5.2	0.0
	Slider_R1_2	1575	0.3	35.6	0.8	0.1	0.1	0.0	62.2	1.1	0.0
	Slider_R1_3	680	0.0	3.8	0.3	0.0	0.0	0.0	95.9	0.0	0.0
Splitter_R1	Splitter_R1_1	4437	24.8	0.0	1.0	5.7	0.4	14.0	1.6	52.3	0.0

Fig. 3 Vertical and horizontal break distance of the VBGM pitch type

characteristics of these pitch types above the difference between the sinker and the fastball.

6.5 Prediction of pitch types

In this section, we investigated whether the classification result of this study could be used to predict pitch types. This was the basis of the development of the automatic pitch type classification system. The flow of analysis was as follows:

1. By VBGM-based classification, assign the VBGM Pitch Type to all pitching data, and this pitch type name is taken as the correct answer data.
2. Randomly divide 80% of all pitching data of each pitcher into learning data and the remaining 20% into test data.
3. Learn the pitching data of all pitchers by the Random Forest method.
4. Predict the VBGM pitch type of the test data and calculate the Accuracy, defined below, for each pitcher. At the same time, the part corresponding to the

Fig. 4 Vertical break distance and speed of the VBGM pitch type

VBGM Pitch Category part is extracted from the VBGM pitch type, and the Accuracy in the VBGM Pitch Category is also calculated.

- Repeat steps two and three 100 times, and calculate the average of the Accuracy of VBGM pitch type and VBGM pitch category for each pitcher.

Fig. 5 Accuracy of all pitchers belonging to each class in the right-handed pitcher group

However, this analysis was done for each pitcher class. In addition, in step three, the objective variable was the VBGMM Pitch Type and the explanatory variables were the same as those shown in Sect. 6.2. Equation (11) shows how to obtain Accuracy when the number of test data is n and the number of correct prediction results is c .

$$\text{Accuracy} = c/n. \quad (11)$$

Figure 5 shows the result of the Accuracy of all pitchers belonging to each class in the right-handed pitcher group in a box plot. In any class, 90% or more of the pitch types of several pitchers were accurately predicted. In addition, the Accuracy of the VBGMM Pitch Category was even higher than the Accuracy of the VBGMM Pitch Type, such that if this was a broad classification, it was classified with high accuracy.

6.6 Necessity to classify pitch types by class

We have classified the right-handed pitchers into eight classes and the left-handed pitchers into five classes. We examined the need for this process. The flow assuming to predict the name of pitch types from the pitching data of the pitchers belonging to Class A in the criterion of Class B was as follows:

1. Learn the pitching data of Class B as the objective variable of the VBGMM Pitch Type by the Random Forest method.
2. Based on the model learned in step one, predict the name of pitch types from the pitching data of Class A. That is, Class B's VBGMM Pitch Type is assigned.
3. Extract the part corresponding to Category and make it the "Predicted Pitch Category."
4. As the pitcher of Class A already has its own VBGMM Pitch Category, compare the VBGMM Pitch Category and Predicted Pitch Category.

In this section, the result of predicting and comparing the pitch types of the pitcher in Class 1 in the right-handed pitcher group with the criterion of Class 6, whose fastball break distance was comparatively similar, is shown in Table 3. Many

Table 3 Fastball's characteristics of right-handed pitchers

		Predict (_R6)			
		Curveball	Fastball	Slider	Splitter
Correct (_R1)	Curveball	6	0	14	1036
	ChangeUp	981	4	674	529
	Cutter	0	696	346	3
	Fastball	0	13,779	4	4
	Slider	974	63	3776	225
	Splitter	0	1009	17	3411

Table 4 Pitch type data of Pitcher A in magazine

	%	Avg. Speed (mph)
Fastball	51.9	88.9
ChangeUp	18.1	79.6
Curveball	17.1	68.4
Slider	12.9	83.2

