Formative vs Reflective constructs: a CTA-PLS approach on a goalkeepers' performance model

Costrutti Formativi vs Riflessivi: un approccio CTA-PLS su un modello di performance dei portieri

Mattia Cefis and Eugenio Brentari

Abstract Nowadays, PLS-SEM is a trend-topic, whereas football is moving towards a data-driven approach; by combining these two worlds, we aim to show a new way for measuring football goalkeepers' performance, by using data provided from EA Sports experts and available on the Kaggle data science platform. Furthermore, another objective is to refine the model, supporting football experts from a statistical point of view. For this purpose, we adopt a confirmatory tetrad analysis (CTA-PLS) to validate and evaluate the nature (e.g. formative or reflective) of each latent variable. Then, a second-order PLS-SEM model is built. We validate and compare this new indicator with a benchmark (the EA *overall*). The final goal is to prove the CTA approach on a real case study and to refine a composite performance indicator for helping football policy makers taking strategic decisions.

Abstract *Al giorno d'oggi, il PLS-SEM è un argomento di tendenza mentre il calcio si sta muovendo verso un approccio data-driven; combinando questi due mondi, vogliamo mostrare un nuovo modo per misurare le abilità dei portieri, utilizzando i dati definiti dagli esperti EA e disponibili sulla piattaforma Kaggle. Come secondo obiettivo vogliamo supportare gli esperti grazie ad un approccio statistico. Con questo fine, applicheremo un'analisi CTA-PLS per valutare la natura (e.g. formativa o riflessiva) di ogni variabile latente. In seguito abbiamo implementato un modello PLS-SEM di secondo ordine. Abbiamo poi confrontato questo nuovo indicatore con un indice di riferimento (l'EA overall). L'obiettivo ultimo è quello di testare la CTA analisi su un reale caso di studio e offrire un indicatore composito di performance per aiutare gli addetti ai lavori a prendere decisioni strategiche.*

Key words: CTA-PLS, PLS-SEM, Latent variables, Football, Performance.

Mattia Cefis
University of Brescia, Department of Economics and Management, e-mail: mattia.cefis@unibs.it

Eugenio Brentari
University of Brescia, Department of Economics and Management, e-mail: eugenio.brentari@unibs.it

1 Introduction

The latest developments in sports research are moving towards a data-driven approach. In particular, focused on football (i.e. soccer for Americans), players' performance measure is becoming a strategic key for football coaches and policy makers, in order to evaluate players impartially. The majority of papers on performance evaluation are focused just on movement players (i.e. defenders, midfielders and forwards, [5]): by this research we want to focalize attention on a singular role, the goalkeepers. We are inspired by Electronic Arts (EA)[1] experts: in their opinion, goalkeepers' performance can be thought as a multidimensional construct made up of 7 performance composite indicators (i.e. the same 6 used for movement players plus a specific one for goalkeepers), each one made up of several specific skills, which combined form an *overall* index that sums up the performance; then, a statistical support is required [2, 4]. Using data provided by the Kaggle data science platform, our goal is to propose the use of an innovative confirmatory tetrad analysis applied in the PLS context (CTA-PLS) to support experts from a statistical point of view regarding the nature of each construct, as formative or reflective. Following the CTA-PLS output, we will build a second order Partial Least Squares - Structural Equation Model (PLS-SEM) model, in order to build a refined composite indicator dedicated to goalkeepers and comparing it with the well-known EA *overall*.

2 Literature overview and data employed

Existing literature focused on players' performance [2, 4] includes different approaches: for example Carpita et al [3] adopted an unsupervised method to classify different area of performance, Cefis and Carpita [5] already proposed a PLS-SEM model considering only movement roles, but without a CTA approach. The aim of this research is to focalize attention on the evaluation of goalkeepers' performance, exploring key performance indices (KPIs), in order to evaluate some different strategic latent variables (LVs) and their theoretical nature (i.e. formative or reflective).

