VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

METODY REDUKCE DIMENZIONALITY STATISTICKÉHO SOUBORU
DIMENSIONALITY REDUCTION OF STATISTICAL DATASET

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Adam Sabo**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Pavel Hrabec, Ph.D.**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání bakalářské práce

Ústav:	Ústav matematiky
Student:	**Adam Sabo**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**Ing. Pavel Hrabec, Ph.D.**
Akademický rok:	2020/21

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Metody redukce dimenzionality statistického souboru

Stručná charakteristika problematiky úkolu:

Vysoká dimenzionalita statistického souboru je velmi často překážkou pro statistické modelování nebo testování statistických hypotéz. Naštěstí lze velmi často využít korelační (kovarianční) strukturu dat k redukci jejich dimenzionality při zachování maximálního množství informace (variability) původních dat (použitím metody hlavních komponent). Alternativou k metodě hlavních komponent je faktorová analýza, která sice nemaximalizuje zachovanou variabilitu, ale (pokud je to možné) zachovává jednoduchou interpretovatelnost výsledků.

Cíle bakalářské práce:

1) Seznámení se s aparátem matematické statistiky a matematiky nutným pro úspěšné použití metody hlavních komponent a faktorové analýzy.
2) Aplikace metod pro redukci dimenzionality na (generovanou nebo reálnou) sadu dat pomocí zvoleného softwaru.

Seznam doporučené literatury:

HEBÁK, Petr, Jiří HUSTOPECKÝ, Eva JAROŠOVÁ a Ivana MALÁ. Vícerozměrné statistické metody. Praha: Informatorium, 2005. ISBN 80-733-3025-3.

MELOUN, Milan a Jiří MILITKÝ. Statistická analýza experimentálních dat. Vyd. 2., upr. a rozš. Praha: Academia, 2004. ISBN 80-200-1254-0.

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2020/21

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt
Tato práce se zabývá představením metod pro redukci dimenzionality a následnou aplikací těchto metod na vybrané sportovní statistické soubory. První část práce pojednává o teoretickém aparátu matematické statistiky, a to konkrétně o metodě hlavních komponent a o její alternativě - faktorové analýze. Druhá část práce stručně vysvětluje pojmy týkající se souborů zvolených fotbalových statistik, na něž jsou metody aplikovány. Třetí část práce seznamuje s výsledky aplikací obou metod na statistické soubory. Data získaná výpočty v programovacím jazyku Python jsou vyjádřena a prezentována formou grafů a tabulkových výstupů.

Summary
This thesis introduces methods which are used to reduce dimensionality and their subsequent application to selected sets of sports statistical data. The first part of the thesis deals with the theoretical apparatus of mathematical statistics, in particular with the Principal Component Analysis and its alternative - the Factor Analysis. The second part provides a brief explanation of the terms related to the selected sets of football statistics where these methods are applied. The third part introduces the results of the application of both methods to statistical files. Data obtained through calculations performed in Python programming language are organized and interpreted by means of graphs and tables.

Klíčová slova
Analýza hlavních komponent, faktorová analýza, redukce dimenzionality, fotbalové statistiky.

Keywords
Principal component analysis, factor analysis, dimensionality reduction, football statistics.

SABO, A.*Metody redukce dimenzionality statistického souboru.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2021. 43 s. Vedoucí Ing. Pavel Hrabec, Ph.D.

Prohlašuji, že jsem bakalářskou práci *Metody redukce dimenzionality statistického souboru* vypracoval samostatně pod vedením Ing. Pavla Hrabce, Ph.D. a Ing. Martina Roseckého s použitím materiálů uvedených v seznamu literatury.

Adam Sabo

Rád bych zde poděkoval vedoucímu bakalářské práce Ing. Pavlu Hrabcovi, Ph.D. a Ing. Martinu Roseckému za odborné vedení mé práce a za cenné rady a čas, který mi při vytváření bakalářské práce věnovali.

Adam Sabo

Obsah

1. Úvod

V dnešním světě existují možnosti získávat velké množství dat a informací z různých oblastí. Už tato skutečnost v mnohém zlehčuje lidem život, nicméně teprve správné zpracování získaných dat umožní využít tuto možnost naplno. Přístup k velkému množství dat s sebou zároveň nese i určité potíže. Je třeba umět posoudit, které informace mohou být pro statistickou analýzu relevantní a které jsou naopak nevhodné.

Fenomén nadbytečně vysoké dimenzionality (označováno jako *„Curse of dimensionality“* [11], česky *„Prokletí dimenzionality“*) je v praxi velmi častým problémem, jenž se v souvislosti s využitím moderních nástrojů, získávajících velké množství dat, vyskytuje. S rostoucí dimenzionalitou roste obtížnost porozumění výsledkům dokonce exponencionální rychlostí, a proto je více než vhodné snažit se co největší měrou rozměr pozorovaného souboru redukovat (zatímco například pro úplné vyjádření šestidimenzionálního prostoru je potřeba 15 2D grafů, na čtyřdimenzionální prostor stačí takových grafů 6).

Práce je členěna do několika kapitol. V úvodní, teoretické kapitole je vysvětlen matematický aparát - metody, které problematiku vysoké dimenzionality řeší. V první části této kapitoly je představena metoda známá jako Analýza hlavních komponent. V druhé části je pak představena metoda, již je možné vnímat jako rozšíření metody hlavních komponent - Faktorová analýza.

Následuje kapitola zaměřující se na podrobné vysvětlení statistických souborů, na něž mají být metody aplikovány. Všechny používané soubory vychází ze světa fotbalu a popisují aspekty hry jednotlivých hráčů, brankářů nebo týmů. Kromě běžných statistik, jakými jsou například góly, přihrávky nebo střely, jsou v práci používány také některé moderní pokročilé statistiky. Těmto pokročilým statistikám je na závěr kapitoly věnována samostatná sekce.

Poslední kapitola se věnuje praktickému využití poznatků vycházejících z předchozích kapitol. Nejprve je představen kompletní postup aplikování metod redukce dimenze. Každý krok je zde patřičně okomentován a případné problémy jsou vysvětleny. Výsledky získané jednotlivými metodami jsou poté prezentovány formou grafů nebo tabulkových výstupů a vzájemně porovnány.

2. Aparát matematické statistiky

V této kapitole budou představeny dvě metody zabývající se problémem vysoké dimenzionality statistického souboru - *metoda hlavních komponent* a její alternativa, *faktorová analýza*.

Předpokladem pro použití těchto metod je podmínka, že původní proměnné jsou korelované. Lineární kombinací těchto proměnných pak vznikají jiné, nové proměnné, poskytující další úhel pohledu na zadaný problém. Cílem těchto metod je zachovat maximum původní informace při minimální dimenzi. *Faktorová analýza* má kromě toho také za cíl vytvořit nové proměnné tak, aby byla jejich interpretace co nejjednodušší.

Stručně vysvětlený rozdíl mezi oběma metodami lze nalézt například v článku [4].

2.1. Metoda hlavních komponent

Jak již bylo zmíněno, *metoda hlavních komponent* (*Principal Component Analysis*), dále zkráceně *PCA*, je metoda, která interpretuje zkoumanou úlohu novým, jednodušším způsobem. Cílem metody *PCA* je nalézt m latentních (skrytých) proměnných tak, aby vysvětlovaly co největší *variabilitu* původních proměnných. Hledané latentní proměnné bývají v terminologii *PCA* označovány jako *hlavní komponenty* nebo také *PC* (zkratka z angl. *principal components*).

Uvedeny budou převážně pojmy a vztahy potřebné pro pochopení souvislostí týkajících se našeho příkladu v kapitole Aplikace metod redukce dimenzionality. Podrobnější vysvětlení problematiky je k nalezení v knihách [5], [9], [6] a [8]. Tato kapitola - Aparát matematické statistiky vychází především z těchto publikací a je z většiny tvořena právě jejich citacemi a parafrázemi.

2.1.1. Model metody hlavních komponent

Základním prvkem, z kterého metoda *PCA* vychází, je kovarianční matice původních znaků. Jednotlivé *komponenty* jsou tvořeny charakteristickými (vlastními) vektory této matice. Bude ukázáno, jak tyto *komponenty* vznikají a proč je výhodné pozorované znaky normalizovat a získat tedy z kovarianční matice *matici korelační*.

Je zkoumán statistický soubor s veličinami X_1, X_2, ... , X_n, pro který jsou hledány *hlavní komponenty* Y_1, Y_2, ... , Y_m, kde $n \geq m$. Teoreticky lze najít tolik *komponent*, kolik je zkoumaných veličin, a v tom případě by byla vysvětlena veškerá celková *variabilita*. Nicméně cílem je kromě zachování co největší *variability* také popsat zkoumaný soubor co nejmenším počtem *hlavních komponent*. Tyto požadavky jsou však navzájem protichůdné. Aby hledaný kompromis vyhovoval oběma těmto požadavkům co nejlépe, je vhodné, aby celková *variabilita* nebyla mezi *komponentami* rozprostřena stejným dílem, ale naopak aby většinu celkové *variability* obsahovalo jen několik *komponent*.

Předpokladem pro úspěšné nalezení hlavních komponent je znalost vektoru středních hodnot $\boldsymbol{\mu}$ i kovarianční matice $\boldsymbol{\Sigma}$. Charakteristická čísla kovarianční matice se seřadí sestupně podle velikosti a označí $\lambda_1, \lambda_2, ..., \lambda_n$. Příslušné charakteristické vektory jsou pak $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_n$. Dále je zadán sloupcový vektor původních znaků (výběrových jednotek) $\boldsymbol{x}$. Komponenty jsou potom vytvářeny postupně. Začneme definicí první z nich:

$$Y_1 = \boldsymbol{\omega}_1^T(\boldsymbol{x} - \boldsymbol{\mu}),$$

kde veličinu Y_1 nazýváme *první hlavní komponenta*. Výraz ($\boldsymbol{x}$ - $\boldsymbol{\mu}$) vyjadřuje odchylky původních znaků od jejich středních hodnot (někdy označován jako $\boldsymbol{x}_c$, tzn. $\boldsymbol{x}_c = (x_1 - \mu_1,\ x_2 - \mu_2,\ ...,\ x_n - \mu_n)^T$). Vektor $\boldsymbol{\omega}_1$ je vektor koeficientů daný normalizační podmínkou

$$\boldsymbol{\omega}_1^T \boldsymbol{\omega}_1 = 1, \tag{2.1}$$

přičemž platí, že variabilita $D(Y_1)$, daná vztahem:

$$D(Y_1) = \boldsymbol{\omega}_1^T \boldsymbol{\Sigma}\, \boldsymbol{\omega}_1$$

je maximální. Lze dokázat, že maximální hodnota rozptylu Y_1 přes všechny vektory $\boldsymbol{\omega}_1$ je při splnění podmínky (2.1) největší charakteristické číslo λ_1 kovarianční matice $\boldsymbol{\Sigma}$, přičemž $\boldsymbol{\omega}_1$ je charakteristický vektor odpovídající λ_1 [10].

Stejným způsobem je definována i *druhá hlavní komponenta*:

$$Y_2 = \boldsymbol{\omega}_2^T(\boldsymbol{x} - \boldsymbol{\mu}),$$

navíc však musí být splněn požadavek nekorelovanosti (kolmosti) komponent. Tzn:

$$\boldsymbol{\omega}_1^T \boldsymbol{\omega}_2 = 0.$$

Druhá hlavní komponenta popisuje největší část doposud nevysvětleného rozptylu. Její rozptyl je pak roven:

$$D(Y_2) = \boldsymbol{\omega}_2^T \boldsymbol{\Sigma}\, \boldsymbol{\omega}_2 = \lambda_2.$$

Analogicky jsou vytvářeny i *další hlavní komponenty* $Y_3, Y_4, ..., Y_m$. Dříve zmíněný požadavek nekorelovanosti pak platí pro všechny dvojice hlavních komponent, tzn.:

$$\boldsymbol{\omega}_i^T \boldsymbol{\omega}_j = 0, \qquad \forall i, j \in \langle 1, 2, ..., m \rangle, i \neq j$$

Na závěr je vhodné zmínit, že maximální počet *komponent*, který lze nalézt, nezávisí pouze na počtu znaků statistického souboru n, ale také na počtu pozorování p. Platí, že maximální počet *komponent* odpovídá menší z těchto dvou hodnot. Případ, kdy je p menší než n, není sice tak obvyklý, ale může nastat. Pro úplnost je třeba dodat, že pokud je $p <= n$, je poslední komponenta triviální.

2.1.2. Interpretace vztahu mezi hlavními komponentami a původními proměnnými

Nyní, když už jsou hlavní komponenty vytvořeny, je třeba stanovit, kolik hlavních komponent bude pro popis zkoumaného statistického souboru vhodné použít. Už dříve bylo uvedeno, že pro uspokojivé výsledky je potřebné zachovat určitou (velmi významnou) část původní variability. Pro podrobnější pozorování této původní variability budou zavedeny další veličiny, a to *komponentní zátěž* a *komunalita*.

Nyní následuje shrnutí několika dříve uvedených skutečností. Z předchozí sekce je známo, že rozptyl i-té komponenty Y_i je roven charakteristickému číslu λ_i. Zároveň platí, že pokud jsou nalezeny všechny komponenty, tedy tolik komponent, kolik je proměnných

původního souboru, tzn. $m = n$, zůstane zachována veškerá celková variabilita. Pro vztah mezi kovarianční (korelační) maticí, rozptyly a charakteristickými čísly pak platí:

$$tr(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2 = D(Y_1) + D(Y_2) + ... + D(Y_m) = \lambda_1 + \lambda_2 + ... + \lambda_m.$$

Relativní variabilita i-té komponenty je pak rovna podílu λ_i / $tr(\mathbf{\Sigma})$, kde $tr(\mathbf{\Sigma})$ značí stopu matice $\mathbf{\Sigma}$.

Pro kovarianci mezi vektorem $\boldsymbol{x}_c$ a i-tou komponentou platí:

$$cov(\boldsymbol{x}_c, Y_i) = cov(\boldsymbol{x}_c, \boldsymbol{\omega}_i^T\ \boldsymbol{x}_c) = E(\boldsymbol{x}_c\ \boldsymbol{x}_c^T)\ \boldsymbol{\omega}_i = \mathbf{\Sigma}\ \boldsymbol{\omega}_i = \lambda_i\ \boldsymbol{\omega}_i,$$

a pro korelaci platí:

$$r(\boldsymbol{x}_c, Y_i) = \frac{\lambda_i\ \omega_{ji}}{\sigma_{x_j}\ \sqrt{\lambda_i}} = \frac{\sqrt{\lambda_i}\ \omega_{ji}}{\sigma_{x_j}}.$$

Pokud je však vektor $\boldsymbol{x}_c$ nejprve standardizován (normován), je vektor odchylek znaků od jejich středních hodnot ve tvaru

$$\boldsymbol{x}_N = \left(\frac{x_1 - \mu_1}{\sigma_{x_1}}, \frac{x_2 - \mu_2}{\sigma_{x_2}}, ..., \frac{x_n - \mu_n}{\sigma_{x_n}}\right).$$

Korelační matice je potom rovna kovarianční matici a pro korelační koeficient platí:

$$cov(\boldsymbol{x}_N, Y_i) = r(\boldsymbol{x}_N, Y_i) = \omega_{ji}^*\ \sqrt{\lambda_i^*},$$

kde ω_{ji}^* a λ_i^* odpovídají rozkladu korelační matice. Díky definici korelační matice jsou na diagonále této korelační matice pouze jedničky. Normalizace tedy zaručuje, že všechny původní znaky mají jednotkové rozptyly. Jinak řečeno, použití korelační matice namísto kovarianční matice odstraňuje závislosti původních znaků na velikosti jednotek, ve kterých jsou měřeny. Tato vlastnost bude požadována i v kapitole Aplikace metod redukce dimenzionality, proto budou i sledované znaky v této práci standardizovány.

