

Pólya urn model for analysis of football passes

Ken Yamamoto

Faculty of Science, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan

(Dated: November 7, 2025)

This study analyzes pass networks in football (soccer) using a stochastic model known as the Pólya urn. By focusing on preferential selection, it theoretically demonstrates that the time evolution of networks can be characterized by a single parameter. Building on this result, a data analysis method is proposed and applied to a large-scale public dataset of professional football matches. The statistical properties of the preferential-selection parameter are examined, demonstrating its correlation with pass accuracy and with mean pass difficulty. This method is applicable to various evolving networks.

Collective ball sports, such as football and basketball, can be regarded as multi-particle systems of tactically interacting players confined to the field. Recently, statistical, nonlinear, and mathematical physics have been extensively applied to sports analysis [1–3]. Examples include scoring events [4], player movements [5], ball possession time [6], and spatial flows of passes [7].

Complex networks have been utilized in various aspects of sports, including player formations [8], player matchups [9], and player transfers [10]. In particular, pass networks for ball sports have been intensively investigated [11–15]. The pass network of a team naturally comprises nodes and edges representing players and passes, respectively. Each edge is directed from the passer to the recipient, and its multiplicity represents the frequency of passes between them.

The time evolution of pass networks often reflects preferential selection, where players are more likely to attempt passes to teammates with whom they have previously completed multiple successful passes. These passes do not create new edges but rather increase the multiplicity of existing ones. A stochastic model describing the emergence of new passes based on a mathematically simplified form of preferential selection was proposed [16]. Here, N denotes the number of nodes, i.e., players in a team, and the number of possible directed edges is given by $M = N(N - 1)$. Initially, all M directed edges have a multiplicity of 0, i.e., N nodes are isolated, and M edges are assigned a uniform statistical weight of 1. At each edge selection step, one of the M directed edges is selected with a probability proportional to its statistical weight, and the statistical weight of the selected edge is increased by a constant α . Thus, edges selected frequently are likely to be selected when $\alpha > 0$, and α represents the strength of the preferential selection. In Ref. [16], the mean number of distinct edges after selecting edges τ times was derived as:

$$m(\tau) = M \left[1 - \frac{\Gamma(M/\alpha)}{\Gamma((M-1)/\alpha)} \frac{\Gamma((M-1)/\alpha + \tau)}{\Gamma(M/\alpha + \tau)} \right], \quad (1)$$

where Γ represents the gamma function. This formula has been shown to accurately describe the evolution of pass networks in football (soccer), rugby, and basketball [16]. In the following analysis of football matches,

the goalkeeper is excluded from the network due to their exceptional role; thus, $N = 10$ and $M = 90$.

This study investigates the preferential-selection model more deeply through its relationship to the Pólya urn model. Compared with a previous work [16], this approach provides a more comprehensive theoretical understanding and introduces an effective method for estimating the preferential-selection parameter α . Using this method, professional football matches are analyzed. The statistical properties of α are then examined, and the result shows that α correlates with pass accuracy and pass difficulty.

Here, the Pólya urn is introduced [17]. Suppose that an urn contains M balls of different colors, and a ball is drawn at random. The drawn ball is returned to the urn, and α balls of the same color as the drawn ball are added to the urn. This process is repeated. The probability that balls of a specific color are drawn k times within τ total draws is given by

$$P(k; \tau) = \binom{\tau}{k} \frac{B(1/\alpha + k, (M-1)/\alpha + \tau - k)}{B(1/\alpha, (M-1)/\alpha)}, \quad (2)$$

where B represents the beta function. This probability distribution is referred to as the beta-binomial distribution [18]. The Pólya urn and its variants have been employed to model a variety of phenomena [17, 19–21].

The pass network model described above relates to the Pólya urn as follows: pass types correspond to ball colors correspond, and both frameworks exhibit a preferential effect in which the statistical weight of the selected edge or ball color increases by α . Accordingly, Eq. (2) represents the probability that a particular pass appears k times out of a total of τ passes.

As the sum of multiplicities for all M directed edges equals the total number of passes τ , the mean edge multiplicity at τ is τ/M , regardless of α . The variance in edge multiplicities at τ is given by

$$\sigma^2(\tau) = \frac{(M-1)}{M^2(M+\alpha)} (\alpha\tau^2 + M\tau). \quad (3)$$

(See Feller [22].) For $\alpha = 0$, corresponding to uniform random selection, $\sigma^2(\tau) = (M-1)\tau/M^2$ scales linearly with τ , indicating that the standard deviation $\sigma(\tau)$ increases as $O(\tau^{1/2})$, in agreement with the central limit

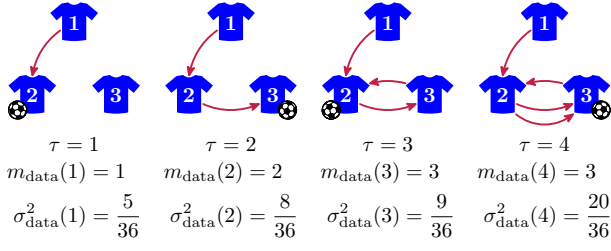


FIG. 1. Example of a pass sequence until $\tau = 4$ among $N = 3$ players and the computation of the number $m_{\text{data}}(\tau)$ of different passes and the variance $\sigma_{\text{data}}^2(\tau)$ of pass multiplicity.

theorem. In contrast, for $\alpha > 0$, $\sigma^2(\tau)$ grows faster than $O(\tau)$, reflecting the preferential effect that makes certain edges more likely to be selected.

The theoretical results presented above introduce two methods for estimating α based on a time series of passes from an actual match. The first estimate, α_m , is determined by counting the number of distinct passes within the initial τ passes, denoted by $m_{\text{data}}(\tau)$, and minimizing the sum of squared differences $\sum_{\tau=1}^T (m(\tau) - m_{\text{data}}(\tau))^2$, where T represents the total number of passes. Similarly, the second estimate, α_{var} , is derived by minimizing $\sum_{\tau=1}^T (\sigma^2(\tau) - \sigma_{\text{data}}^2(\tau))^2$, where $\sigma_{\text{data}}^2(\tau)$ indicates the variance in edge multiplicities calculated from the time series. An example illustrating both $m_{\text{data}}(\tau)$ and $\sigma_{\text{data}}^2(\tau)$ is shown in Fig. 1. At $\tau = 4$ in this figure, the pass $2 \rightarrow 3$ (from player 2 to 3) has multiplicity 2, $1 \rightarrow 2$ and $3 \rightarrow 2$ have multiplicity 1, and the other three empty passes have multiplicity 0, resulting in the variance $\sigma_{\text{data}}^2(4) = 20/36$.

Estimate α_{var} can be expressed in a closed form:

$$\alpha_{\text{var}} = -\frac{60M^2V - 5(M-1)(3T+2)}{60M^2V - (M-1)(12T^2 + 15T + 2)}M, \quad (4)$$

with

$$V = \frac{1}{T+1} \sum_{\tau=1}^T \frac{\tau(\tau-1)}{T(T-1)} \sigma_{\text{data}}^2(\tau) \quad (5)$$

which is calculated from the data (see Sec. S-I in Supplemental Material [23] for its derivation). However, a closed expression for α_m is not feasible owing to the complexity of the gamma functions in Eq. (1).