For this application has been used data from EA experts and available on the Kaggle[2] data science platform; in particular, we will focus on all goalkeepers' stats from the top 5 European Leagues (e.g., Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). This dataset contains 31 variables (e.g. KPIs), with periodic players' performance on a 0-100 scale with respect to different abilities, classified by *sofifa* experts into 6 latent traits: *attacking, skill, movement, power, mentality* and *goalkeeper features*; note that, after a preliminary check, we did not take into account the *defending* block for this model, since its skills are strictly related with movement players. Note that a block is a group of MVs forming a LV: for example the *skill* block is composed by dribbling, curve, fk

[1] www.easports.com

[2] www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

accuracy, long passing and ball control. The classification provided by *sofifa* experts is available online[3]. For our purpose we have chosen to take into account data relying the beginning of the season 2019/2020, so the dataset was composed by stats about 331 goalkeepers.

2.1 The PLS-SEM model and the CTA-PLS approach

PLS-SEM [15], also called PLS-PM, is a tool that offers a valid alternative as compared to the well-known covariance-based model [10]. Its goal is to measure causality relation between concepts (e.g. LVs), starting from some manifest variables (MVs), by an exploratory approach: the explained variance of the endogenous latent variables is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression. Additionally, PLS-SEM does not require any preliminary assumptions for the data, so it's called a soft-modelling technique. In our framework, PLS-SEM estimates simultaneously two models: a measurement (outer) and a structural (inner). In particular, for what concern the measurement model, PLS-SEM allows two types of constructs, respectively reflective and formative: the first one implies that the *q-th* LV exists independently from the measures used (1) (i.e. causality from construct to items, where λ_{pq} is the loading connecting LV q with its MV p, by a simple linear regression, estimated by OLS), whereas the second is determined as a combination of its own indicators (2) (i.e. causality from items to construct, each latent variable ξ_q is considered to be formed by its own MVs following a multiple regression, where the weights are estimated by least squares).

$$x_{pq} = \lambda_{pq}\xi_q + \varepsilon_q \tag{1}$$

$$\xi_q = \sum_{p=1}^{p_q} w_{pq}x_{pq} + \delta_q \tag{2}$$

But there is a lack: while for reflective constructs exist several tests to assess their reliability, for what concern formative constructs researchers are just basing on theory and experts opinion, causing possible measurement misspecifications. As consequence, this can lead a bias in the inner model estimation and lead to incorrect assessments of relationships in PLS-SEM [8]. In order to overstep those limits, some researchers applied the confirmatory tetrad analysis (CTA, [1]) for drawing conclusions about the appropriateness of using formative measurement models as compared to reflective ones [8]. In brief, a tetrad τ is the difference between the product of two pairs of covariances; for instance, the six covariances of a block composed by four MVs permit the formation of three tetrads:

[3] https://sofifa.com/player/192985/kevin-de-bruyne/220030/

$$\begin{aligned}\tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24} \\ \tau_{1342} &= \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32} \\ \tau_{1423} &= \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43}\end{aligned} \tag{3}$$

Note that all tetrads for each block of LV must be tested using a bootstrap procedure (CTA-PLS uses the bias corrected bootstrap by a Bonferroni -nonparametric- approach [8]). If all tetrads confidence intervals (CIs) for that specific LV contain zero (i.e. vanishing tetrads) then the construct can be considered as reflective, otherwise it is formative [8, 1].

Starting from the output of the CTA-PLS, we have built a second order PLS-SEM model, as hierarchical model [12]. In this framework we can include LVs that represent a "higher-order" of abstraction (HOC). In fact, for our purpose, we will assume goalkeepers' macro-composite performance as extra-latent construct of second order, influenced directly from the others 6 lower order constructs (LOCs). Since the HOC is without any apparent MVs, literature suggested us a recent technique in order to modelling this framework: a mixed two-step approach [6]. In the first step we computed the classical repeated-indicators approach, while in the second one we applied the classical PLS-SEM using the computed scores (of LOCs) as MVs for the HOC. For what concern the structural (inner) model, in our framework it links all $R = 6$ LVs (LOCs) with the HOC, by a linear model (4), where the path coefficients (β_{rq}) are estimated by a factorial scheme (i.e. the correlation between the endogenous and the exogenous LV [11]).

$$\xi_q = \sum_{r=1}^{R} \beta_{rq}\xi_r \; + \zeta_q \tag{4}$$

For this project the *smartPLS*[4] software and the R software package *seminr* [13] have been used; we carried out a bootstrap validation (i.e. 5000 resampling) for the model in order to assess the path significance. In the next section, preliminary results are shown.