Vektory $\omega_{ji}\sqrt{\lambda_i}$ označované jako $\boldsymbol{\gamma}_i$ se nazývají *vektory komponentních zátěží* (resp. $\omega_{ji}^*\ \sqrt{\lambda_i^*} = \boldsymbol{\gamma}_i^*$). Stejně jako charakteristické vektory, tak i *vektory komponentních zátěží* určují orientaci (směr) nového prostoru. Rozdíl je v tom, že zatímco velikosti charakteristických vektorů nic nevypovídají o významu odpovídajících komponent (z hlediska rozptylu), tak velikosti vektorů komponentních zátěží už ano. Charakteristická čísla totiž vysvětlují rozptyl sledovaných znaků podél směrů nového prostoru. Pokud jsou tedy prvky charakteristických vektorů vynásobeny odmocninami z odpovídajících charakteristických čísel (= definice *vektorů komponentních zátěží*), vzniká veličina, která dává do souvislosti rozptyly hlavních komponent a sledované znaky. Prvky *vektorů komponentních zátěží* bývají nazývány *komponentní zátěže* a platí pro ně, že jsou zároveň kovariancemi mezi původními znaky a nově vzniklými komponentami. Pokud je tedy vstupní maticí korelační matice, tak komponentní zátěže vyjadřují přímo korelaci mezi hlavními komponentami a původními znaky.

K pozorování celkové variability zachované hlavními komponentami slouží veličina nazývaná *komunalita*. *Komunalita* (původní) proměnné X_j se dá interpretovat jako podíl rozptylu vyjádřeného vybranými hlavními komponentami a rozptylu pozorovaného znaku $D(X_j) = \sigma_j^2$:

$$h_j^2 = \frac{\sum_{i=1}^{P} \gamma_{ji}^2}{\sigma_j^2}, \tag{2.2}$$

kde h_j^2 je komunalita j-té proměnné a γ_{ji} je faktorová zátěž j-té proměnné a i-té komponenty. P značí počet vybraných komponent.

Změna nastane, pokud budou znaky zkoumaného souboru standardizovány (výchozí matice bude korelační matice). Díky jednotkovým rozptylům jednotlivých znaků se vztah pro výpočet komunality (2.2) upraví na tvar

$$h_j^2 = \sum_{i=1}^{P} \gamma_{ji}^{*2}.$$

Pokud tedy úloha vychází z korelační matice, lze veličinu *komunalita* definovat jako sumu čtverců *faktorových zátěží* (díky vlastnostem rozkladu korelační matice zároveň jako sumu čtverců korelačních koeficientů). Množina hlavních komponent použitých pro výpočet *komunality* j-té proměnné je volena libovolně. Většinou je voleno několik prvních komponent a pozoruje se, jakou měrou se vysvětlená variabilita mění. Pokud by na výpočet *komunality* byla zvolena jen jedna komponenta, potom by *komunalita* byla rovna kvadrátu příslušné *faktorové zátěže*. Pokud by naopak byly zvoleny všechny komponenty, výsledná *komunalita* by byla rovna 1. Výpočet *komunalit* statistického souboru vycházejícího z korelační matice bude ukázán na příkladě v kapitole Aplikace metod redukce dimenzionality.

2.2. Faktorová analýza

Další metoda zabývající se problematikou vysoké dimenze statistického souboru je *faktorová analýza* (*Factor analysis*), dále zkráceně *FA*. Cílem této metody je stejně jako u *PCA* nalézt m latentních (skrytých) proměnných (v terminologii *FA* je nazýváme *faktory*). Rozdíl mezi *faktory* z *FA* a *komponentami* z *PCA* je v jejich pojetí. *Faktorová analýza* se na rozdíl od *metody hlavních komponent* snaží kromě vysvětlení celkové *variability* nabídnout i vhodnější interpretaci skrytých proměnných. Je vhodné dodat, že pro *faktorovou analýzu* existují na rozdíl od *metody PCA* také neortogonální modely. Tyto modely však v této práci použity nebudou. Definice a úvahy uvedené v kapitole věnující se *faktorové analýze* jsou určeny ortogonálním modelům. Pro kompletní vysvětlení neortogonálních metod by bylo nutné tuto kapitolu podstatně rozšířit.

Faktorová analýza byla původně využívána pouze v psychologii. Díky moderním technologiím však pronikla do řady dalších oborů.

Stejně jako u *metody PCA* i v této kapitole vychází velká část textu z literatury, která se zabývá definicemi následujících pojmů [5],[9], [7], [8]. Práce je doplněna některými komentáři k jednotlivým pojmům. Kromě toho je zde sjednoceno různé značení z různých literárních pramenů.

2.2.1. Model faktorové analýzy

Nechť je, stejně jako při užití metody hlavních komponent, zkoumán statistický soubor s veličinami $X_1, X_2, ..., X_n$. Vektor středních hodnot tohoto souboru pak je $\boldsymbol{\mu}$ a kovarianční matice $\boldsymbol{\Sigma}$. Model faktorové analýzy předpokládá, že existuje m v pozadí stojících latentních (skrytých) proměnných - faktorů $F_1, F_2, ..., F_m$, kde $n > m$. Vztah mezi veličinami X_j, kde $j = 1, ..., n$ a faktory F je potom vyjádřen jako:

$$X_j = \mu_j + \gamma_{j1}F_1 + \gamma_{j2}F_2 + ... + \gamma_{jm}F_m + \varepsilon_j.$$

Maticově lze zapsat také jako:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma f} + \boldsymbol{\varepsilon}. \tag{2.3}$$

$\boldsymbol{\Gamma}$ je matice *faktorových zátěží* tvořená faktorovými zátěžemi j-té veličiny a i-tého faktoru γ_{ji}, kde $j = 1, ..., n$ a $i = 1, ..., m$. Vektor f je vektor společných faktorů délky m. Prvky vektoru $\boldsymbol{\varepsilon}$ jsou nazývány *chybové složky*. Podle regresní terminologie lze *faktorové zátěže* γ_{ji} označit za regresní koeficienty n pozorovaných veličin na m nepozorovatelných faktorech (při splnění určitých podmínek to jsou zároveň kovariance mezi původními a novými proměnnými).

Způsob, kterým lze získat další interpretaci matice *faktorových zátěží*, je nahrazení vektoru $\boldsymbol{x}$ jeho standardizovanou formou. Nechť vektor $\boldsymbol{x}_c$ je vektor odchylek od jejich středních hodnot, tj. $\boldsymbol{x}_c = \boldsymbol{x} - \boldsymbol{\mu}$, potom lze rovnici (2.3) upravit do tvaru:

$$\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{x}_c = \boldsymbol{\Gamma f} + \boldsymbol{\varepsilon}.$$

Pokud je navíc vektor $\boldsymbol{x}_c$ standardizován, označujeme jej $\boldsymbol{z}$ a platí pro něj:

$$\boldsymbol{z} = \boldsymbol{\Gamma}^* \boldsymbol{f}^* + \boldsymbol{\varepsilon}^*. \tag{2.4}$$

Matice $\boldsymbol{\Gamma}^*$ pak má tu vlastnost, že její prvky (regresní koeficienty) γ_{ji}^* jsou zároveň korelační koeficienty mezi j-tou veličinou X_j a i-tým faktorem F_i.

2.2.2. Alternativní tvar rovnice faktorové analýzy

V této sekci budou využity předpoklady, které z doposud uvedeného textu nemusejí být zřejmé nebo nebyly uvedeny. Následuje tedy stručný výpis těchto předpokladů.

- Faktory F_i, kde $i = 1, ..., m$, jsou nezávislé náhodné veličiny s nulovými středními hodnotami a s jednotkovými rozptyly. V maticovém zápisu to znamená, že vektor středních hodnot $E(\boldsymbol{f}) = \boldsymbol{0}_m$ a kovarianční matice $Cov(\boldsymbol{f}) = \boldsymbol{I}_m$.

- Chybové složky ε_j, kde $j = 1, ..., n$, jsou nezávislé náhodné veličiny s nulovými středními hodnotami a s rozptyly $D(\varepsilon_j) = \Psi_j$. V maticovém zápisu je potom vektor středních hodnot $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}_j$ a kovarianční matice $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$ (tedy diagonální matice j-tého řádu s chybovými rozptyly Ψ znaků X_j na diagonále).

- Faktory F_i a chybové složky ε_j jsou nezávislé náhodné veličiny pro každou kombinaci $i = 1, ..., m$ a $j = 1, ..., n$. Pro kovarianční matici tedy platí $Cov(\boldsymbol{f}, \boldsymbol{\varepsilon}) = \boldsymbol{0}$ (nulová matice s m řádky a n sloupci).

Díky těmto předpokladům lze upravit původní faktorový model (2.3):

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{x}) = Cov(\boldsymbol{\mu} + \boldsymbol{\Gamma f} + \boldsymbol{\varepsilon}) = \boldsymbol{\Gamma} Cov(\boldsymbol{f}) \boldsymbol{\Gamma}^T + \boldsymbol{\Psi} = \boldsymbol{\Gamma} \mathbf{I}_m \boldsymbol{\Gamma}^T + \boldsymbol{\Psi} = \boldsymbol{\Gamma\Gamma}^T + \boldsymbol{\Psi}.$$

To znamená, že k nalezení matic $\boldsymbol{\Gamma}$ a $\boldsymbol{\Psi}$ je možné použít matici $\boldsymbol{\Sigma}$ a *rovnici faktorové analýzy* lze stručně zapsat jako:

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma\Gamma}^T + \boldsymbol{\Psi}, \tag{2.5}$$

resp.

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Gamma}^* (\boldsymbol{\Gamma}^*)^T + \boldsymbol{\Psi}^*, \tag{2.6}$$

pokud je výchozí matice $\boldsymbol{\Sigma}^*$ korelační maticí.

Pro rozptyl j-tého znaku $D(X_j) = \sigma_j^2$ pak z rovnice (2.5) vyplývá:

$$\sigma_j^2 = \sum_{i=1}^{m} \gamma_{ji}^2 + \psi_j^2.$$

Podíl tohoto rozptylu vysvětlený faktory je stejně jako u metody hlavních komponent (2.2) označován jako *komunalita*:

$$h_j^2 = \frac{\sum_{i=1}^{P} \gamma_{ji}^2}{\sigma_j^2}.$$

P značí počet vybraných faktorů.

Pokud bude jako výchozí použita korelační matice, lze díky jednotkovým rozptylům pozorovaných znaků *komunalitu* zapsat ve tvaru:

$$h_j^2 = \sum_{i=1}^{P} \gamma_{ji}^{*2},$$

(analogicky jako u metody PCA, podrobněji viz kapitola Metoda hlavních komponent).

Objektem pozorování bývá ve většině případů právě korelační matice. Jak už bylo naznačeno, interpretace problému při vycházení z korelační matice je jednodušší, pozorované znaky jsou ve stejném měřítku. Korelační matice bude mimo jiné použita i v praktické části této bakalářské práce Aplikace metod redukce a dimenzionality. Z těchto důvodů bude od příští sekce k objasňování některých aspektů řešení použita korelační matice $\mathbf{\Sigma}^*$.

2.2.3. Nejednoznačnost faktorové analýzy

Problémem faktorové analýzy je nejednoznačnost rovnice (2.5) , resp. její analogie pro korelační matici (2.6).

Pokud je pro analýzu použito $m = n$ faktorů, lze korelační matici $\mathbf{\Sigma}^*$ vyjádřit ve tvaru (2.6) tak, aby byla matice chybových rozptylů $\mathbf{\Psi}^*$ nulová. Pokud je však použit menší počet faktorů, což je cíl faktorové analýzy, nelze většinu korelačních matic do požadovaného tvaru rozložit. Nabízí se tedy otázka, jak nalézt matice $\mathbf{\Gamma}^*$ a $\mathbf{\Psi}^*$ tak, aby vyhovovaly modelu (2.4) a zároveň splňovaly předpoklady uvedené na začátku sekce Alternativní tvar rovnice faktorové analýzy.

Díky větě T. W. Andersona a J. Rubina lze vektor $\boldsymbol{z}$ vyjádřit v daném tvaru (2.4). Musí však být splněna nutná a postačující podmínka, a to existence diagonální matice $\mathbf{\Psi}^*$ s nezápornými prvky takovými, že rozdíl $\mathbf{\Sigma}^*$ - $\mathbf{\Psi}^*$ je pozitivně definitní matice s hodností m.

Nechť existuje jediná matice chybových rozptylů $\mathbf{\Psi}^*$. Pak se pro $m = 1$ matice $\mathbf{\Gamma}^*$ redukuje na sloupcový vektor, který je dříve uvedeným modelem a podmínkami jednoznačně určen. Problém nastává, pokud je $m > 1$. V takovém případě nelze faktory, resp. faktorové zátěže jednoznačně určit. Pokud by byl zaveden nový vektor společných faktorů a nová matice faktorů následujícím způsobem:

$$\boldsymbol{f}_1^* = \boldsymbol{T}\boldsymbol{f}^* \qquad a \qquad \mathbf{\Gamma}_1^* = \mathbf{\Gamma}^*\boldsymbol{T}^{-1},$$

kde $\boldsymbol{T}$ je ortogonální matice řádu m, po úpravách by vznikl model nerozlišitelný od modelu (2.4):

$$\boldsymbol{x} = \mathbf{\Gamma}_1^*\boldsymbol{f}_1^* + \boldsymbol{\varepsilon}^* = \mathbf{\Gamma}^*\boldsymbol{T}^{-1}\boldsymbol{T}\boldsymbol{f}^* + \boldsymbol{\varepsilon}^* = \mathbf{\Gamma}^*\boldsymbol{f}^* + \boldsymbol{\varepsilon}^*. \tag{2.7}$$

Oba tyto modely dávají stejnou reprezentaci původní matice $\mathbf{\Sigma}^*$. Ani komunality nejsou vlivem volby matice $\boldsymbol{T}$ nijak ovlivněny. Rozdíl mezi nově vzniklým modelem a původním modelem (2.4) je však v matici faktorových zátěží. Pro $m > 1$ tedy platí, že matici $\mathbf{\Sigma}^*$ lze vyjádřit nekonečně mnoha způsoby lišícími se maticí $\boldsymbol{T}$. Transformace pomocí matice $\boldsymbol{T}$ označujeme pojmem *rotace*.

2.2.4. Extrakce faktorů

Prvním krokem k úspěšnému použití faktorové analýzy je získání matice faktorových zátěží a chybových rozptylů z původního zkoumaného statistického souboru. Faktorová analýza se totiž na rozdíl od metody hlavních komponent nezabývá vznikem těchto matic, ale pouze jejich následnou rotací. A právě použití metody hlavních komponent může být zdrojem k získání požadovaných matic faktorových zátěží a chybových rozptylů. Další metody používající se k odhadu parametrů faktorového modelu neboli k *extrakci* faktorů jsou například metody *Maximum-Likelihood Method* (odhad parametrů, které s maximální věrohodností reprodukují pozorovanou matici) nebo *Least-Squares Method* (odhad založený na minimalizaci součtu čtvercových rozdílů mezi pozorovanou a reprodukovanou maticí). Podrobnější informace o různých metodách extrakce faktorů lze najít na webové stránce [12]. V této práci však bude k *extrakci* faktorů používána výhradně metoda hlavních komponent.