To evaluate the practical utility of α_m and α_{var} , a numerical experiment was conducted. Edge selection was simulated using the Pólya urn model to generate a sequence of T selected edges, and both α_m and α_{var} were calculated from the sequence. Figure 2 displays the results for $\alpha = 0.25, 0.5, 0.75$, and 1, with T ranging from 50 to 500 in increments of 50, and $M = 90$ (corresponding to football). The mean values of α_m and α_{var} over 10^4 independent trials are shown in Figs. 2(a) and (b), respectively. The means of both α_m and α_{var} converge to the true α as T increases. The mean of α_{var} remains close

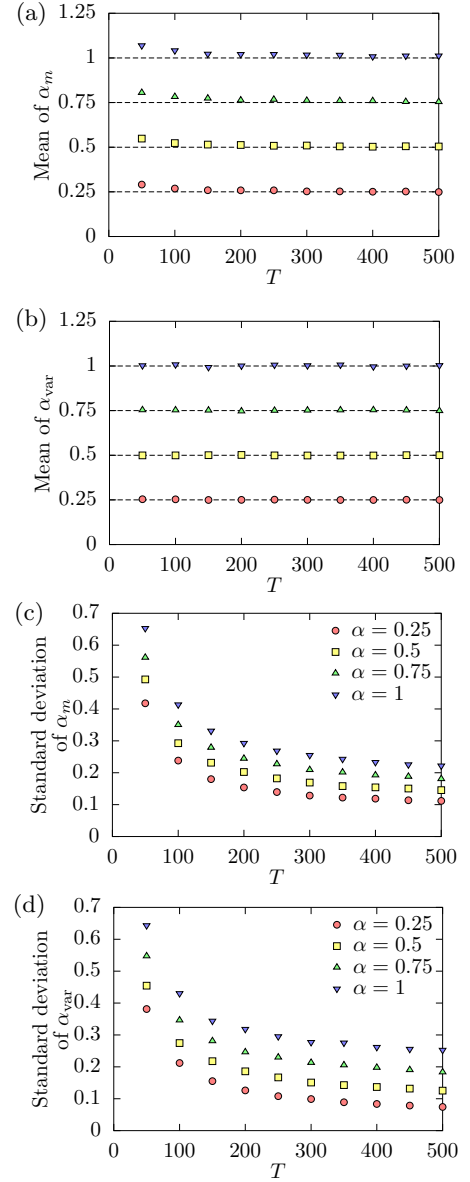


FIG. 2. Numerical result for estimating α_m and α_{var} with $T = 50, 100, \dots, 500$ and true values $\alpha = 0.25, 0.5, 0.75$, and 1. Mean values of α_m (a) and α_{var} (b). Standard deviations of α_m (c) and α_{var} (d).

to true α even at $T = 50$, whereas the mean of α_m deviates evidently for $T = 50$ and 100. Meanwhile, the standard deviations of α_m and α_{var} , displayed in Figs. 2(c) and (d), are comparable. Although these standard deviations decrease with increasing T , they remain between 0.1 and 0.3 even at $T = 500$. Therefore, α_{var} outperforms α_m owing to its lower estimation bias for small T and the availability of the calculation formula composed of Eqs. (4) and (5).

By definition, $m(\tau)$ represents the number of edges with nonzero multiplicity; thus, $m(\tau) = (1 - P(0; \tau))M$. Indeed, Eq. (1) is derived from Eq. (2) using this relation. Notably, $m(\tau)$ depends solely on $k = 0$ in $P(k; \tau)$, whereas $\sigma^2(\tau)$ incorporates all $k \geq 0$ in $P(k; \tau)$. The ad-

TABLE I. Six leagues analyzed in this study. ISL and WSL stand for Indian Super League and Women’s Super League, respectively.

Country	League	Season	Teams	Matches	Networks analyzed
Spain	La Liga	2015/16	20	380	678
England	Premier League	2015/16	20	380	672
Italy	Serie A	2015/16	20	380	657
Germany	Bundesliga	2015/16	18	306	559
India	ISL	2021/22	11	115	193
England	WSL	2020/21	12	132	244

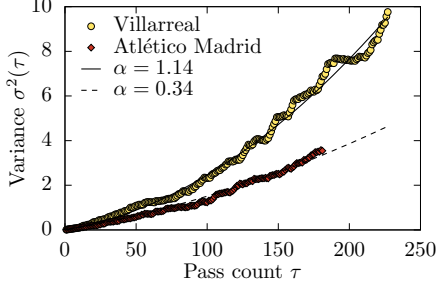


FIG. 3. Example of estimating α_{var} : Villarreal CF (home, circles) vs Atlético Madrid (away, diamonds) in La Liga played on September 26th, 2015. The solid and dashed curves represent Eq. (3) for $\alpha = 1.14$ and $\alpha = 0.34$, respectively.

vantage of α_{var} over α_m can presumably be attributed to this property.

Next, pass networks in actual football matches were analyzed. This study employed the StatsBomb Open Data [24], a publicly accessible and detailed dataset of football matches. As of April 2025, event data for 3433 matches had been uploaded. The dataset includes 21 leagues and competitions at the domestic, binational, confederation, and international levels. European domestic leagues account for the largest portion, comprising 80.5% of the dataset. The oldest matches in the dataset were played in 1958, and the most recent matches were played in 2024. The number of uploaded matches varies substantially by year or season, with the 2015/16 season accounting for 53.1%. Further details about the dataset are presented in Sec. S-II in Supplemental Material [23]. For the six leagues and seasons listed in Table I, all matches are included, whereas the other leagues and seasons are not fully covered. In addition to four European men’s leagues, the Indian Super League (ISL) and Women’s Super League (WSL) are analyzed to prevent geographical and gender biases.

The effect of player substitutions and halftime on the model remains uncertain. Therefore, only first-half passes were analyzed and teams that made substitutions during the first half were excluded. The number of remaining teams is shown in the rightmost column of Table I. Set pieces, such as kick-offs and throw-ins, were treated as passes. As noted earlier, passes made or re-

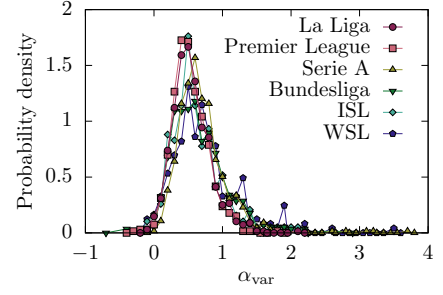


FIG. 4. Probability density of α_{var} for La Liga (circles), Serie A (squares), Premier League (triangles), Bundesliga (inverted triangles), ISL (diamonds), and WSL (pentagons).

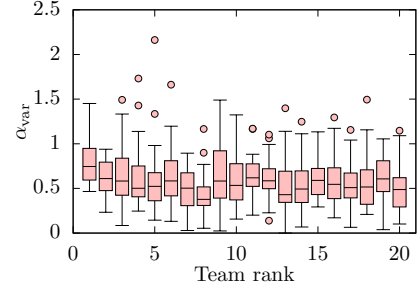


FIG. 5. Box plot of α_{var} for each team of La Liga. The teams are aligned according to their rank in the season.

ceived by the goalkeeper were excluded.