Table 5 VBGMM pitch type data of Pitcher A

	%	Avg. Speed (mph)
Fastball_R51_1	56.1	89.2
ChangeUp_R5_1	17.4	81.4
Curveball_R5_1	19.2	68.6
Slider_R5_1	6.70	83.2
Slider_R5_2	0.20	80.2
Slider_R5_3	0.20	83.3
Splitter_R5_1	0.12	76.8

of the pitched baseballs judged as Curveballs based on Class 1 were judged as Splitters according to the criterion of Class 6. Even when pitch types were predicted by the criterion of any class, the same tendency was found, and few were classified under the same pitch type name. Therefore, it was necessary to classify by pitcher class.

6.7 Comparison of classification result and real pitch types

By comparing with actual pitch types information, we verified the accuracy of the pitch types classification based on VBGMM. As mentioned in Sect. 3.2, only the pitchers and catchers knew the actual name of the pitch types. “Weekly Baseball (issue of August 6, 2018)” featured pitch types and described the pitch types and their percentages as thrown by pitchers representing the NPB. These were actually based on an interview with the players and could be said to hold high credibility.

In this paper, we targeted one right-handed pitcher (Pitcher A) who was featured in the magazine. However, note that the data described in the magazine was the pitch type data of the 2018 season (as of July 22, 2018), and the classification result of this study was the result of using the TrackMan data of the 2017 season. In the classification of pitchers in Sect. 5, Pitcher A was classified as Class 5 in the right-handed pitchers.

For Pitcher A, Table 4 shows the data of only the pitch types thrown more than ten balls in the 2018 season described in the magazine. Table 5 shows the

Fig. 6 Break distance of VBGM pitch type of Pitcher A

classification result of the pitch types based on VBGM. Figures 1 and 6 show the break distance of Pitcher A's Tagged Pitch Type and the VBGM Pitch Type. The pitch types whose legend was “---” indicated that the pitcher A did not pitch the pitch types thrown by other pitchers belonging to the class. As for this pitcher class, as a result of VBGM-based classification of the pitch types, their pitched baseballs were classified into seven pitch types, and Table 5 indicates that Pitcher A also threw the same seven pitch types. However, only four pitch types, Fastball_R5_1, ChangeUp_R5_1, Curveball_R5_1, and Slider_R5_1 accounted for 99% of the pitched baseballs, which was close to the pitch types described in the magazine. Furthermore, it could be said that pitch types were properly classified even by looking at the relationship of the break distance. From Fig. 6, the distribution of the break distance of Slider_R5_2 and Slider_R5_3, which were a small number, deviated from the distribution where Slider_R5_1 spreads widely, and there is a possibility that these pitched baseballs were careless pitch for Pitcher A. Comparing Table 4 with Table 5, there were some pitch types whose pitching ratios were not close; however, considering that the data year differs, this does not pose a problem.

A similar analysis was conducted for one pitcher belonging to Class 3 in the left-handed pitcher group, and this result was also similar to the actual pitch types described in the magazine. The VBGMM-based classification of the pitch types can thus determine the actual pitch type with high accuracy.

7 Analysis of classified pitch types

7.1 Overview

In this section, we analyzed the characteristics of the VBGMM Pitch Type. We calculated the average speed, spin rate, and index related to the batting result. The batting result was analyzed by specifically focusing on xwOBA and xwOBAcon. In addition, we calculated the “Pitch Type Share”, which indicated the extent to which pitchers occupy each pitch type. Furthermore, we examined whether a difference exists in xwOBAcon in the presence or absence of a combination of the VBGMM Pitch Type. We calculated the importance (Breiman et al. 1984) of each variable and illustrated the difference in the top three variables for each pitch type. In this paper, only the result of Class 1 in the right-handed pitcher group is shown.