Prvním krokem při používání metody hlavních komponent jako metody k *extrakci* faktorů je stanovení faktorů v počtu m, které mají být nalezeny. Matice faktorových zátěží je poté rovna matici komponentních zátěží pro prvních m hlavních komponent. Pro vektor faktorových zátěží i-tého faktoru tedy platí:

$$\boldsymbol{\gamma}_i = \omega_{ji}\sqrt{\lambda_i},$$

kde λ_i jsou charakteristická čísla matice znaků a ω_{ji} jsou prvky příslušných charakteristických vektorů prvních m komponent.

Zbylé komponenty budou shrnuty do matice chybových rozptylů. Na diagonále budou prvky matice chybových rozptylů rovny:

$$\psi_j = 1 - \sum_{i=1}^{m} \gamma_{ji}^2. \tag{2.8}$$

Mimo diagonálu budou nuly. Výraz $\frac{\sum_{i=1}^{m}\gamma_{ji}^2}{\sigma_j^2}$ můžeme opět označit jako *komunalitu* j-tého znaku h_j^2.

Pozn.: 1 ve vzorci (2.8) značí rozptyl zkoumaného znaku. Díky standardizaci (použití korelační matice) je však tento rozptyl pro každý znak jednotkový.

2.2.5. Rotace faktorů

Nyní, když už jsou faktory nalezeny, lze konečně přistoupit k hlavnímu kroku faktorové analýzy, tedy k rotaci faktorů. Už dříve bylo uvedeno, že při transformaci rovnice (2.4) pomocí matice transformace $\boldsymbol{T}$ vznikne model ve tvaru (2.7). Vhodným zvolením matice $\boldsymbol{T}$ lze potom získat řešení nabízející strukturu, kterou je možné jednoduše popsat (v literatuře se uvádí jako „jednoduchá struktura“, resp. anglicky „simple structure“).

Cílem je najít takové řešení, které některé faktorové zátěže maximalizuje a jiné naopak minimalizuje.

Existují dva hlavní způsoby provedení rotace. Buď ortogonální rotace (pravoúhlá), nebo neortogonální rotace (šikmá). Každý z těchto způsobů lze dále dělit [13]. V této práci budou předvedeny pouze ortogonální rotace, a to konkrétně metody *Varimax* a *Quartimax*. Mezi neortogonální metody patří například metoda *Promax*. Rozdíl mezi ortogonálními a neortogonálními rotacemi je takový, že faktory získané rotací některé ortogonální metody na sebe zůstávají kolmé i po provedení transformace. Faktory získané metodou neortogonální rotace nikoliv.

První ze zmíněných metod, tedy metoda *Varimax*, volí matici rotace $\boldsymbol{T}$ tak, aby byla maximalizována funkce:

$$V = \frac{1}{n}\sum_{i=1}^{m}\left[\sum_{j=1}^{n}\left(\frac{\gamma_{ji}^{*}}{h_j^2}\right)^4 - \frac{1}{n}\left(\sum_{j=1}^{n}\left(\frac{\gamma_{ji}^{*}}{h_j^2}\right)^2\right)^2\right]. \tag{2.9}$$

Jinak řečeno, metoda *Varimax* provádí transformaci tak, aby součet rozptylů druhých mocnin faktorových zátěží byl pro jednotlivé faktory co největší.

Vyjádření výše uvedené funkce je používáno pro rotaci korelační matice. Je to patrné i díky členu γ_{ji}^{*}, který značí normalizované faktorové zátěže. Kromě vyjádření pro korelační matici existuje také obecnější vyjádření. V praxi však bývá ve většině případů (a to i v pozdější kapitole této práce) použito právě vyjádření 2.9.

Metoda *Varimax* je nejčastěji používanou metodou rotace faktorů. Výsledky získané použitím této metody splňují požadavky, které od faktorové analýzy očekáváme (nově vzniklé faktory korelují silně s několika málo faktory, zatímco s ostatními korelují slabě). Bývá také používána jako mezikrok při použití šikmých (neortogonálních) metod rotace. Je důležité zdůraznit, že použití odlišných metod extrakce ovlivňuje také následnou rotaci faktorů a získané výsledky se pak také odlišují.

Jako druhá metoda k rotaci faktorů (také ortogonální) bude použita metoda *Quartimax*. Výsledek je opět získán maximalizací funkce, tentokrát:

$$V = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(\gamma_{ji}^{2})^2}{nm} - \frac{\left[\sum_{i=1}^{m}\sum_{j=1}^{n}\gamma_{ji}^{2}\right]^2}{n^2m^2}.$$

I v tomto případě je cílem maximalizovat rozptyl druhých mocnin faktorových zátěží. Na rozdíl od metody *Varimax* se zde počítá rozptyl přes celou matici, nikoliv postupně pro všechny faktory. Metoda *Quartimax* se více soustředí na minimalizaci počtu faktorů potřebných k vysvětlení variability původních znaků. Použitím této metody vzniká model, ve kterém se do prvního faktoru dostává i relativně velká část variability prvků, které by jinak s tímto faktorem téměř nekorelovaly. To může být pro analýzu řešení nevhodné. Tato metoda nebývá používána tak často jako metoda *Varimax*.

3. Hráčské a týmové statistiky

V dalších kapitolách bude předmětem zkoumání větší množství různorodých hráčských i týmových statistik z oblasti fotbalu. A protože je toto téma poměrně specifické, bylo by vhodné některé ze statistik podrobněji vysvětlit.

V první sekci budou představeny pojmy, jejichž význam není všem zřejmý nebo které by si mohl laik špatně vyložit. V druhé sekci budou definovány pokročilé statistiky známé jako „očekávané góly“ (angl. expected goals) a „očekávané asistence“ (angl. expected assists).

3.1. Pojmy

Znaky zkoumaného statistického souboru jsou zadávány zkratkami. Z tohoto důvodu zde bude vložena sekce věnující se objasnění těchto zkratek a vysvětlení pojmů s nimi souvisejících. Zkratky budou seřazeny postupně podle pořadí, ve kterém se objevují v kapitole Aplikace metod redukce dimenzionality. Nejprve tedy budou objasněny znaky ze souboru brankářských statistik a později také z hráčských a týmových statistik. Kompletní seznam používaných statistik je k dispozici také v sekci Seznam použitých zkratek a symbolů.

Brankářské statistiky používané v této práci jsou následující:

- **GAA**: Průměrný počet obdržených branek za zápas.
- **Save%**: Úspěšnost zákroků proti střelám mířícím na bránu. Zblokované střely nebo střely mimo bránu se nepočítají.
- **W%**: Procento zápasů, ve kterých brankářův tým zvítězil.
- **CS%** : Procento zápasů, ve kterých brankář neobdržel žádný gól.
- **Psv%**: Úspěšnost při pokutovým kopech.
- **LCmp%**: Přesnost přihrávek delších než 40 yd.
- **Stp%**: Procento křižných přihrávek soupeře do pokutového území, které brankář úspěšně překazil.
- **OPA/90**: Průměrný počet defenzivních zákroků vně pokutového území během zápasu.
- **AvgDist**: Průměrná vzdálenost od brány při defenzivních zákrocích.

Použité hráčské statistiky:

- **Goals**: Počet vstřelených gólů.
- **Assists**: Počet gólových asistencí.
- **xGoals**: viz sekce Expected goals.
- **xAssists**: viz sekce Expected assists.

- **Shoots**: Počet střel na bránu.
- **SoT%**: Procento střel, které míří na bránu.
- **Pass%**: Úspěšnost přihrávek.
- **KP**: Počet přihrávek vedoucích ke střele.
- **Tackles**: Počet odebrání míče soupeři.
- **Press**: Počet napadání soupeře během rozehrávky nebo kontroly míče.
- **Blocks**: Počet zablokování střely nebo přihrávky soupeře vlastním tělem.
- **Touches**: Počet doteků s míčem.
- **Dribbles**: Počet vyhýbání se soupeři snažícímu se zachytit míč.

Použité týmové statistiky:

- **Goals**: Počet vstřelených gólů.
- **Assists**: Počet gólových asistencí.
- **Pkmade**: Počet vstřelených gólů z pokutových kopů.
- **Pkatt**: Počet pokutových kopů.
- **CrdY**: Počet žlutých karet.
- **CrdR**: Počet červených karet.
- **xGoals**: viz sekce Expected goals.
- **xAssists**: viz sekce Expected assists.
- **GoalsAgainst**: Počet obdržených branek.
- **CleanSheet**: Počet zápasů bez obrdženého gólu.
- **Sh**: Počet střel.
- **SoT**: Počet střel na bránu.
- **G/Sh**: Úspěšnost střel.
- **G/SoT**: Úspěšnost střel na bránu.
- **FreeKicks**: Počet střel z přímých kopů.
- **Cmp**: Počet přihrávek.
- **Att**: Počet pokusů o přihrávku.
- **Cmp%**: Úspěšnost přihrávek.

- **KP**: Počet klíčových přihrávek.
- **Tackles**: Počet pokusů o odebrání míče soupeři.
- **TacklesW**: Počet úspěšných odebrání míče soupeři.
- **Press**: Napadání soupeře během rozehrávky nebo kontroly míče.
- **PressW**: Úspěšné napadání soupeře během rozehrávky nebo kontroly míče.
- **Blocks**: Počet zablokovaných střel a přihrávek.
- **Touches**: Počet doteků s míčem.
- **DribblesW**: Úspěšné vyhýbání se soupeři snažícímu se zachytit míč.
- **Dribbles**: Vyhýbání se soupeři snažícímu se zachytit míč.

3.2. Expected stats

Expected stats jsou moderní pokročilé statistiky, které komplexněji popisují herní situace. Později budou v této práci zmíněny statistiky expected goals a expected assists. Proto je zde zařazena krátká sekce, která se bude věnovat jejich vysvětlení.

Existuje několik modelů, podle nichž se expected stats počítají. Databáze [14], z níž jsou čerpány statistiky pro tuto práci, obsahuje data vypočítaná modelem společnosti StatsBomb [23]. Níže popsané statistiky vycházejí z tohoto modelu.

Stručně řečeno, Expected stats je systém, který hodnotí kromě kvantity jednotlivých herních aspektů také jejich kvalitu.

3.2.1. Expected goals

Statistika Expected goal (dále xG), v češtině někdy označováno jako „očekávaný gól“, je pravděpodobnost, že výsledkem střeleckého pokusu bude vstřelený gól. Každá střela je charakterizována proměnnými, z nichž se ona pravděpodobnost vypočítá. Mezi tyto proměnné patří:

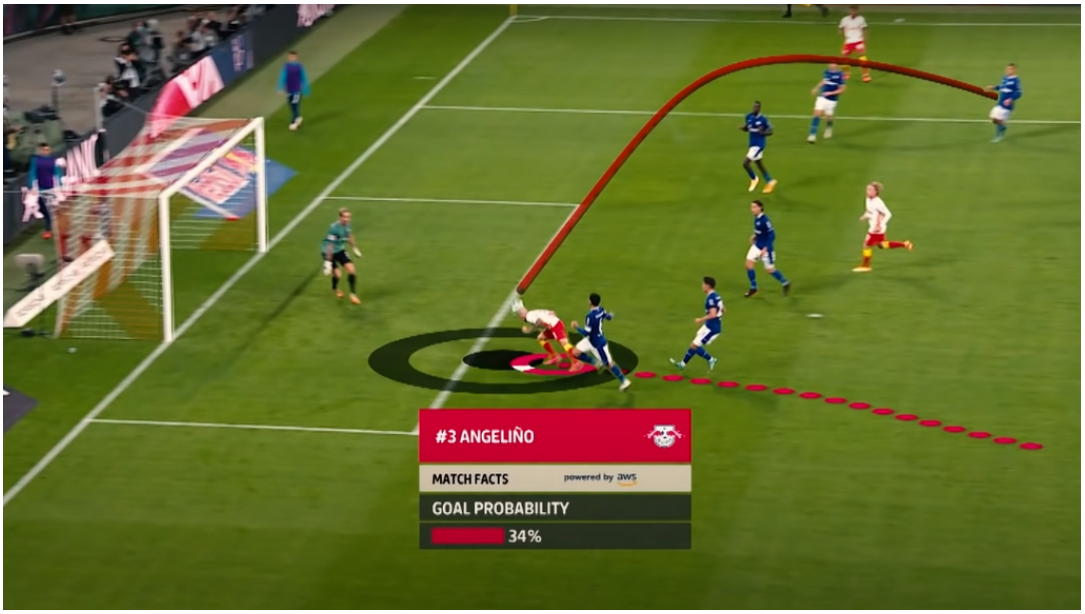
- **Pozice střelce**: Z jaké vzdálenosti a pod jakým úhlem hráč střílel?
- **Část těla**: Jakou částí těla střelec hrál?
- **Typ přihrávky**: Dostal střílející hráč dobrou přihrávku, nebo si musel míč ke střele připravovat?
- **Typ střely**: Byl střelec v dobré pozici, nebo byla obrana připravena zablokovat střelu?
- **Pozice hráčů bránícího mužstva**: Kolik hráčů stojí mezi míčem a bránou? Kolik hráčů je v okolí střelce? Je brankář v bráně?

Každá střela je následně porovnávána s mnoha modelovými situacemi a vyhodnotí se celková pravděpodobnost vstřelení gólu.

Následuje vysvětlení modelu xG, a to na základě obrázků z videa [15]. Kromě čtyř akcí, které jsou zde uvedeny, lze ve videu najít i další situace vysvětlující používaný model.

Obrázek 3.1: xG - Pokutový kop

Zvláštním případem střely je pokutový kop (viz obrázek 3.1). Vzdálenost od brány je při každém pokutovém kopu vždy stejná (11 metrů). Také střelecký úhel je konstantní (37°). Další proměnné, které se pro výpočet xG používají, jsou také vždy stejné (míč je položen na hřišti, obránci nijak střelci nebrání, brankář stojí připraven na brankové čáře). Každý pokutový kop má tedy stejnou hodnotu xG a tou je hodnota $xG = 0{,}77$.