Figure 3 illustrates an example of $\sigma^2_{\text{data}}(\tau)$ derived from an actual match. The solid and dashed curves represent the theoretical $\sigma^2(\tau)$ curves with optimal α for the home and away teams, respectively.

Figure 4 shows the probability density function of α_{var} for the six selected leagues. Across these leagues, the distribution of α_{var} is consistently approximately unimodal, peaking near 0.7. The mean of α_{var} typically ranges between 0.6 and 0.8, regardless of country or player gender. Therefore, $\alpha \approx 0.7$ serves as an effective characterization of football passing behavior. Detailed league-specific statistics are provided in Table S-II in Supplemental Material [23]. In a previous study [16], two rugby matches were analyzed, and α was estimated for four teams, with the lowest value being $\alpha = 2.31$. This comparison suggests that typical α values vary across ball sports.

Only a few networks, such as 0.15% in La Liga, exhibited negative α_{var} . Negative α values may indicate anti-preferential selection, meaning that new or less successful passes are more likely to be chosen. However, negative α_{var} is more likely attributable to estimation errors. Indeed, as shown in Fig. 2(d), α_{var} includes estimation errors particularly for small T . Although the true α is positive, α_{var} may be estimated as negative due to estimation error when α is close to 0.

The distribution of α_{var} for each La Liga team is illustrated in Fig. 5 as a box plot. No significant differences among teams or notable correlations with team rank are

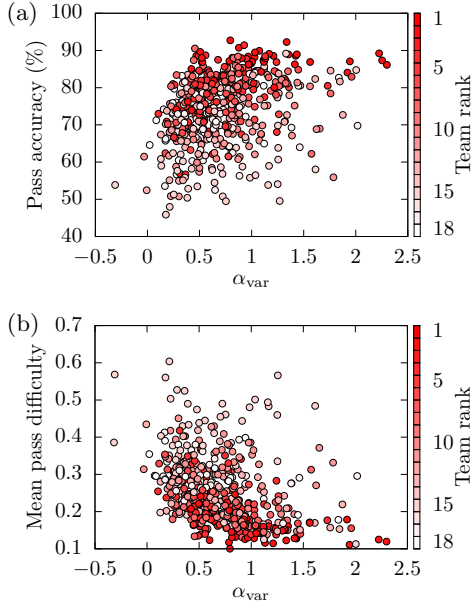


FIG. 6. Scatter plots of (a) pass accuracy versus α_{var} and (b) mean pass difficulty versus α_{var} for the Bundesliga, with Spearman's correlation coefficient ρ of 0.374 for (a) and -0.449 for (b). Color intensity indicates team rank.

found.

The parameter α is likely related to team passing indicators, as it can be estimated from the evolution of the pass network. One such indicator is pass accuracy, defined as the percentage of successful passes made by a team. To ensure consistency with the calculation of α_{var} , pass accuracy was computed using only first-half passes, excluding those involving the goalkeeper.

Figure 6(a) shows the scatter plot of pass accuracy versus α_{var} in the Bundesliga. A weak yet statistically significant correlation was observed, with Spearman's rank correlation coefficient [25] $\rho = 0.374$. The 95% confidence interval, estimated via bootstrap resampling [26], was $[0.293, 0.444]$. Spearman's ρ and corresponding confidence intervals for the other leagues are listed in Table II. Statistically significant positive correlations were observed in all leagues except for the Premier League whose confidence interval includes 0. The scatter plots for each league are displayed in Fig. S-2 in Supplemental Material [23].

To investigate why only the Premier League is exceptional with respect to the correlation between α_{var} and pass accuracy, a further analysis was performed. To visualize the trend in a scatter plot, Fig. 7 illustrates a moving average of pass accuracy along α_{var} ; for each α , the mean of pass accuracy over points whose α_{var} values lie within the window $[\alpha - \delta, \alpha + \delta]$, with $\delta = 0.2$ is plotted. Among the five leagues except for the Premier League (squares), some show upward trends and others rise initially and then plateau, both of which are consistent with positive correlation. In contrast, the graph of the Premier League displays a clear downward trend in $\alpha_{\text{var}} \gtrsim 1$. This “rise and decline” pattern yields virtually

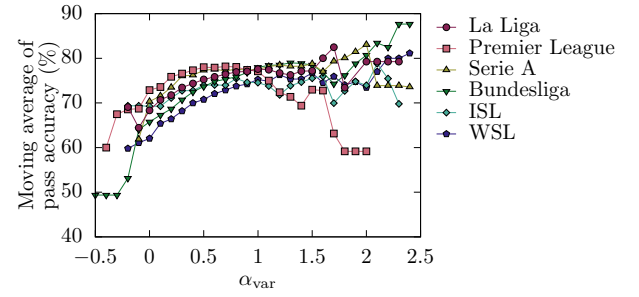


FIG. 7. Moving average of pass accuracy along α_{var} , within a window size of 0.4.

no correlation. A detailed inspection of teams, matches, and leagues will be required to understand the unique trend for the Premier League, which may be outside the scope of this study.

Figure 2(d) illustrates that the estimation of α is prone to error. In general, errors and noise reduce the strength of an observed correlation [27]. To address this issue, the simulation extrapolation (SIMEX) method [27], which introduces artificial noise with varying variance to estimate the error-free value, was employed. Further details on this method are provided in Sec. S-III in Supplemental Material [23]. Following the correction, the estimated correlation coefficient for the Bundesliga rose to $\rho = 0.457$. The corrected values and confidence intervals for the other leagues are listed in Table II. Even after the correction, the Premier League displayed virtually no correlation.

The positive correlation between α and pass accuracy can be explained as follows. A team with high α signifies that passes are concentrated among specific players. Consequently, such a team is more inclined to make safe passes and is less prone to attempt challenging, risky passes, resulting in high pass accuracy. However, real matches involve non-stationary irregular factors, including player spatial configurations, matchups with opposing teams, and weather conditions, which are not directly accounted for in the simple Pólya urn model. These factors likely reduce the correlation between α and pass accuracy.

To validate more directly the above hypothesis that teams with high α tend to make safe passes, the correlation between α_{var} and pass difficulty was investigated. However, the dataset lacked pass difficulty metrics. To address this issue, the success probability of each pass (termed expected pass or xPass [28]) was estimated using machine learning techniques [29] in Python, specifically employing the scikit-learn [30] and xgboost [31] libraries. In this study, pass difficulty was defined as 1 minus xPass, based on the premise that more difficult passes have lower success probabilities. The methodology for estimating xPass is detailed in Sec. S-IV in Supplemental Material [23]. Figure 6(b) shows the scatter plot of the mean difficulty of successful passes versus α_{var} in the Bundesliga, with $\rho = -0.449$. The negative correlation

TABLE II. Spearman's correlation coefficient ρ for α_{var} with pass accuracy and with mean pass difficulty for each league. The 95% confidence interval (CI) was estimated by the bootstrap method, and the corrected value was calculated using the SIMEX algorithm.