7.2 wOBA, xwOBA, and xwOBAcon

The weighted On-Base Average (wOBA) is an index proposed by Tango et al. (2007) and is an extension of the On-Base Percentage (OBP). OBP is defined in (12).

$$\text{OBP} = (\text{H} + \text{BB} + \text{HBP})/(\text{AB} + \text{BB} + \text{HBP} + \text{SF}), \quad (12)$$

where H is the number of hits, BB is the number of bases on balls (walks), HBP is the number of hits by pitches, AB is the number of at bats, and SF is the number of sacrifice flies.

Regarding hits at OBP, the value is the same whether it is a single hit or a home run. For OBP, wOBA is an index that weights the score value for each batting result. The coefficients of the calculation formula of wOBA are different between MPB and NPB, and wOBA in NPB is defined in (13) (1.02 Essence of Baseball).

$$\text{wOBA} = \frac{\left\{ \begin{array}{l} 0.692 \times (\text{BB} - \text{IBB}) + 0.73 \times \text{HBP} + 0.966 \times \text{RBOE} + \\ + 0.865 \times \text{1B} + 1.334 \times \text{2B} + 1.725 \times \text{3B} + 2.065 \times \text{HR} \end{array} \right\}}{\text{AB} + \text{BB} - \text{IBB} + \text{HBP} + \text{SF}}, \quad (13)$$

where IBB is the number of intentional walks, RBOE is the number of reached base on errors, 1B is the number of singles, 2B is the number of doubles, 3B is the number of triples, and HR is the number of home runs.

xwOBA (expected wOBA) is an index that uses a value calculated from the exit velocity and the launch angle as opposed to the batting result in wOBA. The

Fig. 7 Hit ball launch angle dependence on exit velocity in wOBA of MLB data

calculation method of xwOBA is not disclosed. Thus, in this study, the value of xwOBA obtained by MLB data (baseballsavant.com) was learned by the Random Forest method, and the xwOBA for the data of this study was predicted. In the total pitching data (32,975 pitched baseballs) of the top ten pitchers with a large number of pitched baseballs in the 2018 season of MLB, data in which a hit ball has occurred and xwOBA was calculated and set as a target of learning data. Figures 7 and 8 show xwOBA of the acquired MLB data and xwOBA of the estimated NPB data in terms of exit velocity and launch angle, respectively. From these figures, the estimation accuracy can be seen to have no problem.

xwOBAcon (xwOBA on Contact) is calculated for xwOBA only in the scene where a hit ball occurs.

These indices are calculated based on the hit ball information. Therefore, these evaluation indices have been attracting attention in recent years, because these eliminate elements that the pitcher cannot control, such as fielder defense, stadium characteristics, and luck, and can measure the pitcher's true ability. For MLB data, there are many articles that evaluate the players' success by paying attention to xwOBA and xwOBAcon, such as Beneventano et al. (2012),

Fig. 8 Hit ball launch angle dependence on exit velocity in estimated xwOBA of NPB data

Edwards (2018), and Sullivan (2018). However, for the NPB data, the analysis using these indices is merely an individual's hobby.

7.3 Analysis of VBGMM pitch type

Table 6 shows a few calculated indices related to each VBGMM Pitch Type of Class 1 in right-handed pitcher group. However, the speed and the spin rate represent the average values when the pitch types were thrown, and the average value of the exit velocity and the launch angle were calculated only from the ball hit to the fair zone. In addition, Table 7 shows the top five "Pitch Types Share" which indicates the extent to which pitchers occupy each pitch type.

From Table 6, Slider_R1_3 was said to have been thrown 680 balls by two pitchers, but from Table 7, 679 balls were thrown only by one pitcher (Pitcher R1_A) and the Pitch Types Share was very high. Experts evaluated that this pitcher's Slider was a very special pitch type, so it was consistent with the evaluation that the Pitch Types Share of Slider_R1_3 was high in this study.