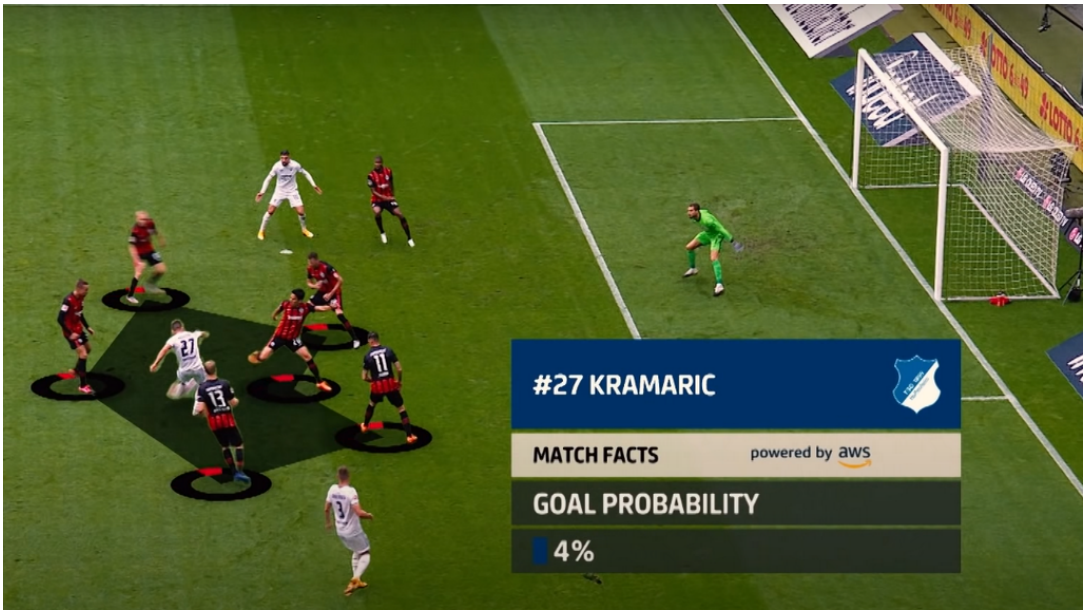
Obrázek 3.2: xG - Zakončení hlavou

Při akci na obrázku 3.2 měl střelec šanci na vstřelení gólu 34%. Mohlo by se zdát, že pozice, ve které se střelec nacházel, je ideální. Míč byl zhruba 6 metrů od brány, brankář

zdaleka nevykrýval většinu brány a v cestě nestál žádný obránce. Jedním z důvodů, proč tomu tak není, je však například to, že hráč musel zakončovat akci hlavou, což je statisticky náročnější. Dalším z důvodů byla nutnost zkoordinovat pohyb s přihrávkou, která sice byla přesná, ale ne jednoduchá na zakončení.

Obrázek 3.3: xG - Blokovaná střela

Na obrázku 3.3 je vidět šance, při které je pravděpodobnost vstřelení gólu 10 %. Pozice střelce je opět poměrně dobrá. Stojí zhruba 9 metrů od brány a střelecký úhel je 37°. Problém je v tom, že většinu tohoto střeleckého úhlu vykrývá svým tělem obránce a střelec tedy musí zamířit velmi přesně.



Obrázek 3.4: xG - Útočná akce proti koncentrované obraně

Na posledním obrázku 3.4 je střelec v pozici, ve které má střelecký úhel pouze 21°. Zároveň je obklíčen 6 bránícími hráči, kteří na něj vyvíjejí tlak a současně blokují případ-

nou střelu. Dalším faktorem snižujícím hodnotu xG je to, že brankář stojí v ideální pozici a je připraven zasáhnout. Šance na vstřelení gólu je tedy pouze 4%.

Další informace o modelu xG používáném v této práci lze najít na webové stránce [24].

3.2.2. Expected assists

Statistika Expected assist (dále xA) vyjadřuje pravděpodobnost, že výsledkem přihrávky bude gólová asistence. Stejně jako je každé střele přiřazena hodnota xG, tak i každé přihrávce je přiřazena hodnota xA. Proměnné, jež charakterizují jednotlivé přihrávky a z nichž se celková pravděpodobnost počítá, jsou například:

- **Pozice přihrávajících si hráčů**: Jak vzdálení od sebe hráči jsou? Brání jim někdo ve hře? V jaké části hřiště se tito hráči nacházejí?
- **Rychlost hry**: Jednalo se o běžnou přihrávku, nebo o přihrávku do protiútoku?
- **Typ přihrávky**: Je přihrávka přesná, nebo si ji bude muset hráč obtížně zpracovávat?

Může se zdát nejasné, proč je hodnota xA přiřazena každé přihrávce, a ne jen přihrávkám, po kterých následuje střela. Důvod je však jednoduchý. V některých situacích může po přihrávce potencionální střelec vlastní vinou namísto střely míč ztratit. Přihrávka by potom měla hodnotu $xA = 0$, i když ve skutečnosti mohla být kvalitní. Přiřazení hodnoty xA všem přihrávkám však tento problém řeší. Statistika xA tedy umožňuje identifikovat hráče, kteří jsou schopní nahrávači, nicméně kvůli méně kvalitnímu týmu nemají tolik asistencí jako hráči z kvalitnějších týmů. Pokud by tedy tento schopný nahravač přestoupil do kvalitního týmu a zachoval si stejnou výkonnost, měl by se počet jeho asistencí výrazně zvýšit.

4. Aplikace metod redukce dimenzionality

Další kapitola této práce se věnuje praktickému využití metod PCA a FA. Vstupním statistickým souborem, který bude objektem sledování, je databáze [14], která obsahuje fotbalové statistiky. Konkrétně budou analyzovány hráčské a brankářské statistiky a statistiky týmů z německé Bundesligy v ročníku 2018/2019. Nejprve bude podrobně předveden celý postup na brankářských statistikách. V další části budou okomentovány i výsledky hráčských statistik a statistik týmů.

Je třeba zdůraznit, že hráči, jejichž statistiky byly použity a kteří byli do následujících výpočtů zahrnuti, odehráli v průběhu sezóny nejméně 900 minut [2]. Důvod použití tohoto omezení je takový, že pro lepší statistické výsledky je vhodné namísto celkových hodnot za celou sezónu použít průměrné hodnoty za určitý časový interval (např. průměrný počet přihrávek za 90 minut). V opačném případě by výsledky byly zkresleny hráči, kteří strávili ve hře méně času.

Veškeré výpočty předvedené v této kapitole byly prováděny v programovacím jazyku Python s pomocí webového rozhraní Jupyter [16]. Kromě základní knihovny Pythonu bylo použito také několik doplňkových balíčků - knihoven.

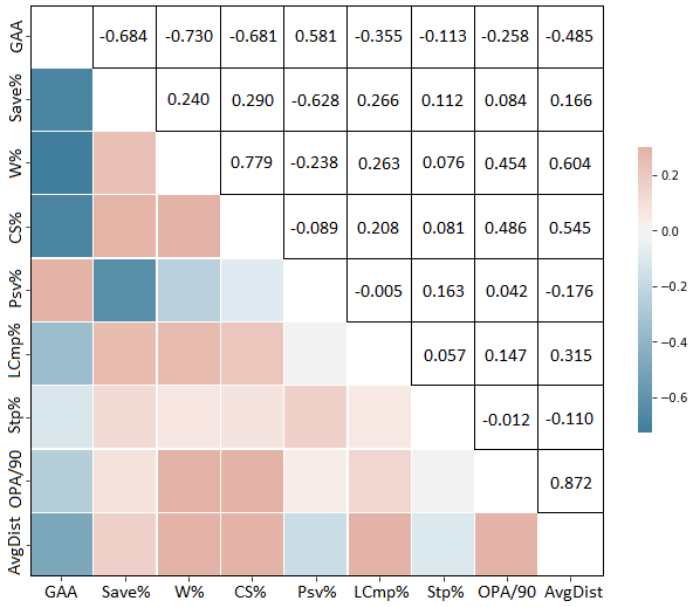
Prvním z nich je knihovna NumPy [17]. Tato knihovna slouží pro numerické výpočty v Pythonu. Mimo jiného umí pracovat s vícerozměrnými poli a obsahuje základy lineární algebry. Knihovna pandas [18] je statistická knihovna umožňující uživateli vytvářet datové struktury (DataFrame). S daty je díky ní možné jednoduše manipulovat a analyzovat je. K vizualizaci dat byla použita knihovna Plotly Express [19], část knihovny Plotly. Dalšími balíčky na tvorbu vizualizací jsou knihovna matplotlib [20] a seaborn [21]. K nejdůležitějším balíčkům použitým v této práci patří knihovna scikit-learn [22], a to konkrétně její modul sklearn.decomposition, který obsahuje veškeré nástroje pro tvorbu hlavních komponent, a knihovna factor_analyzer [1], která se zabývá tvorbou faktorů.

4.1. Soubor brankářských statistik

4.1.1. Analýza PCA

Analýza hlavních komponent je prvním krokem a zároveň i jedním z nejlepších nástrojů k redukci dimenze vícerozměrných úloh. Jako statistický soubor, na kterém budou postupy analýzy hlavních komponent představeny, bude použit soubor brankářských statistik. Jak již bylo zmíněno, do tohoto souboru jsou zahrnuti pouze brankáři, kteří odehráli minimálně 900 minut. Brankářů, kteří tento požadavek v německé Bundeslize v sezóně 2018/2019 splnili, bylo 23. Kromě osmnácti týmových „jedniček“ (tj. nejdůležitějších brankářů, již jsou ve svém týmu nejvíce vytěžováni), je do tohoto souboru zahrnuto ještě pět dalších brankářů, kteří také hráli ve svém klubu výraznou roli.

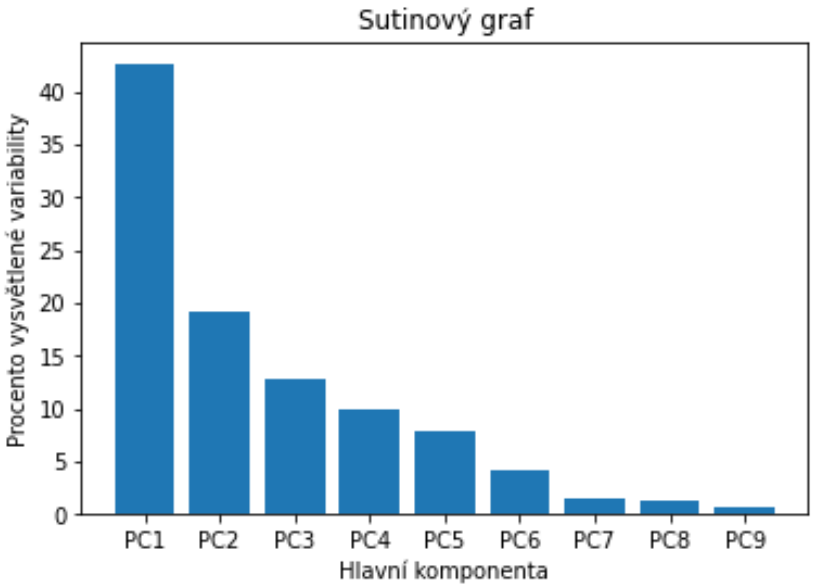
Před samotnou tvorbou hlavních komponent je vhodné připomenout, jakou strukturu komponenty vlastně mají. Hlavní komponenty vycházejí z korelační matice původních proměnných a jsou tvořeny jako lineární kombinace původních proměnných. Pro ilustraci budou nyní uvedeny korelační koeficienty mezi zkoumanými znaky.



Obrázek 4.1: Korelační koeficienty brankářských statistik

Zabývat se maticí korelace má smysl před řešením každého příkladu. I když teoreticky se dají hledat hlavní komponenty i pro soubor se slabými korelacemi, pro řešení, jehož výsledky mají mít reálné využití, bude požadováno, aby alepsoň některé znaky byly mezi sebou korelovány významněji. Ve výše uvedené tabulce korelačních koeficientů (tab. 4.1) lze tyto korelace pozorovat. Nejsilnější korelace je mezi znaky *OPA/90* a *AvgDist*, a to zaokrouhleně 0,87. Silnou korelaci mají ještě znaky *W%* a *CS%* nebo *W%* a *GAA*. Záporná hodnota korelačního koeficientu (v případě korelace mezi *W%* a *GAA*) značí negativní korelaci. Slabě korelované jsou například znaky *OPA/90* a *Stp%* s korelačním koeficientem zaokrouhleně 0,01.

Dalším krokem již bude hledání komponent. Nejdříve budou nalezeny všechny komponenty, poté se na základě výsledků rozhodne, které komponenty budou použity. Výpočet proběhl v Pythonu (viz příloha).



Obrázek 4.2: Variabilita brankářských statistik vysvětlená jednotlivými komponentami

Na (sutinovém) grafu (obr. 4.2) lze vidět podíl celkové variability, který tyto komponenty vysvětlují. První hlavní komponenta vysvětluje 42,6 % rozptylu, druhá hlavní komponenta vysvětluje 19,1 % rozptylu a tak dále.

Nyní je třeba rozhodnout, kolik komponent z celkového počtu (9) bude vybráno za reprezentanty zkoumaného souboru. Nelze jednoznačně určit, jaký počet reprezentantů je považován za správný. Obecně platí, že je volen takový počet komponent, jenž vysvětlí dostatečně velkou část variability. Část variability, která se dá považovat za dostatečně velkou, se však liší příklad od příkladu. Počet reprezentantů je volen tak, aby výsledné řešení bylo co nejjednodušší. Pokud je za reprezentanty vybrán velký počet komponent, celá metoda redukce dimenzionality ztrácí význam. V praxi bývá často používáno pravidlo, že za reprezentanty daného souboru jsou voleny ty komponenty, které mají vlastní číslo větší než 1. Množství vysvětlené variability každým zkoumaným znakem je ekvivalentní právě komponentě s vlastním číslem 1. Toto pravidlo tedy říká, že je vhodné použít jen ty komponenty, které vysvětlí více variability než původní znaky. Z hlediska grafického vyjádření je ideální, aby byl počet reprezentantů 2-3. Řešení lze potom vyjádřit pomocí dvourozměrného nebo trojrozměrného grafu. V ilustračním příkladě splňují pravidlo vlastního čísla většího než 1 první tři komponenty. Obsah celkové variability, kterou vysvětlují, je zhruba 75 %, což lze považovat za dostatečné. Zároveň bude možné výsledky vhodně zobrazit.

Komponentní zátěže a komunalita

Zbývá podrobněji analyzovat nalezené komponenty. V následující tabulce jsou k nahlédnutí komponentní zátěže a komunalita prvních tří komponent.

	PC1	PC2	PC3	Komunalita
GAA	-0.88	0.36	0.09	0.92
Save%	0.57	-0.67	-0.05	0.77
W%	0.84	0.15	-0.09	0.74
CS%	0.81	0.20	-0.15	0.71
Psv%	-0.44	0.71	-0.39	0.86
LCmp%	0.42	-0.00	-0.28	0.26
Stp%	0.06	-0.05	-0.89	0.80
OPA/90	0.63	0.61	0.17	0.79
AvgDist	0.78	0.45	0.25	0.87

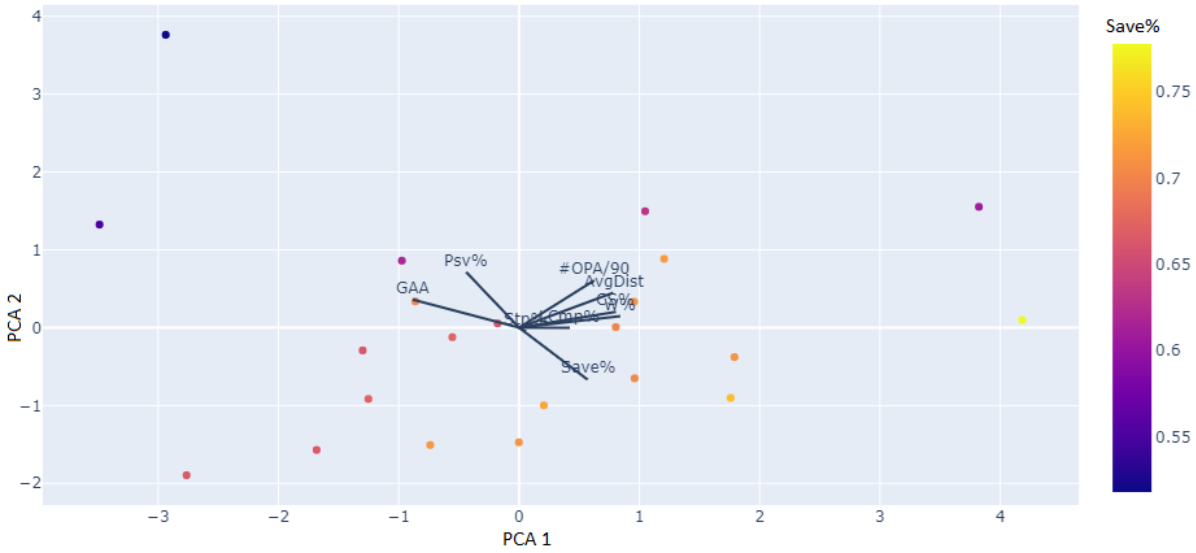
Tabulka 4.1: Komponentní zátěže a komunalita brankářských statistik

Hodnoty v prvních třech sloupcích tabulky 4.1 jsou komponentní zátěže pro jednotlivé hlavní komponenty. Řádky pak označují pozorované znaky. V posledním sloupci je komunalita pro první tři komponenty. Způsob, jakým lze v Pythonu získat komponentní zátěže celého zkoumaného souboru, je uveden na odkazu [3].