League	Correlation between α_{var} and pass accuracy				Correlation between α_{var} and mean pass difficulty			
	ρ	95% CI	SIMEX correction		ρ	95% CI	SIMEX correction	
			ρ	95% CI			ρ	95% CI
La Liga	0.269	[0.199, 0.339]	0.332	[0.220, 0.424]	-0.377	[-0.446, -0.317]	-0.483	[-0.566, -0.388]
Premier League	0.063	[-0.013, 0.142]	0.056	[-0.051, 0.159]	-0.166	[-0.235, -0.087]	-0.218	[-0.327, -0.107]
Serie A	0.164	[0.086, 0.239]	0.181	[0.078, 0.291]	-0.290	[-0.362, -0.214]	-0.351	[-0.457, -0.246]
Bundesliga	0.374	[0.293, 0.444]	0.457	[0.352, 0.555]	-0.449	[-0.522, -0.372]	-0.544	[-0.633, -0.434]
ISL	0.155	[0.010, 0.289]	0.189	[0.011, 0.402]	-0.364	[-0.486, -0.229]	-0.459	[-0.628, -0.259]
WSL	0.374	[0.249, 0.488]	0.426	[0.245, 0.555]	-0.406	[-0.514, -0.278]	-0.473	[-0.608, -0.287]

suggests that teams with higher α_{var} tend to choose safer passes. Spearman's ρ values for the other leagues are listed in Table II, exhibiting stronger correlations than those with pass accuracy. Although errors in mean pass difficulty are negligibly small compared with those in α_{var} (see Sec. S-V in Supplemental Material [23]), both errors were incorporated into the SIMEX correction.

The correlation between the preferential-selection parameter α and hands-on football characteristics such as pass accuracy and difficulty suggests the practical utility of α . Although pass difficulty is estimated from a large, extensive dataset using machine learning, α_{var} can be derived solely from sequential player-to-player pass data, offering a practical advantage.

This study investigated football passes using the Pólya urn model and discussed the statistical properties of pa-

rameter α . Other ball sports can also be analyzed using the Pólya urn, provided that pass data are available. This facilitates a comparative study of sports based on the preferential-selection characteristics. Additionally, the methodology developed in this study can be applied to the time evolution of systems represented as weighted networks [32, 33], such as email communication [34], information spread on online social media [35], and human mobility networks [36].

ACKNOWLEDGMENTS

This study was supported by a Grant-in-Aid for Scientific Research (C) 23K03264 from Japan Society for the Promotion of Science.

-
- [1] W. L. Winston, S. Nestler, and K. Pelechrinis, *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports* (Princeton University Press, Princeton, NJ, 2022).
 - [2] D. Sumpter, *Soccermatics: Mathematical Adventures in the Beautiful Game* (Bloomsbury, London, 2017).
 - [3] J. D. Barrow, *Mathletics: A Scientist Explains 100 Amazing Things about the World of Sports* (W. W. Norton & Company, New York, 2012).
 - [4] A. Clauset, M. Kogan, and S. Redner, Phys. Rev. E **91**, 062815 (2015).
 - [5] B. Kadoch, W. J. T. Bos, and K. Schneider, Phys. Rev. Fluids **2**, 064604 (2017).
 - [6] K. Yamamoto, S. Uezu, K. Kagawa, Y. Yamazaki, and T. Narizuka, Phys. Rev. E **109**, 014305 (2024).
 - [7] T. Morishita, Y. Aruga, M. Nakayama, A. Kijima, and H. Shima, Physica A **666**, 130507 (2025).
 - [8] T. Narizuka and Y. Yamazaki, Sci. Rep. **9**, 13172 (2019).
 - [9] F. Radicchi, PLoS One **6**, e17249 (2011).
 - [10] X. F. Liu, Y.-L. Liu, Q.-X. Wang, and T.-X. Wang, PLoS One **11**, e0156504 (2016).
 - [11] A. Chacoma, Phys. Rev. E **111**, 044313 (2025).
 - [12] K. Yamamoto and T. Narizuka, Phys. Rev. E **98**, 052314 (2018).
 - [13] J. M. Buldú, J. Busquets, J. H. Martinez, J. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque, Frontiers in Psychology **9**, 1900 (2018).
 - [14] F. M. Clemente, F. M. L. Martins, and R. S. Mendes, *Social Network Analysis Applied to Team Sports Analysis* (Springer, Cham, 2016).
 - [15] J. H. Fewell, D. Armbruster, J. Ingraham, A. Petersen, and J. S. Waters, PLoS One **7**, e47445 (2012).
 - [16] K. Yamamoto and T. Narizuka, Phys. Rev. E **103**, 032301 (2021).
 - [17] H. M. Mahmoud, *Pólya Urn Models* (CRC, Boca Raton, 2009).
 - [18] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions* (Wiley, New York, 2011).
 - [19] A. Bellina, G. De Marzo, and V. Loreto, Phys. Rev. Res. **7**, 023127 (2025).
 - [20] I. Iacopini, G. Di Bona, E. Ubaldi, V. Loreto, and V. Latora, Phys. Rev. Lett. **125**, 248301 (2020).
 - [21] E. Baur and J. Bertoin, Phys. Rev. E **94**, 052134 (2016).
 - [22] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., Vol. 1 (Wiley, New York, 1968).
 - [23] See Supplemental Material for details.
 - [24] StatsBomb, StatsBomb Open Data, <https://github.com/statsbomb/open-data> (2021), accessed: April 2025.
 - [25] S. Boslaugh, *Statistics in a Nutshell*, 2nd ed. (O'Reilly, Sebastopol, 2013).

- [26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).
- [27] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models*, 2nd ed. (Chapman & Hall/CRC, Boca Raton, 2006).
- [28] G. Anzer and P. Bauer, *Data Mining and Knowledge Discovery* **36**, 295 (2022).
- [29] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. (Packt, Birmingham, 2019).
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *J. Machine Learning Res.* **12**, 2825 (2011).
- [31] T. Chen and C. Guestrin, in *Proc. 22nd ACM SIGKDD International Conf. Knowledge Discovery and Data Mining* (ACM, 2016) pp. 785–794.
- [32] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, 2008).
- [33] F. Menczer, S. Fortunato, and C. A. Davis, *A First Course in Network Science* (Cambridge University Press, Cambridge, 2020).
- [34] M. E. J. Newman, S. Forrest, and J. Balthrop, *Phys. Rev. E* **66**, 035101(R) (2002).
- [35] Z. Ruan, G. Iniguez, M. Karsai, and J. Kertész, *Phys. Rev. Lett.* **115**, 218702 (2015).
- [36] H. Nilforoshan, W. Looi, E. Pierson, B. Villanueva, N. Fishman, Y. Chen, J. Sholar, B. Redbird, D. Grusky, and J. Leskovec, *Nature* **624**, 586 (2023).