Table 6 Indices for each VBGM pitch type of Class 1 in right-handed pitchers

VBGM pitch type	Pitcher	<i>n</i>	%	RS (mph)	SR (rpm)	SC% (%)	SS% (%)	EV (mph)	LA (°)	xwOBA	xwOBAAcon
ChangeUp_R1_1	40	1056	3.8	78.2	1254.5	9.7	17.9	79.7	6.7	0.226	0.292
Curveball_R1_1	44	1651	6.0	71.9	2337.3	24.4	6.4	82.2	13.3	0.307	0.325
Curveball_R1_2	29	537	1.9	82.8	2115.8	14.9	13.0	82.3	1.7	0.303	0.370
Cutter_R1_1	35	1045	3.8	87.9	2142.8	14.4	6.6	80.4	3.5	0.310	0.291
Fastball_R1_1	63	13,787	50.0	90.2	2110.3	18.4	4.7	87.1	9.7	0.339	0.336
Slider_R1_1	57	2783	10.1	83.4	2153.6	18.0	12.0	83.5	8.6	0.292	0.316
Slider_R1_2	51	1575	5.7	80.3	2251.0	18.3	11.1	83.1	9.2	0.289	0.313
Slider_R1_3	2	680	2.5	80.8	3135.7	24.1	10.7	86.1	6.4	0.274	0.394
Spitter_R1_1	61	4437	16.1	81.1	1519.4	10.1	14.0	82.2	5.5	0.270	0.316

RS RelSpeed, *SR* SpinRate, *SC*% percent of strike called, *SS*% percent of strike swinging, *EV* exit velocity, *LA* launch angle, *mph* mile per hour, *rpm* revolutions per minute, °degree

Table 7 Pitch types share of Class 1 of right-handed pitchers

VBGMM pitch type	Pitcher	<i>n</i>	%
Slider_R1_3	Pitcher_R1_A	679	99.9
Curveball_R1_2	Pitcher_R1_B	295	54.9
Cutter_R1_1	Pitcher_R1_C	279	26.7
ChangeUp_R1_1	Pitcher_R1_D	208	19.7
Slider_R1_2	Pitcher_R1_B	305	19.4

Table 8 Difference in xwOBAcon by combination of pitch types

TPT	CPT	Y_N	xwOBAcon	<i>p</i> value
Splitter_R1_1	Curveball_R1_1	0	0.254	0.004
		1	0.331	
Splitter_R1_1	ChangeUp_R1_1	0	0.277	0.015
		1	0.334	
Curveball_R1_1	Curveball_R1_2	0	0.284	0.023
		1	0.370	
ChangeUp_R1_1	Curveball_R1_1	0	0.549	0.036
		1	0.268	
Slider_R1_2	Slider_R1_3	0	0.316	0.045
		1	0.136	

*TPT*target pitch type, *CPT*compare pitch type

Here, attention should to be paid to each index in Table 6 for Slider_R1_3 with high Pitch Types Share. The value of xwOBA of Slider_R1_3 was not only the lowest among Slider_R1, but also low among all pitch types, thus it can be judged as an excellent pitch type. By contrast, the xwOBAcon calculated only for the scenario where a ball is hit is the highest in all pitch types. Therefore, it is a pitch type that can get many strikeouts, while being considered as a dangerous pitch type if hit. As shown in Table 6, the strike swinging and strike called rates were at high levels, but the fast exit velocity affects large xwOBAcon when compared with other pitch types.

Welch's *t* test was performed for the presence or absence of a combination of pitch types, and Table 8 shows the combinations where the difference of xwOBAcon is significant with a significance level of 5%. It is possible to determine whether the Target Pitch Type (TPT) is a pitch type that can be easily hit, whether or not the pitcher exhibits a Compare Pitch Type (CPT) (Y_N = 1 or 0).