Tabulka komponentních zátěží a komunality představuje zřejmě nejlepší zdroj informací pro interpretaci hlavních komponent. Můžeme pozorovat vzájemné korelace mezi libovolnými komponentami a pozorovanými znaky. Zároveň díky komunalitě vidíme, jak velká část rozptylu původních znaků je ve vybraných komponentách obsažena. Za zmínku stojí komunalita znaku *LCmp%* (úspěšnost přihrávek na dlouhou vzdálenost). Oproti ostatním znakům je jeho komunalita výrazně nižší. Důvodem je, že tento znak nemá s ostatními znaky mnoho společného, a analýza PCA se snaží tvořit komponenty tak, aby bylo zachováno co nejvíce celkové variability. Obecně je za nízkou komunalitu považována komunalita nižší než 0,4 - 0,5.

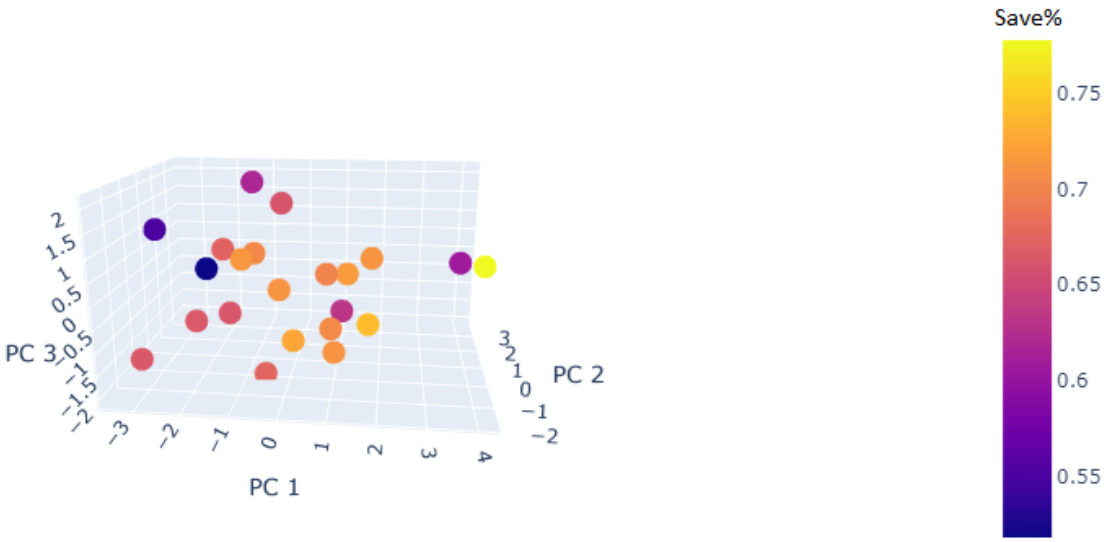
Znaménka u komponentních zátěží mají význam vztahu mezi komponentami a vstupními znaky. Na množství zachované variability však nemají žádný vliv. A protože je veličina komunalita definovaná jako suma čtverců komponentních zátěží, nemají vliv ani na ni.

Závěr této sekce bude věnován grafickému vyjádření výsledků. Na obrázcích 4.3 a 4.4 jsou zobrazeny všechny prvky zkoumaného statistického souboru a jejich ohodnocení nově vzniklými komponentami. Brankáři jsou označeni barevně podle jejich úspěšnosti zákroků.



Obrázek 4.3: Brankáři a jejich hodnoty 1. a 2. komponenty

Na prvním obrázku 4.3 jsou zobrazena měření (tedy brankáři) v prostoru prvních dvou komponent. Uprostřed obrázku je pro lepší představu přidána legenda vysvětlující váhu původních znaků (komponentní zátěže) pro výpočet hodnot těchto komponent. K lepšímu pochopení vztahů slouží také „obarvení“ prvků původního souboru podle jednoho z jejich původních znaků (*Save%*). Lze vidět, že brankáři, kteří měli hodnotu *Save%* vyšší, mají obecně také vyšší hodnotu u první komponenty a nižší hodnotu u druhé komponenty. Což souhlasí s přidanou legendou. Vysoká hodnota *Save%* nemusí být nutně důvodem vysoké hodnoty první komponenty. Například brankář mající druhou největší hodnotu první komponenty, zhruba 3,8, (Manuel Neuer) má úspěšnost zákroků (*Save%*) pouze asi 61 %. Působí však v týmu, který je v lize dominantní a nedovoluje soupeři mnoho střeleckých pokusů (Bayern Mnichov). Tím pádem má vysoce nadprůměrné statistiky *W%*, *CS%* a naopak velmi malou statistiku *GAA* - tedy faktory způsobující vysokou hodnotu první komponenty. Dá se proto říct, že hodnoty první komponenty jsou dány nejen schopnostmi daného brankáře, ale jsou také silně ovlivňovány schopnostmi celého týmu.



Obrázek 4.4: Brankáři a jejich hodnoty 1.,2. a 3. komponenty

Obrázek 4.4 ukazuje navíc hodnoty zkoumaných prvků statistického souboru pro třetí hlavní komponentu.

4.1.2. Faktorová analýza

Cílem této sekce bude ukázat použití faktorové analýzy na konkrétním statistickém souboru. Jelikož je jako metoda extrakce faktorů použita analýza hlavních komponent, lze tuto sekci brát jako rozšíření dosavadní části kapitoly. Následná rotace bude provedena na prvních třech faktorech (komponentách). Důvody, proč jsou za reprezentanty zkoumaného souboru voleny právě 3 komponenty, jsou vysvětleny v sekci 4.1.

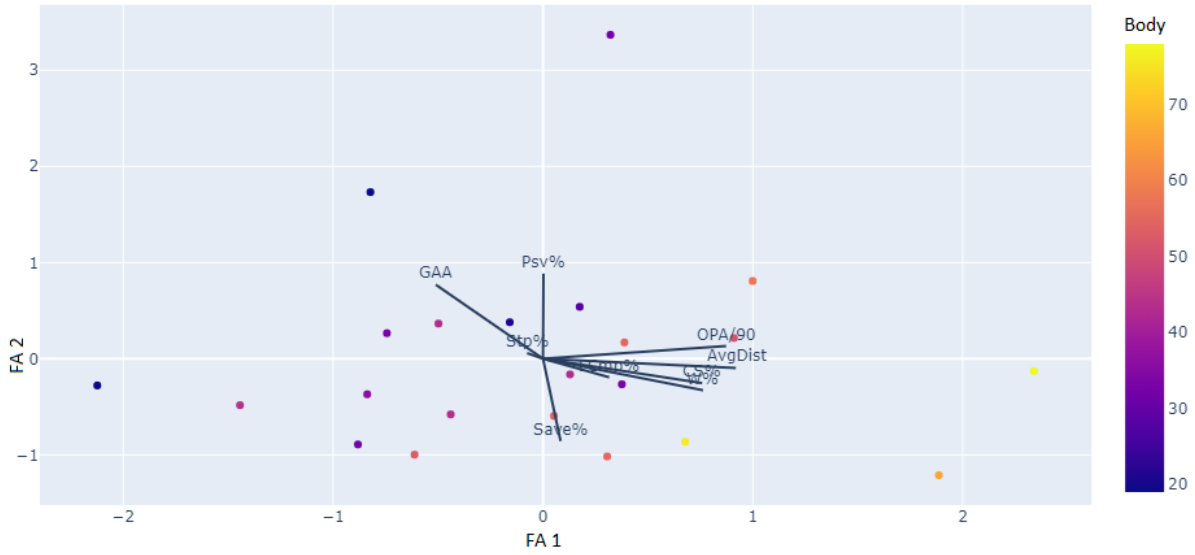
Varimax

Nyní už k samotným rotacím. Nejprve budou představeny a analyzovány výsledky získané ortogonální metodou varimax.

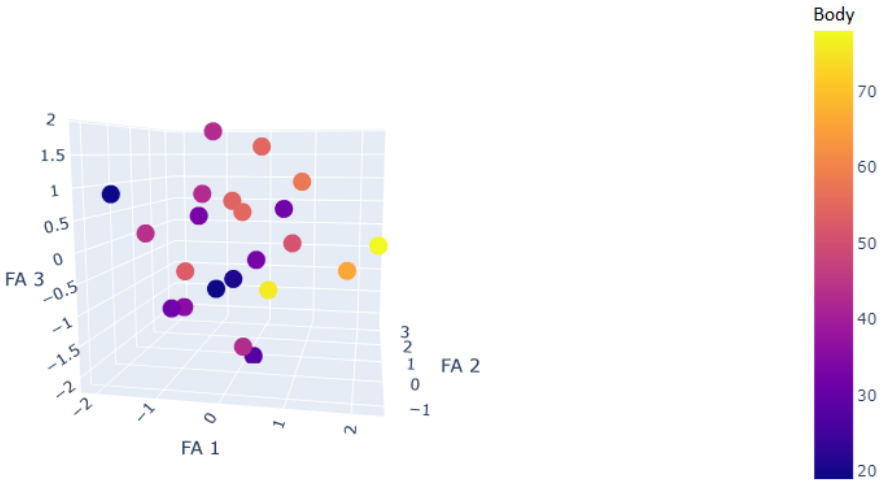
	F1	F2	F3	Komunalita
GAA	-0.51	0.77	-0.26	0.92
Save%	0.08	-0.86	0.19	0.77
W%	0.76	-0.33	0.22	0.74
CS%	0.76	-0.25	0.27	0.71
Psv%	0.00	0.89	0.27	0.86
LCmp%	0.32	-0.19	0.35	0.26
Stp%	-0.08	0.06	0.89	0.80
OPA/90	0.87	0.13	-0.10	0.79
AvgDist	0.92	-0.10	-0.14	0.87

Tabulka 4.2: Faktorové zátěže a komunalita brankářských statistik po rotaci varimax

V tabulce 4.2 lze pozorovat faktorové zátěže rotovaných faktorů a zkoumaných znaků. Díky použití faktorové analýzy nyní faktory korelují s některými zkoumanými znaky výrazněji a s ostatními naopak téměř vůbec. Jelikož je cílem faktorové analýzy zjednodušit jejímu uživateli interpretaci původního souboru, bude ukázka takové interpretace později uvedena i v této práci. Nyní však ještě zbývá vyjádřit výsledky metody varimax graficky. Na rozdíl od grafického znázornění metody hlavních komponent, kde byli brankáři označováni podle úspěšnosti zákroků, budou při znázornění výsledků faktorové analýzy brankáři označováni podle počtu bodů týmu, v němž hráli.



Obrázek 4.5: Brankáři a jejich hodnoty 1. a 2. faktoru



Obrázek 4.6: Brankáři a jejich hodnoty 1., 2. a 3. faktoru

Výhody faktorové analýzy jsou patrné jak z tabulky 4.2, tak i z obrázku 4.5 . Směry faktorových zátěží různých znaků se nyní mnohem více blíží směrům jednotlivých faktorů. Tím pádem je pro uživatele jednodušší danému modelu porozumět.

Konkrétně pro zkoumanou úlohu platí, že po rotaci ztrácí první faktor velkou část závislosti na znacích *Save%* a *GAA*. Znak *Psv%* pak má s tímto faktorem faktorovou zátěž dokonce menší než 0,005. Zjednodušeně lze tedy první faktor považovat za faktor, jenž je dán kombinací 4 znaků - *W%*, *CS%*, *OPA/90* a *AvgDist*. První dva zmíněné znaky *W%* a *CS%* jsou obecné statistiky, které jsou ale z velké části dány týmem, ve kterém jednotliví brankáři hrají. Další dva znaky *OPA/90* a *AvgDist* pak závisí na vzdálenosti od brány, ve kterých brankář provádí zákroky. Do jisté míry i tyto dva znaky vycházejí ze síly týmu, ve kterém brankář působí. Silnější týmy kontrolují většinu zápasu hru a jejich brankáři tedy nemusejí stát v bráně, ale mohou pomáhat s rozehrávkou. Pokud už mají provést obranný zákrok, většinou jde o akci z protiútoku nebo střelu ze střední vzdálenosti, kdy se snaží vyběhnutím snížit střelecký úhel. První faktor se tedy dá interpretovat jako síla týmu, v němž brankář působí. O tom svědčí i fakt, že mezi 5 brankáři, již mají hodnotu u prvního faktoru nejvyšší, je brankář Mnichova (Manuel Neuer), Dortmundu (Roman Bürki), Lipska (Péter Gulácsi) i Leverkusenu (Lukáš Hradecký) - tedy čtyř nejúspěšnějších týmů Bundesligy v sezóně 2018/2019.

Druhý faktor je závislý zejména na znacích *Save%*, *Psv%* a *GAA*. Statistika *GAA* udává, kolik dostal průměrně daný brankář za zápas gólů - tato statistika je opět ovlivněna týmem. Statistiky *Save%* a *Psv%* už jsou však dány hlavně schopnostmi brankáře. Jedná se tedy o faktor, který vyjadřuje převážně individuální dovednosti brankářů.

V tabulce faktorových zátěží je vidět, že třetí faktor výrazněji závisí pouze na jednom znaku. Důvodem je, že většina rozptylu původních znaků, které spolu do nějaké míry korelují, byla už vysvětlena. V tomto případě se jedná o původní znak *Stp%*, tj. procento křižných přihrávek soupeře do pokutového území, které brankář úspěšně překazil. Žádný jiný znak není tímto faktorem výrazněji vysvětlen.

Quartimax

Na závěr budou výsledky srovnány s další metodou rotace faktorů, a to s rotací quartimax.

	F1	F2	F3	Komunalita
GAA	-0.53	0.77	-0.19	0.92
Save%	0.11	-0.86	0.14	0.77
W%	0.78	-0.32	0.17	0.74
CS%	0.77	-0.25	0.22	0.71
Psv%	0.00	0.87	0.31	0.86
LCmp%	0.33	-0.20	0.32	0.26
Stp%	-0.03	0.01	0.90	0.80
OPA/90	0.86	0.15	-0.13	0.79
AvgDist	0.91	-0.07	-0.19	0.87

Tabulka 4.3: Faktorové zátěže a komunalita brankářských statistik po rotaci quartimax

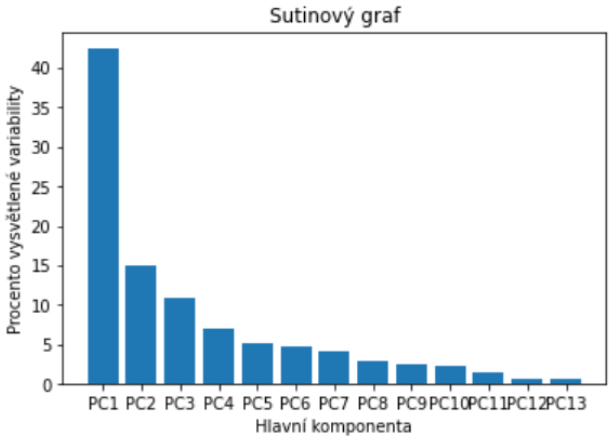
Výsledky získané metodou quartimax jsou velmi podobné výsledkům získaným metodou varimax. Největší rozdíl je u znaku *GAA* a třetího faktoru, kde se číselné hodnoty zátěží obou metod liší o 0,07. Vzhledem k tomu, že obě tyto metody fungují na podobném principu, nejde o nic neobvyklého. Jak metoda varimax, tak i quartimax se snaží maximalizovat rozptyl druhých mocnin faktorových zátěží. Rozdíl je v postupu, který daná metoda na výpočet používá. Příkladem souboru, u kterého se liší výsledky získané oběma metodami, je soubor týmových statistik v sekci 4.3.