Supplemental Material of “Pólya urn model for analysis of football passes”

Ken Yamamoto

Faculty of Science, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan

S-I. DERIVATION OF EQ. (4) FOR α_{var}

The optimal parameter α_{var} for given $\sigma_{\text{data}}^2(\tau)$ can be obtained by minimizing

$$S(\alpha) = \sum_{\tau=1}^T (\sigma^2(\tau) - \sigma_{\text{data}}^2(\tau))^2 = \sum_{\tau=1}^T \left(\frac{M-1}{M^2(M+\alpha)} (\alpha\tau^2 + M\tau) - \sigma_{\text{data}}^2(\tau) \right)^2.$$

A straightforward calculation yields

$$\left. \frac{dS}{d\alpha} \right|_{\alpha=\alpha_{\text{var}}} = \frac{2(M-1)}{M(M+\alpha_{\text{var}})^2} \left[\frac{M-1}{M^2(M+\alpha_{\text{var}})} \sum_{\tau=1}^T (\alpha_{\text{var}}\tau^2 + M\tau)(\tau^2 - \tau) - \sum_{\tau=1}^T (\tau^2 - \tau)\sigma_{\text{data}}^2(\tau) \right] = 0,$$

and the solution becomes

$$\alpha_{\text{var}} = - \frac{M^2 \sum_{\tau=1}^T (\tau^2 - \tau)\sigma_{\text{data}}^2(\tau) - (M-1) \sum_{\tau=1}^T (\tau^3 - \tau^2)}{M^2 \sum_{\tau=1}^T (\tau^2 - \tau)\sigma_{\text{data}}^2(\tau) - (M-1) \sum_{\tau=1}^T (\tau^4 - \tau^3)} M.$$

By using relations

$$\sum_{\tau=1}^T (\tau^3 - \tau^2) = \frac{1}{12} (T-1)T(T+1)(3T+2), \quad \sum_{\tau=1}^T (\tau^4 - \tau^3) = \frac{1}{60} (T-1)T(T+1)(12T^2 + 15T + 2),$$

the formula for α_{var} , Eq. (4) in the main text, is derived.

S-II. STATISTICAL RESULTS

A. Details of the dataset

We present detailed statistics for the dataset [S-1] as of April 2025. Table S-I lists leagues and competitions included in the dataset, exhibiting a geographical bias. The concentration on European matches is evident, while Asian matches are limited to the Indian Super League (ISL) and African matches are only from the Africa Cup. No matches from Oceania are available. In addition, the number of women’s matches is notably smaller than that of men’s matches. StatsBomb, the provider of the dataset, is a company based in England, and likely focuses commercially on European men’s leagues, which may explain these biases. In addition, the leagues and competitions included in this public dataset may have been selected for the purpose of promotions aimed at European users. Meanwhile, the dataset has been updated by adding new matches; as of April 2025, the most recent update was in 2024. Therefore, further enrichment of the dataset and the correction of geographical and gender biases are anticipated.

The years or seasons in which the matches in the dataset were played are also biased. Figure S-1(a) illustrates the number of matches in each year for the 3433 matches in the dataset. For ease of presentation, matches are counted by calendar year rather than by season; for example, the matches in the 2020/21 season are classified as either 2020 or 2021. The distribution is highly uneven, with 54.2% of the matches concentrated in 2015 or 2016. Figure S-1(b) presents a stacked bar chart showing the number of matches by year or season for each league and competition with more than 100 matches. The figure indicates that matches in the 2015/16 season for European domestic leagues have been extensively uploaded.

B. Statistics about α_{var}

Basic statistical results related to α_{var} for each league are detailed in Table S-II. As mentioned in the main text, the analysis focused solely on first-half passes by teams without substitutions during the first half and excluded passes

TABLE S-I. Details of leagues and competitions included in the dataset. NWSL stands for National Women’s Soccer League.

League or Competition	Matches	Level	Continent or Country	Gender	Available years or seasons
La Liga	868	Domestic	Spain	Men	1973/74, 2004/05–2020/21
Ligue 1	435	Domestic	France	Men	2015/16, 2021/22, 2022/23
Premier League	418	Domestic	England	Men	2003/04, 2015/2016
Serie A	381	Domestic	Italy	Men	1986/87, 2015/16
Bundesliga	340	Domestic	Germany	Men	2015/16, 2023/24
WSL	326	Domestic	England	Women	2018/19–2020/21
World Cup	147	International	–	Men	1958, 1962, 1970, 1974, 1986, 1990, 2018, 2022
Women’s World Cup	116	International	–	Women	2019, 2023
ISL	115	Domestic	India	Men	2021/22
Euro	102	Confederation	Europe	Men	2020, 2024
Africa Cup	52	Confederation	Africa	Men	2023
NWSL	36	Domestic	USA	Women	2018
Copa America	32	Confederation	South America	Men	2024
Women’s Euro	31	Confederation	Europe	Women	2022
Champions League	18	Confederation	Europe	Men	1970/71–1972/73, 1999/00, 2003/04–2004/05, 2006/07, 2008/09–2018/19
Major League Soccer	6	Binational	USA and Canada	Men	2023
UEFA Europa League	3	Confederation	Europe	Men	1988/89
Copa de Rey	3	Domestic	Spain	Men	1977/78, 1982/83, 1983/84
Liga Profesional	2	Domestic	Argentina	Men	1979, 1997/98
U20 World Cup	1	International	–	Men	1979
North American League	1	Binational	USA and Canada	Men	1977

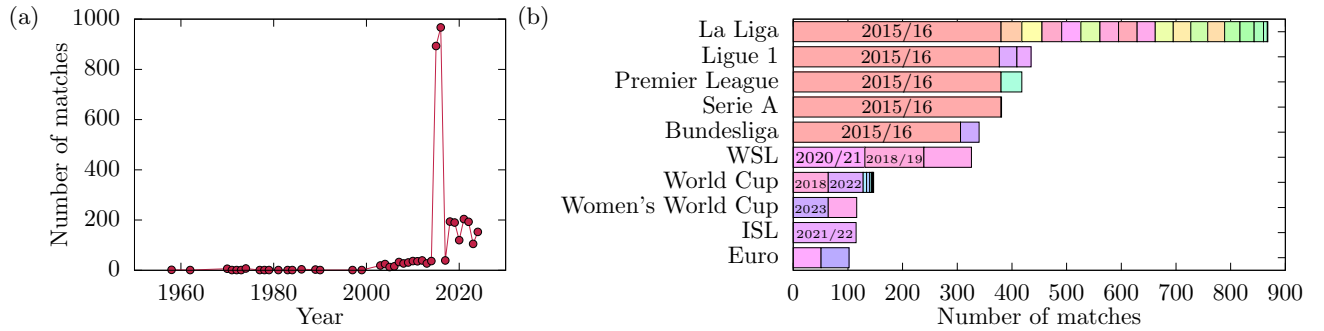


FIG. S-1. (a) The number of matches in the dataset in each year. (b) The number of matches by year or season for leagues and competitions with more than 100 matches. Year and season names are written in prominent bars.

TABLE S-II. Basic characteristics related to the number of passes and α_{var} . Some columns overlap with those in Table I in the main text.

League	Country	Gender	Season	Networks analyzed	Number of passes T (mean \pm sd)	α_{var} (mean \pm sd)
La Liga	Spain	Men	2015/16	678	169 \pm 60	0.60 \pm 0.28
Premier League	England	Men	2015/16	672	178 \pm 59	0.58 \pm 0.26
Serie A	Italy	Men	2015/16	657	180 \pm 65	0.75 \pm 0.38
Bundesliga	Germany	Men	2015/16	559	169 \pm 78	0.71 \pm 0.39
ISL	India	Men	2021/22	193	144 \pm 53	0.67 \pm 0.37
WSL	England	Women	2020/21	244	152 \pm 73	0.79 \pm 0.49

made or received by the goalkeeper. The number of passes T in Table S-II includes only first-half passes not involving the goalkeeper, and the value of α_{var} was calculated accordingly. T is smaller for the ISL and Women’s Super League (WSL) than for the other four leagues; however, α_{var} shows no significant differences across leagues.