Many pitchers belonging to this class (61 out of 63) threw Splitter_R1_1, but pitchers who did not exhibit Curveball_R1_1 or ChangeUp_R1_1 had lower xwOBAcon values. The value of xwOBAcon of Slider_R1_2, whose pitcher that also throws Slider_R1_3, was also very low; however, there still were few pitchers throwing Slider_R1_3. For them, there was a possibility that Slider_R1_2 was a careless

Fig. 9 $az0$ of each VBGM pitch type. For an explanation of $az0$, see TrackMan Overview

Fig. 10 $RelSpeed$ of each VBGM pitch type. For an explanation of $RelSpeed$, see TrackMan Overview

Fig. 11 ax_0 of each VBGM pitch type. For an explanation of ax_0 , see TrackMan Overview

pitch, and possibly the value of $xwOBA_{con}$ may be low because of a different trajectory of the pitched baseball than usual. We need to pay attention not only to the combination of the two pitch types, but also to the combination of all pitch types, and further studies are needed to this end. Nevertheless, this analysis acts as a hint when a pitcher masters the new pitch types.

Resultantly, from the calculation of the variable importance of each variable, az_0 , $RelSpeed$, vy_0 , $SpinAxis$, $ZoneSpeed$, and $SpinRate$ were the top seven variables. It is understood that the speed and acceleration of pitched baseball greatly contribute to classification of pitch types. The $SpinAxis$ was also on a higher level, but as described in Sect. 3.1, this did not represent the exact direction of the spin axis. However, the fact that this $SpinAxis$ was arranged at the top of the variable importance, and that the variable importance of the $SpinRate$ was not low, certainly followed the discussion of Jinji and Sakurai (2007) and Bahill and Baldwin (2007) stating that the spin rate, speed, and the direction of the spin axis of the pitched baseball affected the trajectory of the pitched baseball.

Figures 9, 10, and 11 show the information of each pitch type with respect to the top three variables (az_0 , $RelSpeed$, and ax_0) of the variable importance. Curveball_R1_1 and Curveball_R1_2 had no big difference in az_0 , but Curveball_R1_2 had large $RelSpeed$, and ax_0 exhibited values close to zero. By contrast, ax_0 of Curveball_R1_1 indicated a positive value. In other words, we can predict that Curveball_R1_2 had a characteristic of falling vertically fast, and Curveball_R1_1 had a characteristic of falling while turning sideways at a slow speed. Such characteristics were also shown in Figs. 3 and 4. As for Slider_R1_1, Slider_R1_2, and Slider_R1_3,

there were no big differences in the speed, but differences in $az0$ and $ax0$ can be seen. In particular, `Slider_R1_3` had big absolute values for both $az0$ and $ax0$, and the batter will receive the impression that its trajectory changes most rapidly. From this analysis, `Slider_R1_3` could be considered as a very special pitch type.

8 Application to next season data

8.1 Overview

Until Sect. 7, we used the data of the 2017 season to classify pitchers, classified and analyzed the pitch types, and evaluated classification accuracy. In Sect. 6.5, the VBGMM Pitch Type was set as the correct data and it was possible to predict it accurately by verifying whether the name of pitch types can be correctly predicted within the same pitcher class. In this section, we examined whether classification results of the 2017 season were utilized to predict pitching data of the 2018 season. In this paper, only the results for the right-handed pitchers were shown.

8.2 Pitcher class prediction

Before predicting the pitch types for the 2018 season's pitching data, as in Sect. 5, we first classified the pitchers thrown in the 2018 season into each class (right-handed pitcher: eight classes, left-handed pitcher: five classes). We learned the class of pitcher based on the characteristics of fastball ($RelSpeed$, $SpinRate$, $InducedVerticalBreak$, and $HorzBreak$) in the 2017 season via Support Vector Machines and predicted the class of pitcher throwing in the 2018 season. In the 2018 season, 240 right-handed pitchers and 95 left-handed pitchers threw, out of which 165 right-handed pitchers and 69 left-handed pitchers also had a pitching record in the 2017 season.