4.2. Soubor hráčských statistik

V této části kapitoly budou představeny výsledky obou metod redukce dimenze na souboru hráčských statistik. Postupy používaných metod byly vysvětleny už v minulé sekci, proto se sekce venované souborům hráčských a týmových statistik těmito postupy zabývat nebudou a zaměří se především na interpretaci získaných výsledků. Soubor hráčských statistik je specifický svým rozsahem. Během sezóny 2018/2019 zasáhlo do hry na více než 900 minut (viz kapitola Hráčské a týmové statistiky) celkem 305 hráčů.

4.2.1. Analýza PCA

Stejně jako u souboru brankářských statistik i v tomto případě se začne tvorbou hlavních komponent a následným zvolením několika komponent za reprezentanty zkoumaného souboru. Z důvodu většího počtu zkoumaných znaků zde nebude uvedena tabulka vzájemných korelací, ale až sutinový graf samotných komponent.



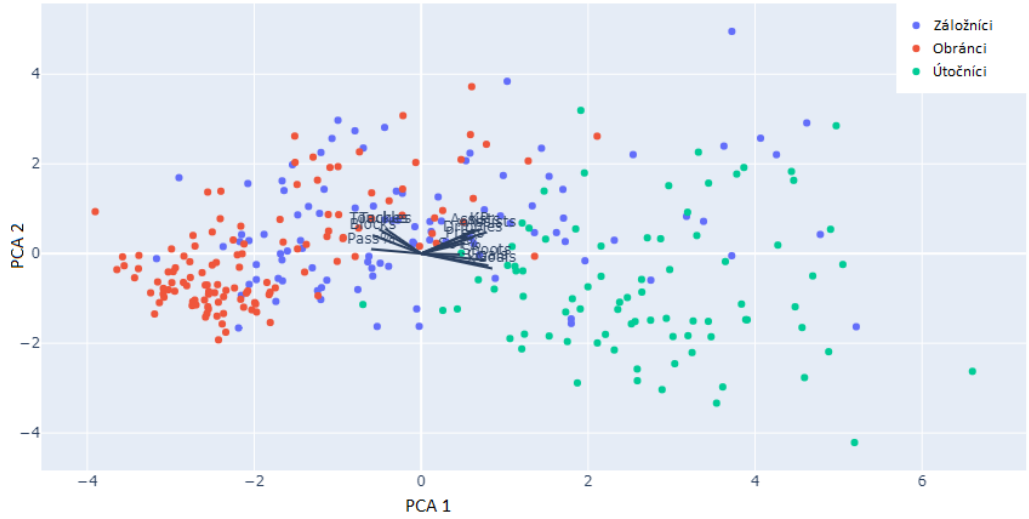
Obrázek 4.7: Variabilita hráčských statistik vysvětlená jednotlivými komponentami

Už na první pohled je zřejmé, že první hlavní komponenta bude mít při interpretaci souboru hráčských statistik ještě významnější roli než při interpretaci souboru brankářských statistik. I přes vyšší počet zkoumaných znaků je touto komponentou stále možné vysvětlit více než 40 % celkové variability. Zbytková variabilita je však více rozdělena mezi ostatní komponenty a druhá a třetí hlavní komponenta už nevysvětlují tolik celkové variability jako ekvivalentní komponenty v předcházejícím souboru. I přesto jsou to opět právě první tři komponenty, které mají hodnotu vlastního čísla větší než 1. Dohromady je jimi vysvětleno 68 % variability celého souboru. Interpretace tohoto souboru bude tedy probíhat podobně jako u souboru brankářských statistik.

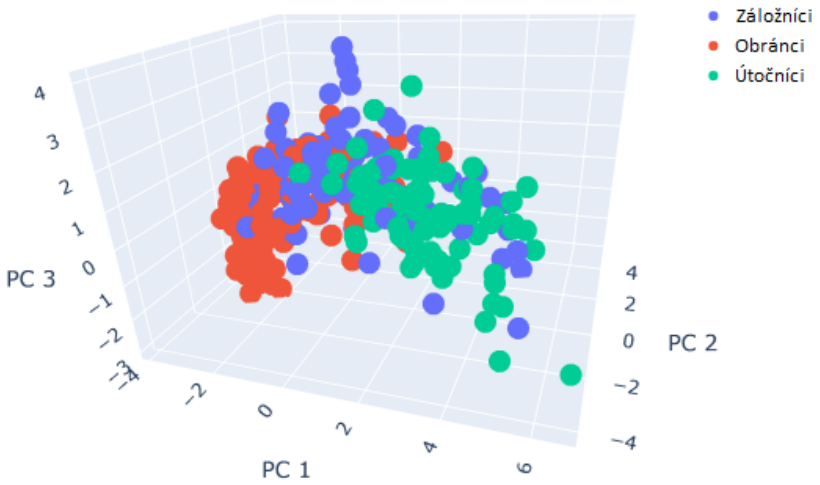
	PC1	PC2	PC3	Komunalita
Goals	0.80	-0.27	-0.16	0.74
Assists	0.64	0.52	-0.18	0.71
xGoals	0.85	-0.33	-0.10	0.85
xAssists	0.79	0.48	-0.12	0.86
Shoots	0.78	-0.15	-0.21	0.67
SoT%	0.46	-0.02	0.02	0.22
Pass%	-0.60	0.10	-0.59	0.72
KP	0.70	0.54	-0.08	0.79
Tackles	-0.43	0.55	0.40	0.65
Press	0.53	0.23	0.63	0.73
Blocks	-0.58	0.41	0.25	0.57
Touches	-0.50	0.54	-0.57	0.87
Dribbles	0.62	0.36	-0.04	0.52

Tabulka 4.4: Komponentní zátěže a komunalita hráčských statistik

Z tabulky 4.4 lze vyčíst vztahy mezi původními znaky a vzniklými komponentami. Většina původních znaků je v nově vzniklém prostoru vysvětlena poměrně dobře - pouze 3 znaky z 13 mají komunalitu nižší než 0,6. Výrazně nižší komunalitu pak má pouze znak *SoT%*, jehož variabilita je novým modelem vysvětlena jen z 22 %. Na obrázcích 4.8 a 4.9 jsou k vidění prvky souboru hráčských statistik (hráči) a jejich hodnoty pro jednotlivé komponenty. Barevné značení vyjadřuje pozice na kterých hráči hrají.



Obrázek 4.8: Hráči a jejich hodnoty 1. a 2. komponenty



Obrázek 4.9: Hráči a jejich hodnoty 1.,2. a 3. komponenty

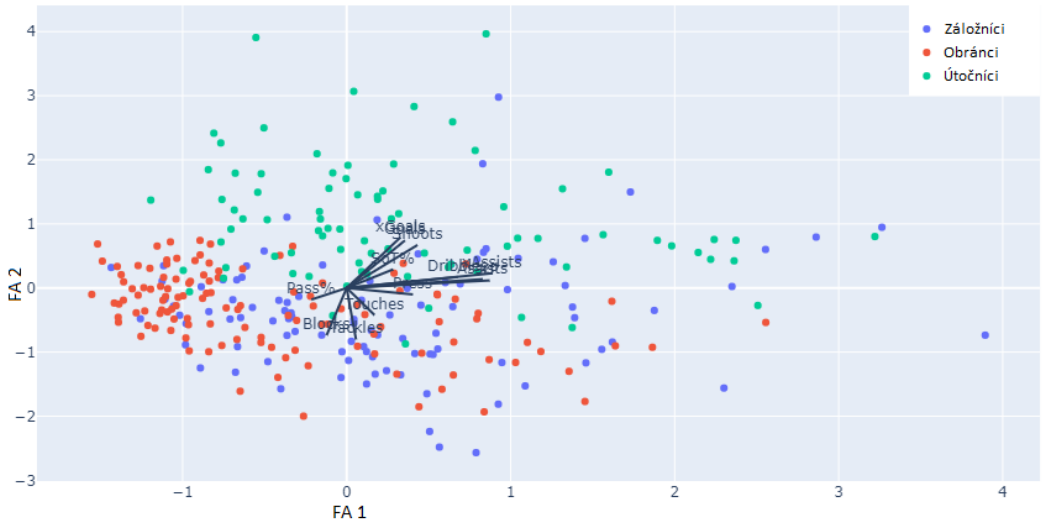
4.2.2. Faktorová analýza

Po úspěšném použití metody hlavních komponent následuje rotace nalezených faktorů. Takto získané výsledky budou poté opět podrobněji rozebrány a porovnány s výsledky získanými metodou hlavních komponent.

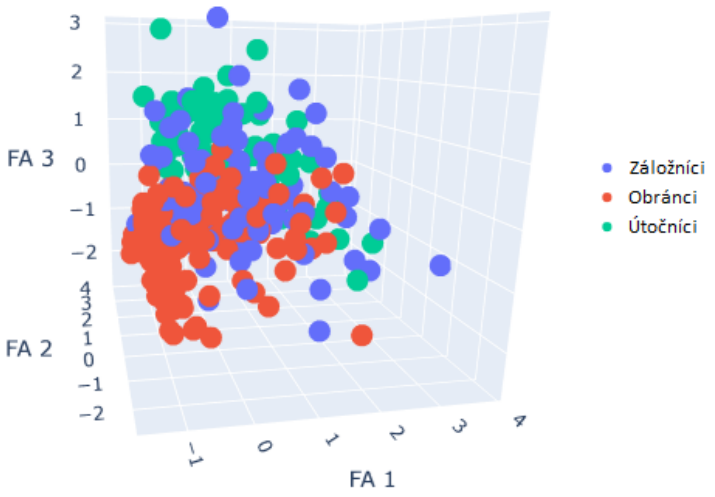
Varimax

	F1	F2	F3	Komunalita
Goals	0.36	0.74	0.26	0.74
Assists	0.83	0.14	0.02	0.71
xGoals	0.33	0.79	0.34	0.85
xAssists	0.89	0.23	0.15	0.86
Shoots	0.43	0.67	0.18	0.67
SoT%	0.28	0.29	0.23	0.22
Pass%	-0.21	-0.18	-0.80	0.72
KP	0.87	0.12	0.14	0.79
Tackles	0.06	-0.80	0.07	0.65
Press	0.40	-0.10	0.75	0.73
Blocks	-0.12	-0.74	-0.10	0.57
Touches	0.17	-0.42	-0.81	0.87
Dribbles	0.68	0.17	0.17	0.52

Tabulka 4.5: Faktorové zátěže a komunalita hráčských statistik po rotaci varimax



Obrázek 4.10: Hráči a jejich hodnoty 1. a 2. komponenty



Obrázek 4.11: Hráči a jejich hodnoty 1.,2. a 3. komponenty

První - nejdůležitější - faktor nově vzniklého modelu vysvětluje zhruba 26,5 % celkové variability. Další dva faktory vysvětlují 25 % resp. 17 % celkové variability. Oproti modelu získanému analýzou hlavních komponent je tedy variabilita mezi faktory rozdělena rovnoměrněji. První faktor nejvíce závisí na statistikách *Assists* (asistence), *xAssists* (očekávané asistence) a *KP* (klíčové přihrávky). Vysoký podíl rozptylu těchto tří znaků v jednom faktoru není žádným překvapením. Statistiky *Assists*, *xAssists* i statistika *KP* spolu totiž úzce souvisí. Základním prvkem, který všechny tyto tři statistiky ovlivňuje, je přihrávání spoluhráčům při útočných akcích. Faktorové zátěže všech tří znaků dosahují hodnot vyšších než 0,8. Zajímavé je pozorovat znak *Pass%* (přesnost přihrávek). Díky vysokým faktorovým zátěžím znaků *Assists*, *xAssists* a *KP* by se mohlo zdát, že první faktor obecně závisí na statistikách zabývajících se přihrávkami. Hodnota faktorové zátěže znaku *Pass%* je však pro první faktor malá. Možným vysvětlením tohoto výsledku je to, že první faktor se nezabývá ani tolik přihrávkami obecně, ale spíše přihrávkami vedoucími k zakončení. Hráči s vyšším počtem asistencí totiž často zkouší i riskantní přihrávky a jejich úspěšnost tedy nemusí být vždy vysoká. Kromě tří výše zmíněných znaků *Assists*, *xAssists* a *KP* má větší faktorovou zátěž u prvního faktoru ještě znak *Dribbles* - úspěšné kličkování. Největší hodnotu prvního faktoru pak mají hráči, kteří dokáží vymyslet rozhodující přihrávky, ale zároveň jsou schopní obehrát protihráče. Hráči, kteří tyto požadavky splňují, jsou převážně ofenzivní záložníci a techničtí útočníci, naopak obránci mají hodnotu prvního faktoru nízkou (viz obrázek 4.10). První faktor tedy lze interpretovat jako hodnocení kreativních schopností hráčů.

Druhý faktor závisí hlavně na znacích *Goals*, *xGoals*, *Shoots*, *Tackles* a *Blocks*. Znaménka faktorových zátěží prvních tří zmíněných znaků jsou kladná, znaménka faktorových zátěží znaků *Tackles* a *Blocks* jsou záporná. Statistiky *Goals*, *xGoals* a *Shoots* jsou čistě útočné statistiky. Statistiky *Tackles* a *Blocks* naopak čistě defenzivní. Největší hodnotu druhého faktoru tedy mají hráči, kteří jsou zvyklí často a úspěšně střílet a nevěnují příliš pozornosti obranným činnostem - převážně střední útočníci. Nejmenší hodnotu (nejvíce zápornou) druhého faktoru naopak mají hráči, kteří se věnují bránění a do útočných akcí se nezapojují - obránci a defenzivní záložníci. Druhý faktor lze tedy jednoznačně interpretovat jako škálu, která rozděluje hráče na ofenzivní hráče (kladná hodnota FA 2) a defenzivní hráče (záporná hodnota FA 2).

Třetí faktor závisí hlavně na znacích *Pass%*, *Touches* a *Press*. Korelace znaků *Pass%*, *Touches* a třetího faktoru má kladné znaménko, korelace znaku *Press* s tímto faktorem pak má znaménko záporné. Interpretace tohoto znaku už není tak jednoduchá. Znaky *Pass%*, *Touches* a *Press* mezi sebou nekorelují tak silně jako některé znaky u předchozích faktorů a určité souvislosti je potřeba si domyslet. Vysokou hodnotu *Pass%* a *Touches* mají obecně hráči, kteří tvoří jádro týmu a kteří se často podílí na tvorbě hry. Statistika *Press* pak závisí na fyzické kondici hráče, roli v týmu a celkově na týmové strategii. Největších hodnot třetího faktoru tedy dosahují hráči, kteří hrají poziční hru. Nejmenších hodnot dosahují naopak hráči, kteří využívají naplno své fyzické vytrvalosti. Problémem je interpretovat hráče, kteří mají průměrné hodnoty třetího faktoru. Kvůli tomu, že faktorové zátěže znaků *Pass%*, *Touches* a znaku *Press* směřují opačnými směry, nelze pouze z hodnoty třetího faktoru posoudit, zda je daný hráč schopen tvořit hru a zároveň mít dobrou fyzickou kondici, nebo je naopak v těchto atributech slabý.