C. Correlation of α_{var} with pass accuracy and mean pass difficulty

The scatter plot of pass accuracy versus α_{var} is shown in Fig. S-2. The six panels, corresponding to the leagues analyzed, are aligned in descending order of Spearman's rank correlation coefficient ρ . To assess the correlation strength, this study utilized Spearman's ρ instead of Pearson's correlation coefficient, as a linear relation between α_{var} and pass accuracy was not assumed. Indeed, the data points in each graph generally lie along a curve rather than a straight line.

Figure S-3 displays the scatter plot of mean pass difficulty against α_{var} , with the estimation method for pass difficulty detailed in the next section. The graphs for the six leagues are arranged in the same sequence as in Fig. S-2. For the Bundesliga, the results presented in Figs. S-2 and S-3 correspond to Figs. 6(a) and (b) in the main text,

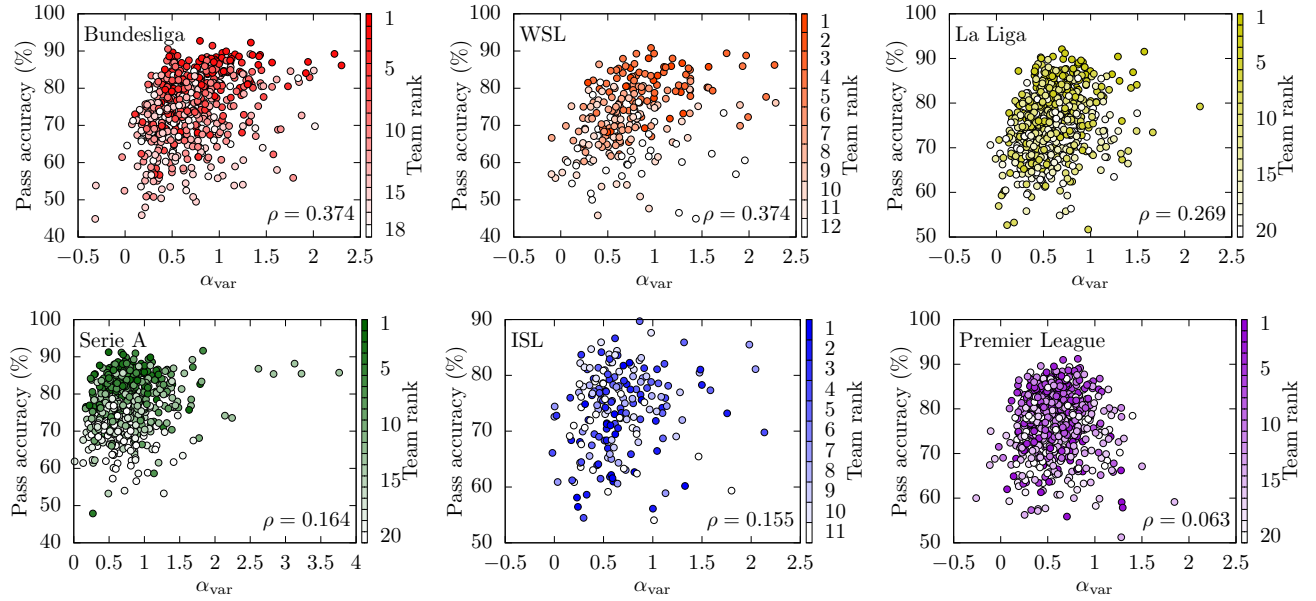


FIG. S-2. Scatter plot of pass accuracy against α_{var} . Color intensity indicates team rank. Panels corresponding to each league are aligned in descending order of Spearman's ρ , from left to right, top to bottom.

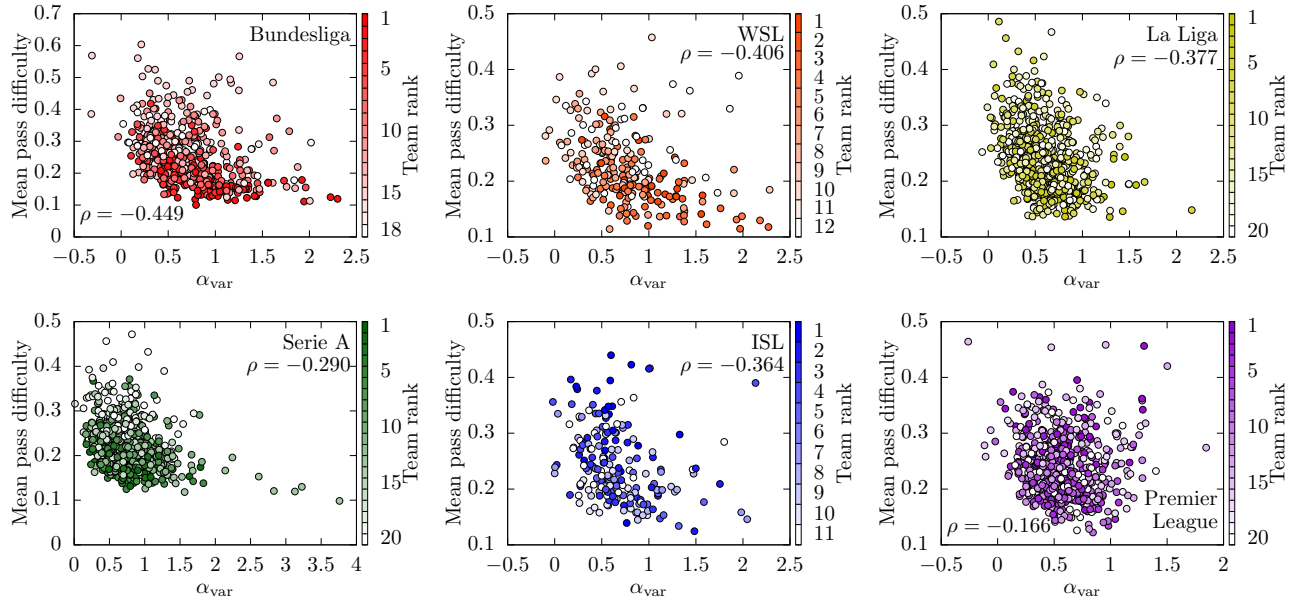


FIG. S-3. Scatter plot of mean pass difficulty against α_{var} . Color intensity indicates team rank. Six leagues are aligned in the same order as in Fig. S-2.

respectively.

S-III. SIMULATION EXTRAPOLATION METHOD

The simulation extrapolation (SIMEX) method [S-2] numerically corrects bias in statistical estimates caused by measurement errors. This corrected estimate is obtained by introducing artificial noise with varying variance into the data and extrapolating the resulting estimates to the error-free limit. This section describes the SIMEX procedure for Spearman's rank correlation coefficient ρ between α_{var} and pass accuracy, in which α_{var} contains estimation error, as illustrated in Fig. 2(d) of the main text. The same procedure is applicable to Spearman's ρ between α_{var} and mean pass difficulty.