As for the 165 right-handed pitchers who threw in both seasons in 2017 and 2018, Table 9 shows the class of the 2017 season and the predicted class of the

Table 9 Changes in pitcher class in the 2017 and 2018 season

		2018 season class							
		1	2	3	4	5	6	7	8
2017 season class	1	41	1	0	0	2	3	0	0
	2	5	18	0	1	2	2	0	0
	3	2	0	11	0	0	0	1	0
	4	0	1	0	8	1	1	0	0
	5	2	0	0	0	22	2	0	0
	6	2	1	0	0	6	16	0	0
	7	3	0	1	0	0	1	8	0
	8	0	0	0	0	0	0	0	1

Fig. 12 Pitch type predictions of Class 1 in right-handed pitchers in the 2018 Season

2018 season with the number of pitchers. Among the 165 pitchers, 126 pitchers were predicted to be in the same class, and 39 pitchers were predicted to be in different classes. Among these 39 pitchers, Pitcher R1_A mentioned in the previous section was also included. Their Slider_R1_3 had a high Pitch Type Share, so he was a very peculiar pitcher. While predicting the pitch types in a class different from the previous year, it is difficult to predict the pitch type in which the Pitch Types Share is high. Thus, one should discuss which class to predict for the pitch types of the pitcher whose class has changed from the previous year, and this is left to the judgment of the analyst.

8.3 Prediction of pitch types for pitchers with no pitching record in the 2017 season

There were 75 right-handed pitchers who had pitching records only in the 2018 season. We predicted the pitch types of these pitchers and evaluated the prediction performance by showing the relationship of the break distances. In this section, we

predicted the pitch types of 22 pitchers predicted as Class 1 among 75 right-handed pitchers.

The pitch type data of Class 1 in right-handed pitcher group of the 2017 season was learned by the Random Forest method and the pitch types of 22 pitchers were predicted. Figure 12 shows the relationship between the vertical and horizontal break distances of the predicted pitch types, exhibiting a distribution similar to that shown in Fig. 3, and thus the precision of the 2018 season's pitch types prediction using the data of the 2017 season seems effective. Additionally, when compared to Fig. 3, the number of plots of a few pitch types, such as Slider_R1_3 and Curveball_R1_2, was low. However, when compared with Table 7, the results indicated that these pitch types correspond to pitch types with a high Pitch Types Share. From this result, one can conclude that the classification that fully reflects the characteristics of the pitch types was completed.

9 Conclusion and future studies

9.1 Conclusion

In the game of baseball, the actual pitch type of a given pitch is determined by a self-declaration of the pitcher, and thus there is a possibility that kinematic characteristics can differ even with the same pitch type name. However, conventionally, if the name of the pitch types is the same, there is no further classification, and the difference in characteristics is left to subjective judgment. For example, if there are pitchers with unique pitch types, the expression “Pitcher A's slider” was frequently used.

In this study, we classified pitchers into classes depending on the characteristics of fastball, and the VBGMM-based classification was used to quantitatively classify the pitch types, which was conventionally subjective. The results showed that the pitch types were not classified only by the kinematic characteristics, but with consideration of the characteristics of the fastball that the pitcher throws. We assigned each pitched baseball with a pitch type based on the TaggedPitchType, and even if the VBGMM Pitch Category was the same, there were also pitch types that were further classified. To this point, we found that there were some pitch types with different kinematic characteristics even though they were considered to be the same pitch type conventionally.

In Sect. 7, we analyzed the characteristics of each classified pitch type. By associating the Sabermetrics indices, such as *xwOBA* and *xwOBAcon*, excellent pitch types and dangerous pitch types emerged if the balls were hit. We also calculated the difference of *xwOBAcon* between the presence and absence of a combination of pitch types, and we would think that this analysis could provide pitchers with a hint when a pitcher masters new pitch types. Furthermore, we calculated the variable importance, presented the variables contributing to the classification, and showed what kind of difference actually existed among the pitch types with the box plot.