Quartimax

Na závěr budou výsledky opět srovnány s rotací quartimax.

	F1	F2	F3	Komunalita
Goals	0.38	0.75	0.21	0.74
Assists	0.83	0.13	-0.04	0.71
xGoals	0.35	0.80	0.29	0.85
xAssists	0.90	0.23	0.08	0.86
Shoots	0.45	0.67	0.13	0.67
SoT%	0.30	0.30	0.20	0.22
Pass%	-0.27	-0.21	-0.78	0.72
KP	0.88	0.12	0.08	0.79
Tackles	0.06	-0.80	0.10	0.65
Press	0.45	-0.08	0.72	0.73
Blocks	-0.13	-0.74	-0.06	0.57
Touches	0.11	-0.45	-0.81	0.87
Dribbles	0.69	0.17	0.12	0.52

Tabulka 4.6: Faktorové zátěže a komunalita hráčských statistik po rotaci quartimax

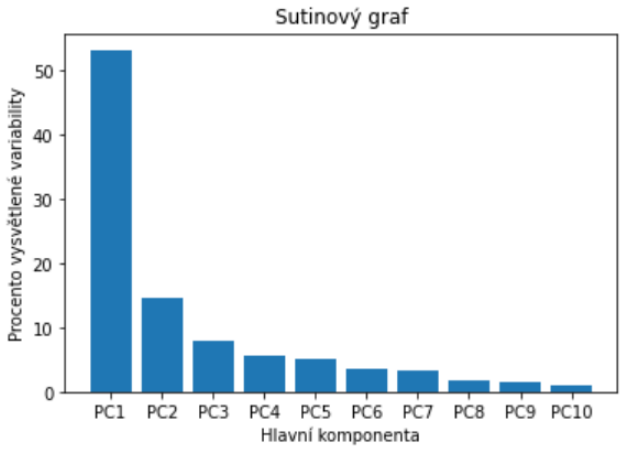
Stejně jako u souboru brankářských statistik i u souboru hráčských statistik vyšly výsledky získané rotací quartimax velmi podobně jako výsledky získané metodou varimax. Největší rozdíl lze pozorovat u znaku *xAssists* a třetího faktoru, kde se číselné hodnoty zátěží obou metod liší o 0,07.

4.3. Soubor týmových statistik

Poslední soubor, na kterém budou metody použity, je soubor týmových statistik. Tento soubor je svým rozsahem nejmenší. Pozorováno bude pouze 18 statistických jednotek - týmů. Na druhou stranu je tento soubor popsán největším počtem znaků - 27.

4.3.1. Analýza PCA

Jako u předchozích souborů, tak i u souboru týmových statistik se začne nejpve metodou hlavních komponent. Z důvodu velkého množství pozorovaných znaků je v sutinovém grafu zobrazeno pouze prvních 10 komponent a opět zde nebude uvedena tabulka vzájemných korelací. Soubor týmových statistik je specifický tím, že má větší počet znaků n než počet pozorování p. V takovém případě se projeví úkaz zmíněný na konci sekce 2.1.1, kdy je maximální počet komponent roven počtu pozorování.



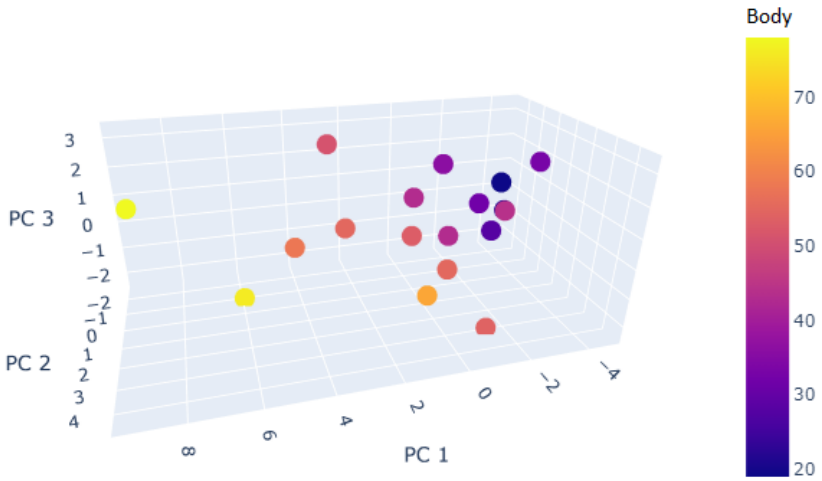
Obrázek 4.12: Variabilita týmových statistik vysvětlená jednotlivými komponentami

První komponenta má z hlediska vysvětlené variability ještě významnější roli než u předchozích souborů. Přestože je pozorováno 27 znaků a celkový rozptyl je rozdělen mezi 18 komponent, dokáže první komponenta vysvětlit celých 53 % původní variability. Další komponenty, které mají vlastní číslo větší než 1, jsou druhá komponenta (vysvětluje zhruba 15 % celkové variability), třetí komponenta (8 %), čtvrtá komponenta (5,5 %) a pátá komponenta (5 %). Dohromady je těmito komponentami vysvětleno zhruba 87 % celkové variability, což je ze všech 3 zkoumaných souborů největší hodnota. Použití pěti hlavních komponent namísto tří však mírně snižuje jednoduchost interpretace. Číselně lze výsledky popsat stejným způsobem jako u předchozích souborů. Graficky je však složitější vyjádřit všechny vztahy, a proto budou ukázány pouze výsledky pro tři nejdůležitější komponenty.

	PC1	PC2	PC3	PC4	PC5	Komunalita
Goals	0.93	0.27	-0.07	0.12	-0.09	0.96
Assists	0.94	0.20	-0.13	0.09	-0.08	0.96
Pkmade	0.07	0.57	0.55	0.29	0.31	0.82
Pkatt	0.00	0.39	0.63	0.26	0.56	0.94
CrdY	-0.67	0.42	0.39	-0.21	-0.22	0.87
CrdR	-0.38	-0.19	-0.05	-0.68	0.09	0.66
xGoals	0.90	0.25	0.20	-0.04	-0.21	0.96
xAssists	0.91	0.22	0.09	-0.06	-0.26	0.96
GoalsAgainst	-0.77	-0.39	0.07	-0.05	-0.20	0.78
CleanSheet	0.59	0.40	-0.27	-0.03	0.29	0.66
Sh	0.82	0.04	0.41	-0.04	-0.35	0.97
SoT	0.89	0.11	0.24	0.13	-0.29	0.97
G/Sh	0.73	0.34	-0.45	0.23	0.00	0.90
G/SoT	0.70	0.36	-0.55	0.13	0.13	0.96
FreeKicks	0.61	-0.16	0.26	-0.22	0.06	0.52
Cmp	0.94	-0.18	-0.09	-0.06	0.07	0.93
Att	0.94	-0.13	-0.13	-0.07	0.03	0.93
Cmp%	0.86	-0.31	0.04	0.01	0.19	0.87
KP	0.81	0.12	0.40	-0.06	-0.34	0.96
Tackles	-0.56	0.72	0.10	-0.21	-0.03	0.89
TacklesW	-0.38	0.69	-0.13	-0.22	-0.03	0.69
Press	-0.74	0.54	-0.15	0.18	0.07	0.89
PressW	0.12	0.89	-0.25	-0.23	-0.14	0.94
Blocks	-0.61	0.46	-0.11	-0.25	-0.14	0.68
Touches	0.94	-0.10	-0.13	-0.11	0.02	0.92
DribblesW	0.77	0.01	0.11	-0.46	0.33	0.92
Dribbles	0.71	0.02	0.13	-0.46	0.34	0.84

Tabulka 4.7: Komponentní zátěže a komunalita týmových statistik

V tabulce 4.7 lze pozorovat komponentní zátěže a komunalitu znaků u prvních pěti komponent. Jak už bylo dříve zmíněno, model tvořený prvními pěti komponentami vysvětluje velkou část celkové variability (87 %). S tím souvisí i vysoké hodnoty komunality pro jednotlivé znaky. Ani jeden ze zkoumaných znaků nemá komunalitu nižší než 0,5 a pouze jeden znak má komunalitu nižší než 0,6. Na druhou stranu hned 15 znaků má komunalitu větší než 0,9.



Obrázek 4.13: Týmy a jejich hodnoty 1.,2. a 3. komponenty

Na obrázku 4.13 jsou znázorněny týmy (barevně označené podle počtu bodů získaných v sezóně 2018/2019) a jejich hodnoty prvních tří komponent.

4.3.2. Faktorová analýza

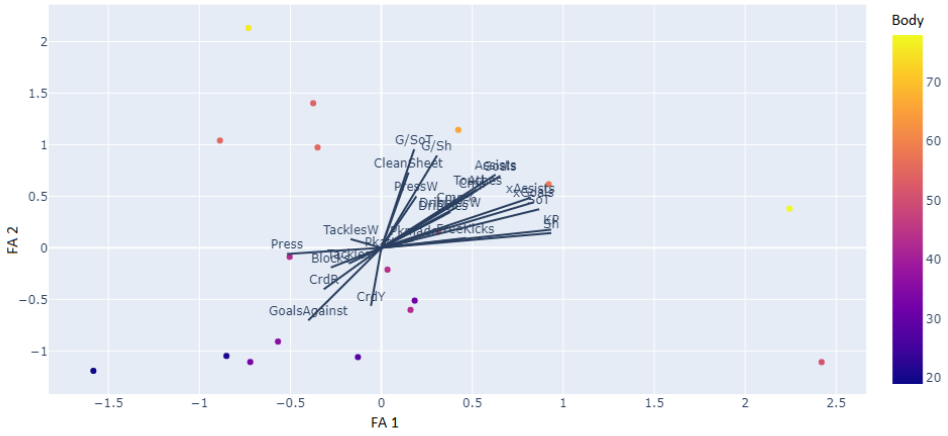
Po použití metody hlavních komponent následuje aplikace faktorové analýzy. V celé práci je jako metoda extrakce faktorů volena metoda hlavních komponent. Tak tomu bude i v tomto případě, kdy je po použití metody hlavních komponent extrahováno hned pět hlavních komponent (faktorů).

Varimax

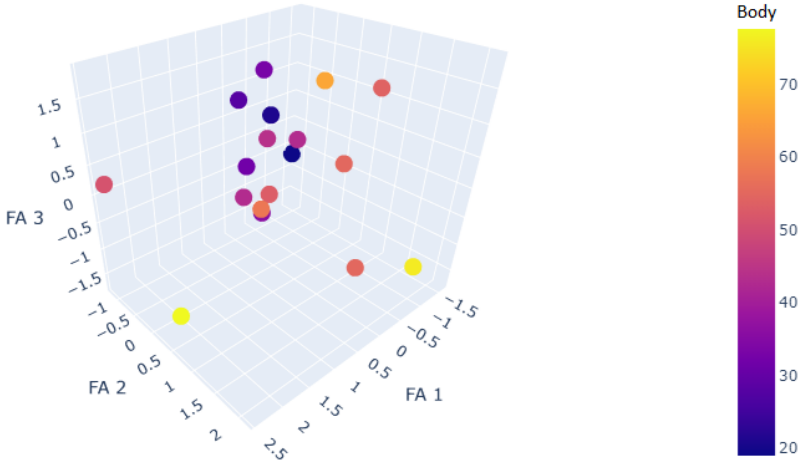
Nyní bude provedena rotace varimax a její výsledky budou opět interpretovány.

	F1	F2	F3	F4	F5	Komunalita
Goals	0.65	0.70	-0.16	0.12	0.08	0.96
Assists	0.63	0.71	-0.21	0.15	-0.00	0.96
Pkmade	0.18	0.07	0.24	-0.04	0.85	0.82
Pkatt	-0.00	-0.04	0.07	0.11	0.96	0.94
CrdY	-0.06	-0.56	0.71	-0.13	0.18	0.87
CrdR	-0.32	-0.40	0.23	0.48	-0.33	0.66
xGoals	0.84	0.44	-0.10	0.22	0.11	0.96
xAssists	0.82	0.49	-0.09	0.20	-0.01	0.96
GoalsAgainst	-0.40	-0.70	0.04	-0.25	-0.25	0.78
CleanSheet	0.15	0.73	0.06	0.29	0.14	0.66
Sh	0.93	0.14	-0.21	0.16	0.08	0.97
SoT	0.87	0.37	-0.24	0.05	0.10	0.97
G/Sh	0.31	0.89	-0.06	-0.03	-0.05	0.90
G/SoT	0.18	0.96	-0.02	0.08	-0.07	0.96
FreeKicks	0.46	0.09	-0.31	0.44	0.06	0.52
Cmp	0.50	0.53	-0.49	0.38	-0.09	0.93
Att	0.52	0.56	-0.44	0.36	-0.12	0.93
Cmp%	0.42	0.39	-0.63	0.39	0.04	0.87
KP	0.93	0.18	-0.14	0.17	0.10	0.96
Tackles	-0.18	-0.15	0.88	-0.07	0.23	0.89
TacklesW	-0.17	0.08	0.80	-0.03	0.07	0.69
Press	-0.52	-0.06	0.65	-0.40	0.20	0.89
PressW	0.19	0.50	0.81	0.05	0.01	0.94
Blocks	-0.28	-0.19	0.74	-0.10	-0.08	0.68
Touches	0.53	0.56	-0.40	0.38	-0.13	0.92
DribblesW	0.38	0.35	-0.20	0.78	0.08	0.92
Dribbles	0.34	0.31	-0.17	0.77	0.11	0.84

Tabulka 4.8: Faktorové zátěže a komunalita týmových statistik po rotaci varimax



Obrázek 4.14: Týmy a jejich hodnoty 1. a 2. faktoru



Obrázek 4.15: Týmy a jejich hodnoty 1.,2. a 3. faktoru

Po rotaci je prvním faktorem vysvětleno zhruba 26 % celkové variability. Největší vliv na hodnotu prvního faktoru pro dané týmy mají znaky: *Goals*, *Assists*, *xGoals*, *xAssists*, *Sh*, *SoT* a *KP* - znaménka faktorových zátěží jsou vždy kladná. Všechny tyto znaky jsou ofenzivní statistiky a první faktor lze tedy jednoznačně označit jako faktor vypovídající o ofenzivní síle týmu. Kromě 7 zmíněných znaků ještě mnoho dalších znaků přispívá značným dílem k celkové hodnotě první komponenty (11 znaků má faktorové zátěže mezi hodnotami 0,3 - 0,6). Snažit se však slovně interpretovat tak velké množství znaků je dosti obtížné, proto je interpretace tvořena jen z nejvýznamnějších znaků.

Druhý faktor, vysvětlující 23,9 % celkové variability, závisí převážně na znacích *Goals*, *Assists*, *GoalsAgainst*, *CleanSheet*, *G/Sh* a *G/SoT*. Kromě útočných statistik, kterými jsou *Goals* a *Assists*, jsou zde defenzivní statistiky *GoalsAgainst* a *CleanSheet* (vyjma znaku *GoalsAgainst* mají všechny znaky kladné faktorové zátěže). Namísto střeleckých statistik *Sh* a *SoT* pak druhý faktor zohledňuje úspěšnost těchto pokusů. Druhý faktor je tedy určen kombinací různorodých statistik a lze jej interpretovat jako celkovou sílu týmu.