The SIMEX mechanism is detailed as follows. The i th sample of $\alpha_{\text{var},i}$ can be expressed as $\alpha_{\text{var},i} = \alpha_i + \varepsilon_i$, where α_i denotes the true value and ε_i is a random variable representing the estimation error. Referring to Fig. 2(b) in the main text, the mean estimation error of α_{var} is assumed to be sufficiently close to 0. The exact value of ε_i cannot be determined from real data in principle, nor can its distribution be exactly obtained. Nevertheless, assume that the distribution is approximately estimable. For $\zeta \geq 0$, this study numerically generates $\alpha_{\text{var},i}(\zeta) = \alpha_{\text{var},i} + \sqrt{\zeta}\varepsilon'_i$, where ε'_i is a random number sampled from the estimated distribution of ε_i and is independent of ε_i . Then, the correlation coefficient $\rho(\zeta)$ between $\alpha_{\text{var}}(\zeta)$ and pass accuracy is calculated. By definition, $\alpha_{\text{var},i}(0) = \alpha_{\text{var},i}$ and $\rho(0)$ is identical to the original correlation coefficient between α_{var} and pass accuracy. By setting σ_i^2 for the variance of ε_i , the variance of $\sqrt{\zeta}\varepsilon'_i$ becomes $\zeta\sigma_i^2$ and that of $\alpha_{\text{var},i}(\zeta)$ becomes $(1 + \zeta)\sigma_i^2$. Naively, the estimation error can be eliminated by substituting $\zeta = -1$. In reality, however, negative variance $\zeta\sigma_i^2 = -\sigma_i^2$ corresponding to $\zeta = -1$ cannot be added. Alternatively, SIMEX estimates the corrected correlation coefficient $\rho(-1)$ by extrapolating from values of $\rho(\zeta)$ for $\zeta \geq 0$.

In the aforementioned method, the key step is to accurately estimate the distribution of ε'_i . However, the difficulty is that the distribution of estimation error for α_{var} varies with α and T [Fig. 2(d) in the main text], although the true value of α is unknown in the data analysis. Therefore, this study proposes the following method to approximately generate ε'_i . For each estimated $\alpha_{\text{var},i}$, the Pólya urn model is simulated with parameter $\alpha_{\text{var},i}$ to generate a sample of time series with length equal to the total number T of passes for the original data. The estimate $\alpha'_{\text{var},i}$ is obtained from this simulated time series using Eqs. (4) and (5) in the main text, and we let $\varepsilon'_i = \alpha'_{\text{var},i} - \alpha_{\text{var},i}$. In this estimation of ε'_i , the estimation error is approximated by substituting estimate $\alpha_{\text{var},i}$ for the unknown true α_i .

To verify whether the above estimation for ε'_i is reasonable, the numerical results are presented. Numerically, a time series with known α can be generated, which allows the estimation error $\varepsilon = \alpha_{\text{var}} - \alpha$ to be calculated exactly. Hence, the distributions of ε and ε' can be compared. Figure S-4(a) shows the probability densities of ε (histograms) and ε' (points) for $T = 50$ and 500 with true value $\alpha = 1$. Each graph was calculated using 10^5 samples. This figure shows that ε and ε' with the same T have probability densities close to each other. Instead of presenting the probability densities for additional α and T values, which would be repetitive, the numerical results of the Kullback-Leibler (KL) divergence are provided. The KL divergence

$$D_{\text{KL}}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

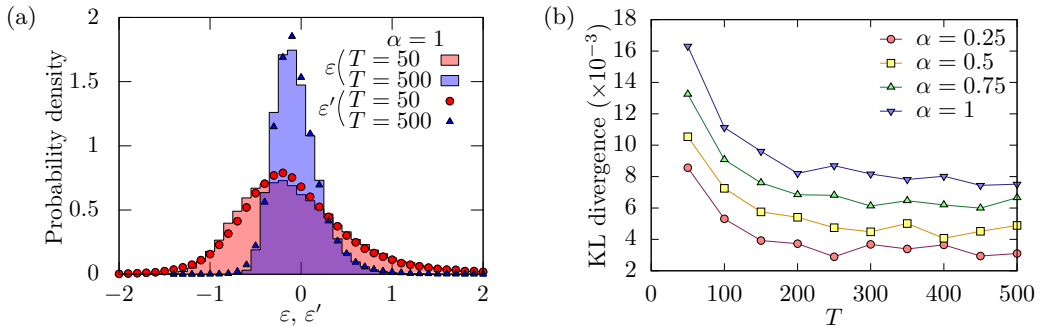


FIG. S-4. Numerical results for the distributions of ε and ε' . (a) Distributions for $\alpha = 1$ with $T = 50$ and 500. The distributions of ε are shown by histograms, and the distributions of ε' are shown by circles ($T = 50$) and squares ($T = 500$). (b) KL divergence between the distributions of ε and ε' for $T = 50, 100, \dots, 500$ and $\alpha = 0.25$ (circles), 0.5 (triangles), 0.75 (triangles), and 1 (inverted triangles).

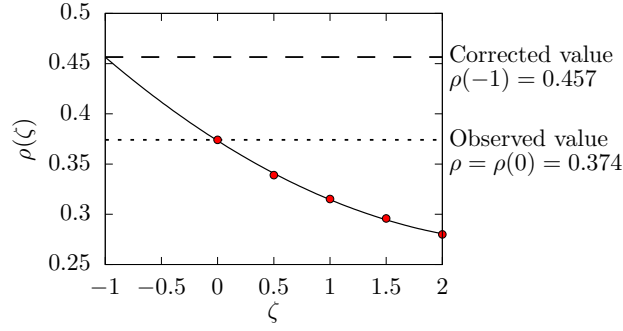


FIG. S-5. SIMEX method for the correlation coefficient ρ between α and pass accuracy in the Bundesliga. Circles represent the numerical result of $\rho(\zeta)$ averaged over 100 samples each. Solid curve represents the optimal quadratic function. Original value $\rho = \rho(0) = 0.374$ and corrected value $\rho(-1) = 0.457$ are shown by the horizontal dotted and dashed lines, respectively.

measures how much the distribution Q diverges from P , where $p(x)$ and $q(x)$ represent the probability densities of P and Q , respectively. Here, P and Q denote the distributions of ε and ε' , respectively. To numerically compute the KL divergence, continuous functions $p(x)$ and $q(x)$ were derived from discrete data samples using kernel density estimation, followed by numerical integration. Both kernel density estimation and numerical integration were executed with the SciPy library [S-3] in Python. Figure S-4(b) illustrates the KL divergence for $\alpha = 0.25, 0.5, 0.75$, and 1 and $T = 50, 100, 150, \dots, 500$. The KL divergence increases significantly with larger α and smaller T . However, Figure S-4(a) does not clearly show that the discrepancy between the distributions of ε (histogram) and ε' (circles) for $T = 50$ exceeds that for $T = 500$. Therefore, it is reasonable to infer that the distributions of ε and ε' are quite similar.