In Sect. 8, based on the classification result of the 2017 season, we classified the pitch types in the 2018 season, considering that it could serve as the basis of development for the automatic pitch type classification system. The pitch types that had a high Pitch Type Share did not appear to be many in the classification of the pitch types of the pitchers throwing only in the 2018 season. Consequently, we believe that the classification of the pitch types in this study is a result that fully reflects the characteristics.

9.2 Future studies

Future studies include setting a reasonable class number, sophistication of the mechanism of predicting pitch types, and examination of other methods to classify pitch types.

In this study, we divided the right- and left-handed pitchers into eight and five classes, respectively, and analyzed the data. However, the number of classes was set independently. In the classification of pitch types, a method in which the number of clusters is automatically determined is used, but even in this case, a method in which the class number is automatically determined is recommended. Many of the methods for automatically determining the number of clusters, such as X-means (Ishioka 2000) and VBGMM used in this study, have the shortcoming that the result may depend on the initial clusters. We believe that these methods are not suitable in the classification of pitchers and set out to decide the class number beforehand and adopted the Ward method. As mentioned in Sect. 8.2, it is left to the analyst to decide which class is appropriate for the pitch types of the pitchers whose class has changed in the following year.

For example, it is of interest to analyze the combination of pitch types, such as pitch type B right after the pitch type A is hard to hit. In Sect. 7, the value of *xwOBA* of *Slider_R1_3* is low and the value of *xwOBACon* is high, so we considered it a pitch type that can obtain many strikeouts, while being also a dangerous pitch type if hit. Furthermore, if many of these pitch types were thrown to the slugger, their speed would tend to be faster. Thus, it may be worthwhile to analyze the pitch types depending on the opponents or the count. In addition, analyzing the characteristics of the pitch types of a pitcher who succeeds is also a very meaningful research topic.

In the prediction of pitch types in the following year using data of the previous year analyzed in Sect. 8, it is fully conceivable that a unique pitch type with a high Pitch Type Share will be thrown in the next season as pitch types with high Pitch Type Share existed in the previous season. However, in this study, the pitch type is always predicted as one of the pitch types that appeared in the previous year. If the probability of belonging to any of the pitch types is determined as low, then a mechanism is required such that the name of pitch types corresponds to “other”.

In this study, we obtained good results by using VBGMM as a method to classify pitch types. However, since there are many other clustering methods including UMAP and t-SNE, analysis and comparison using other methods may also be necessary.

Acknowledgements We would like to thank the anonymous referees for their valuable comments. This research was funded by Seibu Lions Co., Ltd. The publication of this data is approved by Seibu Lions Co., Ltd. and TrackMan. This work was partly supported by JSPS Grants-in-Aid for Scientific Research Grant no. 18K11202. We would like to thank Editage (www.editage.com) for English language editing.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bahill, A. T., & Baldwin, D. G. (2007). Describing baseball pitch movement with right-hand rules. *Computers in Biology and Medicine*, 37(7), 1001–1008.
- Baseball Savant: Trending MLB Players, Statcast and Visualizations. [baseballsavant.com](http://baseballsavant.mlb.com/). (2019). <https://baseballsavant.mlb.com/>. Accessed 01 February 2019.
- Beneventano, P., Berger, P. D., & Weinberg, B. D. (2012). Predicting run production and run prevention in baseball: The impact of sabermetrics. *International Journal of Business, Humanities and Technology*, 2(4), 67–75.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Freidman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification and regression trees. Wadsworth.
- Corduneanu, A., & Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. *Artificial intelligence and Statistics*, 2001, 27–34.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Edwards, C. (2018). FIP vs. xwOBA for assessing pitcher performance. Resource document. FanGraphs Baseball. <https://www.fangraphs.com/blogs/fip-vs-xwoba-for-assessing-pitcher-performance/>. Accessed 1 Feb 2019.
- Fast, M. (2010). What the Heck is PITCHf/x? *The Hardball Times Baseball Annual, 2010*, 153–158.
- Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. *ACM Sigmod Record*, 27(2), 73–84.
- Hayashi, C. (1951). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematical-statistical point of view. *Annals of the Institute of Statistical Mathematics*, 3(1), 69–98.
- INDEX. 1.02 Essence of Baseball. (2019). https://1point02.jp/op/gnav/glossary/gls_index.aspx. Accessed 1 Feb 2019.
- Ishioka, T. (2000). Extended K-means with an efficient estimation of the number of clusters. *Intelligent Data Engineering and Automated Learning IDEAL, 2000*, 17–22.
- Jinji, T., & Sakurai, S. (2007). Direction of spin axis and spin rate of the pitched baseball. *Sports Biomechanics*, 5(2), 197–214.
- Jinji, T., Sakurai, S., & Hirano, Y. (2011). Factors determining the spin axis of a pitched fastball in baseball. *Journal of Sports Sciences*, 29(7), 761–767.
- Kamishima, T. (2003). A Survey of recent clustering methods for data mining (part 1): Try clustering! *Journal of Japanese Society for Artificial Intelligence*, 18(1), 59–65. (in Japanese).
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14), 281–297.
- Nagami, T., Higuchi, T., Nakata, H., Yanai, T., & Kanosue, K. (2015). Relation between lift force and ball spin for different base-ball pitches. *Journal of Applied Biomechanics*, 32(2), 196–204.