Zajímavé je pozorovat vztahy prvního a druhého faktoru a počtu získaných bodů - nejdůležitější statistiky, která hodnotí, zdali je tým úspěšný, nebo ne. Přestože první faktor vysvětluje největší část celkového rozptylu, a dalo by se tedy očekávat, že bude nejlépe vystihovat celkovou sílu týmu, tak korelace mezi prvním faktorem a získanými body je poměrně slabá. Důvodem je, že v původním statistickém souboru je velké množství znaků hodnotících ofenzivní schopnosti. Tyto statistiky spolu silně korelují a první faktor jejich zahrnutím vysvětlí velkou část celkového rozptylu. Ve skutečnosti však velmi záleží i na obranných statistikách. Např. statistika *GoalsAgainst*, která udává počet obdržených branek, je pro úspěšný tým absolutně klíčová, avšak první faktor má s tímto znakem faktorovou zátěž pouze 0,4 (dalším důležitým znakem je např. znak *CleanSheet*, který má faktorovu zátěž s prvním faktorem dokonce pouze 0,15). Druhý faktor nezachovává takové množství celkové variability, avšak znaky na kterých nejvíce závisí, jsou zřejmě nejdůležitější statistiky pro úspěch týmu. To dokládá i korelace mezi druhým faktorem a počtem získaných bodů (tabulka 4.9).

Korelace	*Body*	*F1*	*F2*
Body	1	0,501949	0,813061
F1	0,501949	1	-1,4E-07
F2	0,813061	-1,4E-07	1

Tabulka 4.9: Korelace prvních dvou faktorů týmových statistik a počtu získaných bodů

Třetí faktor, vysvětlující 18,6 % celkové variability, je závislý hlavně na znacích *CrdY*, *Cmp%*, *Tackles*, *TacklesW*, *Press*, *PressW* a *Blocks* (vyjma znaku *Cmp%* jsou všechny faktorové zátěže kladné). Kromě znaku *Cmp%* se jedná o čistě defenzivní statistiky. Více než o obranných schopnostech týmu však třetí faktor hovoří o tom, jak velkou část zápasů se tým musel bránit. Důkazem je znak *GoalsAgainst*, jeden z nejlepších znaků k popisu kvality obrany, který s třetím faktorem koreluje velmi slabě. Tvrzení, že třetí faktor popisuje více dobu, kdy se tým brání, než schopnost obrany, potvrzují záporné faktorové zátěže znaků souvisejících s držením míče jako např. *Cmp* nebo *Touches*.

Čtvrtý a pátý faktor silně korelují s dvěma na sobě závislými znaky, s ostatními znaky je korelace poměrně slabá. Pro interpretaci to znamená, že výpovědní hodnota čtvrtého a pátého faktoru je velmi podobná příslušným závislým znakům. S čtvrtým faktorem mají velké faktorové zátěže znaky *DribblesW* a *Dribbles*. S pátým faktorem mají velké faktorové zátěže znaky *Pkmade* a *Pkatt*.

Quartimax

Na závěr budou výsledky opět srovnány s rotací quartimax.

	F1	F2	F3	F4	F5	Komunalita
Goals	0.95	0.11	0.05	-0.11	-0.15	0.96
Assists	0.96	0.07	-0.02	-0.14	-0.11	0.96
Pkmade	0.11	0.26	0.85	0.07	-0.11	0.82
Pkatt	0.00	0.05	0.96	0.01	0.07	0.94
CrdY	-0.59	0.51	0.21	0.47	0.02	0.87
CrdR	-0.39	0.07	-0.29	0.13	0.63	0.66
xGoals	0.94	0.13	0.10	0.21	-0.05	0.96
xAssists	0.95	0.14	-0.02	0.16	-0.06	0.96
GoalsAgainst	-0.78	-0.20	-0.23	0.27	-0.03	0.78
CleanSheet	0.60	0.28	0.13	-0.44	0.12	0.66
Sh	0.85	-0.05	0.08	0.48	-0.08	0.97
SoT	0.92	-0.03	0.09	0.26	-0.21	0.97
G/Sh	0.75	0.22	-0.08	-0.47	-0.25	0.90
G/SoT	0.72	0.26	-0.09	-0.59	-0.12	0.96
FreeKicks	0.59	-0.21	0.07	0.19	0.29	0.52
Cmp	0.90	-0.26	-0.10	-0.13	0.14	0.93
Att	0.92	-0.20	-0.13	-0.13	0.12	0.93
Cmp%	0.80	-0.44	0.03	-0.11	0.17	0.87
KP	0.86	0.03	0.10	0.46	-0.07	0.96
Tackles	-0.46	0.78	0.24	0.11	0.05	0.89
TacklesW	-0.29	0.77	0.08	-0.07	0.05	0.69
Press	-0.68	0.54	0.20	-0.21	-0.23	0.89
PressW	0.24	0.93	0.01	-0.14	-0.02	0.94
Blocks	-0.53	0.62	-0.07	0.06	0.05	0.68
Touches	0.91	-0.17	-0.14	-0.12	0.14	0.92
DribblesW	0.74	-0.05	0.10	-0.04	0.60	0.92
Dribbles	0.68	-0.03	0.13	-0.04	0.60	0.84

Tabulka 4.10: Faktorové zátěže a komunalita týmových statistik po rotaci quartimax

Na rozdíl od souborů brankářských a hráčských statistik vychází u souboru týmových statistik výsledky získané metodami varimax a quartimax odlišně. Důvodem je velký počet zkoumaných znaků vůči počtu zkoumaných jednotek. V takovém případě se projeví vlastnost rotace quartimax, která se na rozdíl od metody varimax více zaměřuje na minimalizaci počtu faktorů potřebných k vysvětlení variability původních znaků. Model získaný rotací quartimax je pak dosti podobný modelu získanému analýzou hlavních komponent. Odchylky faktorových zátěží jsou pro většinu znaků a faktorů / komponent menší než 0,1. Největší rozdíl lze pozorovat u znaku *CrdR* a čtvrtého faktoru, kde se číselné hodnoty zátěží obou metod liší o 0,81.

5. Závěr

Cílem práce bylo představit metody sloužící k redukci dimenzionality statistických souborů a následně tyto metody na zvolených souborech aplikovat. Konkrétně byla práce zaměřena na popis metody hlavních komponent a na její alternativu - faktorovou analýzu. Jako zkoumaný statistický soubor byly využity datové sady ze světa fotbalu, které jsou díky svému rozsahu vhodným příkladem k demonstraci metod redukce dimenzionality.

Úvodní kapitola sloužila k seznámení s matematickým aparátem, který byl pro úspěšné použití obou metod potřebný. Většina kapitoly byla věnována detailnímu vysvětlení modelů obou používaných metod. V části zabývající se faktorovou analýzou byly navíc zmíněny různé postupy extrakce a rotace faktorů, které ukazují množství variant, kterými lze faktorovou analýzu dále směrovat (v této práci byl používán výhradně postup vycházející z metody hlavních komponent).

Další kapitola se věnovala objasnění souborů, na kterých měly být metody redukce aplikovány. Kromě hlavní sekce zabývající se stručným vysvětlením všech použitých znaků zde byla zařazena i kratší sekce vysvětlující podrobněji některé statistiky, jež jsou z hlediska matematiky zajímavé.

V závěrečné kapitole byly názorně představeny postupy, které jsou při aplikaci metod redukce dimenzionality na konkrétní soubory využívány. Aplikace metody hlavních komponent i aplikace faktorové analýzy zde byly krok po kroku vysvětleny a jejich výsledky byly pomocí grafů a tabulek prezentovány.

Výsledky práce poukazují na odlišnosti obou metod, zároveň však potvrzují, že jejich kombinace je vhodná pro zpracovávání souborů vysoké dimenze. Použitím metody hlavních komponent a následně faktorové analýzy lze získat z často nepřehledného souboru dat řešení, které je snadno interpretovatelné při minimální ztrátě informací.

Tato práce se snažila poskytnout základ pro další analýzu fotbalových statistik. Při volbě vhodného statistického souboru lze vytvořit plně funkční model, který dokáže hodnotit hráče podle jejich schopností, dovedností, rolí nebo stylu hry. Práci však lze brát jako inspiraci nejen pro zkoumání souborů fotbalových statistik, ale v podstatě jakýchkoliv statistických souborů vyšší dimenze.

Literatura

[1] BIGGS, Jeremy a MADNANI, Nitin. *Factor_analyzer* [online]. 2017 [cit. 2021-5-19]. Dostupné z: https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html

[2] BRANSEN, Lotte & VAN HAAREN, Jan. (2018). Measuring Football Players' On--the-ball Contributions From Passes During Games.

[3] CENTELLEGHER, Simone. How to compute PCA loadings and the loading matrix with scikit-learn. In: *scentellegher.github.io* [online]. 2020 [cit. 2021-5-19]. Dostupné z: https://scentellegher.github.io/machine-learning/2020/01/27/pca-loadings-sklearn.html

[4] GRACE-MARTIN, Karen. The Fundamental Difference Between Principal Component Analysis and Factor Analysis [online]. 2017 [cit. 2021-5-19]. Dostupné z: https://www.theanalysisfactor.com/the-fundamental-difference-between--principal-component-analysis-and-factor-analysis

[5] HEBÁK, Petr. *Vícerozměrné statistické metody (3)*. Praha: Informatorium, 2005. ISBN 80-7333-039-3.

[6] JOLLIFFE, Ian. *Principal component analysis*. New York: Springer Verlag, 2002. ISBN 0-387-95442-2

[7] KOMENDA, Martin. *Faktorová analýza* [online]. Brno, 2015 [cit. 2021-5-19]. Dostupné z: https://is.muni.cz/www/98951/41610771/43823411/43823458/Analyza_a_hodnoc/44563155/00_Faktorova_analyza.pdf

[8] KURDYUKOVA, Anna. *Factor Models* Budapest, 2010 [cit. 2021-5-19]. Dostupné z: http://mathematics.ceu.edu/sites/mathematics.ceu.hu/files/attachment/basicpage/29/annakurdyukova-2010.pdf. Master's Thesis. Central European University.

[9] MELOUN, Milan a Jiří MILITKÝ. *Statistická analýza experimentálních dat*. Vyd. 2., upr. a rozš. Praha: Academia, 2004. ISBN 80200-1254-0.

[10] Arnab Auddy (https://math.stackexchange.com/users/451712/arnab-auddy). Why eigenvectors with the highest eigenvalues maximize the variance in PCA? [online]. 2019-05-02 [cit. 2021-5-19]. Dostupné z: https://math.stackexchange.com/q/3211640

[11] Curse of dimensionality. In: *Wikipedia* [online]. 2021 [cit. 2021-5-19]. Dostupné z: https://en.wikipedia.org/w/index.php?title=Curse_of_dimensionality-dir=prevaction=history

[12] Factor Analysis Extraction. In: *ibm.com* [online]. ©1989-2021 [cit. 2021-5-19]. Dostupné z: https://www.ibm.com/docs/en/spss-statistics/SaaS?topic=analysis-factor-extraction

[13] Factor Analysis Rotation. In: *ibm.com* [online]. ©1989-2021 [cit. 2021-5-20]. Dostupné z: https://www.ibm.com/docs/en/spss-statistics/23.0.0?topic=analysis-factor-rotation

[14] *Football Stats and History* [online]. Philadelphia, United States: Sports Reference, 2000 [cit. 2021-5-6]. Dostupné z: https://fbref.com/en/

[15] How Did These Goals Go In? - We Explain How Goal Probability Works. In: *YouTube* [online]. [cit. 2021-5-6]. Dostupné z: https://youtu.be/_vGhocyvKhA?

[16] *Jupyter* [online]. 2014 [cit. 2021-5-19]. Dostupné z: https://jupyter.org/about

[17] *Numpy* [online]. 2008 [cit. 2021-5-19]. Dostupné z: https://numpy.org/doc/stable/user/whatisnumpy.htmlwhatisnumpy

[18] *pandas* [online]. 2008 [cit. 2021-5-19]. Dostupné z: https://pandas.pydata.org/about/index.html

[19] *Plotly* [online]. 2013 [cit. 2021-5-19]. Dostupné z: https://plotly.com/python/plotly-express/

[20] *Matplotlib* [online]. 2002 [cit. 2021-5-19]. Dostupné z: https://matplotlib.org/

[21] *Seaborn* [online]. 2012 [cit. 2021-5-19]. Dostupné z: https://seaborn.pydata.org/

[22] *Scikit-learn* [online]. 2007 [cit. 2021-5-19]. Dostupné z: https://scikit-learn.org/stable/about.html

[23] *StatsBomb* [online]. Bath, Somerset, United Kingdom: StatsBomb, 2016 [cit. 2021-5-6]. Dostupné z: https://statsbomb.com/

[24] xG Explained. In: *fbref.com* [online]. 2000 [cit. 2021-5-6]. Dostupné z: https://fbref.com/en/expected-goals-model-explained/

6. Seznam použitých zkratek a symbolů

Assists: Počet gólových asistencí.

Att: Počet pokusů o přihrávku.

AvgDist: Průměrná vzdálenost od brány při defenzivních zákrocích.

Blocks: Počet zablokovaných střel a přihrávek.

CleanSheet: Počet zápasů bez obrdženého gólu.

Cmp: Počet přihrávek.

Cmp%: Úspěšnost přihrávek.

CrdR: Počet červených karet.

CrdY: Počet žlutých karet.

CS% : Procento zápasů, ve kterých brankář neobdržel žádný gól.

Dribbles: Počet vyhýbání se soupeři snažícímu se zachytit míč.

DribblesW: Počet úspěšného vyhýbání se soupeři snažícímu se zachytit míč.

FreeKicks: Počet střel z přímých kopů.

GAA: Průměrný počet obdržených branek za zápas.

Goals: Počet vstřelených gólů.

GoalsAgainst: Počet obdržených branek.

G/Sh: Úspěšnost střel.

G/SoT: Úspěšnost střel na bránu.

KP: Počet klíčových přihrávek (přihrávek vedoucích ke střele).

LCmp%: Přesnost přihrávek delších než 40 yd.

OPA/90: Průměrný počet defenzivních zákroků vně pokutového území během zápasu.

Pass%: Úspěšnost přihrávek.

Pkatt: Počet pokutových kopů.

Pkmade: Počet vstřelených gólů z pokutových kopů.

Press: Počet napadání soupeře během rozehrávky nebo kontroly míče.

PressW: Úspěšné napadání soupeře během rozehrávky nebo kontroly míče.

Psv%: Úspěšnost při pokutovým kopech.

Save%: Úspěšnost zákroků proti střelám mířícím na bránu. Zblokované střely nebo střely mimo bránu se nepočítají.

Sh: Počet střel.

Shoots: Počet střel.

SoT: Počet střel na bránu.

SoT%: Procento střel, které míří na bránu.

Stp%: Procento křižných přihrávek soupeře do pokutového území, které brankář úspěšně překazil.

Tackles: Počet pokusů o odebrání míče soupeři.

TacklesW: Počet úspěšných odebrání míče soupeři.

Touches: Počet doteků s míčem.

W%: Procento zápasů, ve kterých brankářův tým zvítězil.

xGoals: viz sekce Expected goals.

xAssists: viz sekce Expected assists.