Figure S-5 presents the results of the SIMEX method for the correlation coefficient between α and pass accuracy in the Bundesliga. The circles represent simulated $\rho(\zeta)$ for $\zeta = 0, 0.5, 1, 1.5$, and 2 , each averaged over 100 independent samples. A quadratic function is conventionally used for the extrapolation function, and the solid curve shows the optimal function for $\rho(\zeta)$. The extrapolation to $\zeta = -1$ results in the corrected correlation coefficient $\rho(-1) = 0.457$, shown by the horizontal dashed line. This value is 22% higher than the original correlation coefficient $\rho = 0.374$, represented by the horizontal dotted line.

S-IV. ESTIMATION OF PASS DIFFICULTY

Machine learning techniques were employed to estimate the difficulty of each pass in the matches listed in Table I of the main text. General background of machine learning can be found in Raschka and Mirjalili [S-4]. Out of the 3433 matches in the dataset [S-1], those pertaining to the six leagues and seasons specified in Table I of the main text (also listed in Table S-II) were designated as the test set, while the remaining 1741 matches were used for training.

The input features (predictor variables) are listed in Table S-III. The features in the last four lines, separated by the horizontal line, are not provided directly but can easily be calculated. To prevent discontinuities caused by angle wrapping, pass direction was represented using its cosine and sine values instead of the raw angle. Features such as `end_location` and `duration` are determined simultaneously with the pass outcome (success or failure). Therefore, if the outcome of each pass is predicted based on the situation at the onset of the pass, these features cannot be utilized. However, since the aim is to assess the difficulty of each pass in previously played match, these features were retained. The outcome and difficulty of passes are strongly influenced by the configuration and movement of all players on the field; however, this information is not available in the dataset. A binary classifier was developed to predict pass outcomes, employing extreme gradient boosting (XGBoost). Computations were performed using Python along with the scikit-learn [S-5] and xgboost [S-6] libraries.

The confusion matrix for the first-half passes in the test set is presented in Table S-IV, which summarizes the prediction performance. “TP,” “FN,” “FP,” and “TN” denote “true positive,” “false negative,” “false positive,” and “true negative,” respectively. For instance, FN represents the number of passes predicted to fail but actually succeeded. The confusion matrix is presented as mean counts per team per match. From this matrix, $TP + FN = 170.3$ indicates the number of successful passes, comparable to the mean T shown in Table S-II, and $(TP + FN) / (FP + FN + TP + TN) = 0.774$ represents the overall pass accuracy of the analyzed teams. Furthermore, the following characteristics were

TABLE S-III. Features of passes used for predicting pass difficulty.

Feature name	Description	Data type	Notes
(Features directly provided in the data)			
location	x and y coordinates of the pass origin	[float, float]	$0 \text{ m} \leq x \leq 120 \text{ m}$, $0 \text{ m} \leq y \leq 80 \text{ m}$
length	Length of the pass	float	
duration	Time elapsed during the pass	float	
end_location	x and y coordinates of the pass destination	[float, float]	Same normalization as location
height	Height of the pass	integer	1: ground pass, 2: low pass, 3: high pass
under_pressure	Whether the pass was made under pressure	boolean	
position_id	Position ID of the passer	integer	from 1 (goalkeeper) to 25 (second striker)
bodypart_id	Body part used for the pass	integer	e.g., 38: left foot, 40: right foot
type_id	Type ID of the pass	integer	e.g., 65: kick off, 67: throw-in
(Features computed from the data)			
cos_angle, sin_angle	Cosine and sine of the angle of the pass	float	Calculated from angle data
dx, dy	x and y displacements of the pass	float	Difference of end_location and location
team	Team affiliation of the passer	binary	0: home and 1: away, obtained by the team name
gender	Players' gender	binary	0: men and 1: women

TABLE S-IV. Confusion matrix on the test set per match per team.

	Predicted positive	Predicted negative
Actual positive	TP = 141.6	FN = 28.7
Actual negative	FP = 6.5	TN = 43.1

calculated:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} = 0.839, \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} = 0.956, \\
\text{Recall} &= \frac{\text{TP}}{\text{FN} + \text{TP}} = 0.831, \\
\text{F1} &= 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.889.
\end{aligned}$$

S-V. ESTIMATION ERROR IN PASS DIFFICULTY

The estimation of α_{var} involves an error owing to one-sample fluctuation, as illustrated in Fig. 2(d) in the main text. Pass difficulty values, estimated by machine learning, also contain an error. In this section, we evaluate the estimation error in pass difficulty.

From the difficulty value of each pass, the mean pass difficulty of successful passes can be calculated for each team in each match. To estimate uncertainty in the mean pass difficulty, XGBoost was fit 50 independent times with different random seeds, with hyperparameters fixed at values obtained via grid search, yielding 50 difficulty estimates per pass. As a result, 50 samples for the mean difficulty were obtained for each team in each match. The variation of machine learning estimates for each team in each match can be measured using the sample standard deviation divided by the sample mean (i.e., the coefficient of variation or relative standard deviation). Figure S-6(a) shows the probability density of this coefficient of variation for each league. The distribution of the coefficient of variation peaks at approximately 6×10^{-3} .

The estimation error for α_{var} can be reasonably simulated as ε' described in the previous section, and the coefficient of variation for α_{var} can be estimated as the sample standard deviation of ε' divided by α_{var} . Figure S-6(b) shows the probability density of the coefficient of variation for α_{var} . The absolute value was taken to make the coefficient of variation positive when $\alpha_{\text{var}} < 0$. The distributions in Fig. S-6(b) have a peak at approximately 0.3 regardless of league. Thus, the typical magnitude 6×10^{-3} of the coefficient of variation for mean pass difficulty is 50 times smaller than that for α_{var} ; errors in mean pass difficulty are almost negligible relative to those in α_{var} .

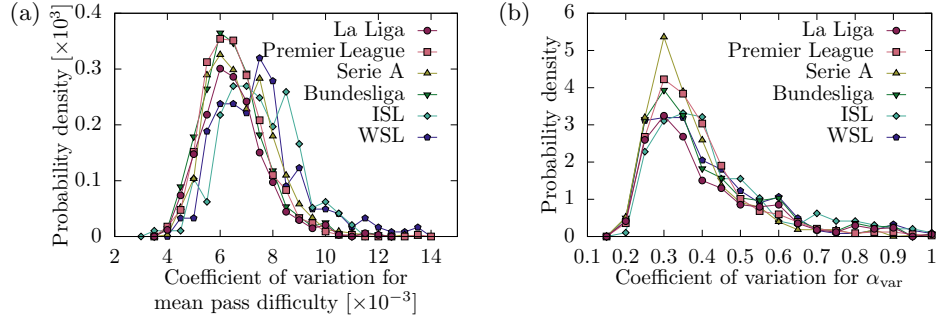


FIG. S-6. Probability density of the coefficient of variation for (a) mean pass difficulty and (b) α_{var} for each league.

-
- [S-1] StatsBomb, StatsBomb open data, <https://github.com/statsbomb/open-data> (2021).
- [S-2] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models*, 2nd ed. (Chapman & Hall/CRC, Boca Raton, 2006).
- [S-3] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, *Nature Methods* **17**, 261 (2020).
- [S-4] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. (Packt, Birmingham, 2019).
- [S-5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).
- [S-6] T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016) pp. 785–794.