- Nagami, T., Kimura, Y., Kanosue, K., & Yanai, T. (2016). Kinematic characteristics of various types of baseball pitches. *Japan Journal of Physical Education, Health and Sport Sciences*, 61(2), 589–605.
- Nagami, T., Morohoshi, J., Higuchi, T., Nakata, H., Naito, S., & Kanosue, K. (2011). Spin on fastballs thrown by elite baseball pitchers. *Medicine and Science in Sports and Exercise*, 43(12), 2321–2327.
- Nagata, D., & Minami, M. (2017). Factors Affecting batters' contact with a four-seam fastball. *Proceedings of the Institute of Statistical Mathematics*, 65(2), 185–200. (in Japanese).
- Nathan, A. M. (2008). The effect of spin on the flight of a baseball. *American Association of Physics*, 76(2), 119–124.
- Nathan, A.M. (2018). Determining the 3D Spin Axis from statcast data. Resource document. <https://baseball.physics.illinois.edu/trackman/SpinAxis.pdf>. Accessed 01 Feb 2019.
- Sandomir, R. (2015). Statcast arrives, offering way to quantify nearly every move in game. Resource document. The New York Times, April 21, 2015. <https://www.nytimes.com/2015/04/22/sports/baseball/statcast-offers-a-way-to-quantify-baseballs-every-move.html>. Accessed 01 Feb 2019.
- Sullivan, J. (2018). The greatest hope in Kansas City. Resource document. FanGraphs Baseball. <https://www.fan-graphs.com/blogs/the-greatest-hope-in-kansas-city/>. Accessed 01 Feb 2019.
- Tango, T., Lichtman, M., & Dolphin, A. (2007). *The book: Playing the percentages in baseball*. Sterling: Potomac Books Inc.
- TrackMan Overview. TrackMan. (2019). <https://trackman.zendesk.com/hc/en-us/sections/115000931467-TrackMan-Overview>. Accessed 01 Feb 2019.
- Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portraits. *Automation and remote control*, 24(6), 774–780.
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 234–244.
- Weekly Baseball. (2018). Baseball magazine SHA Co., Ltd, 73(38). (in Japanese)
- Whiteside, D., McGinnis, R. S., Deneweth, J. M., Zernicke, R. F., & Goulet, G. C. (2016). Ball flight kinematics, release variability and in-season performance in elite baseball pitching. *Scandinavian Journal of Medicine and Science in Sports*, 26(3), 256–265.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.