3 Results and discussion

Preliminary CTA-PLS output suggests us the following classification for the LOCs:

- Reflective constructs (i.e. all vanishing tetrads in each block): *attacking, mentality* and *power*.
- Formative constructs (i.e. at least one tetrad does not vanish in each block): *gk_features, movement* and *skill*.

At this point we run the model following the CTA-PLS advice and then we assessed each LV removing problematic MVs [14]:

[4] www.smartpls.com

- Reflective constructs: we removed some MVs with reliability problems (i.e. loadings < 0.7), in particular crossing, heading accuracy and short passing that refers to the *attacking* LV, aggression, vision and penalties relying *mentality*, and jumping, strength and long shot for *power*.
- Formative constructs: here we removed MVs with collinearity problems (i.e. VIF> 5) or outer weights non-significant; agility relying the *movement* construct, whereas diving, positioning and speed for the *gk_features* block.

The final model is showed in Fig. 1: in the light blue circle there are formative constructs, whereas in the light blue rectangles there are reflective constructs; finally, in the white circle there is the HOC. We can see how *GK_Features* (as we expected) have the strongest impact on the macro-composite indicator (i.e. beta coefficient significant and equal to 0.28 for the inner model). It's interesting to note how for each LV the strongest MV (i.e. with highest weight or loading) is a typical variable strictly related with the goalkeepers ability [9], for example: long passing for *skill*, reaction for *movement*, shot power for *power*, positioning for *mentality*, short passing for *attacking*. Other comforting results derived from the GoF index, that is 0.792 (i.e. the geometric mean between the inner and the outer model performances) and from the SRMR (standardized root mean square residual, the difference between the observed correlations and the model-implied correlation matrix), equals to 0.096 (i.e. under the threshold of 0.10) [14].

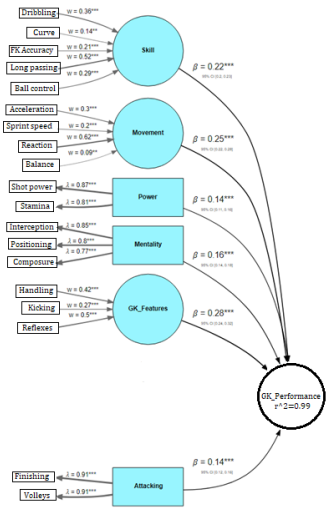

Fig. 1 PLS-SEM GK performance model after 5000 bootstrap resampling.

In order to check the concurrent validity, we compared our scores with some criteria measures (Tab. 1), such as the EA *overall*, wage and players' market value, with interesting results: all medium-high correlations and significant (no one CI 95% contains the zero), the highest between our indicator and the EA *overall*.

Table 1 Correlations of the GK Performance Indicators with three criterion variables.

	GK performance Sept. 2019	CI 95%
EA *overall* Sept. 2019	0.858	$[0.826 - 0.884]$
Wage Sept. 2019	0.605	$[0.532 - 0.669]$
Market Value Sept. 2019	0.585	$[0.509 - 0.652]$

Finally, this model seems to provide comforting results, and at this point for future projects it could be interesting to integrate it in some predictive modelling, such as the expected goal model used in football analytics [7], or to apply CTA-PLS also for movement roles [5]; it should be interesting to compare our model performance respect to a model that considers all constructs as formative or reflective, too.

References

1. Bollen, K.A., Ting, K.f.: A tetrad test for causal indicators. Psychological methods **5**(1), 3 (2000)
2. Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the kaggle european soccer database. Statistical Modelling **19**(1), 74–101 (2019)
3. Carpita, M., Ciavolino, E., Pasca, P.: Players' role-based performance composite indicators of soccer teams: A statistical perspective. Social Indicators Research **156**(2), 815–830 (2021)
4. Carpita, M., Golia, S.: Discovering associations between players' performance indicators and matches' results in the european soccer leagues. Journal of Applied Statistics **48**(9), 1696–1711 (2021)
5. Cefis, M., Carpita, M.: Football analytics: a higher-order pls-sem approach to evaluate players' performance. Book of Short Papers SIS 2021 pp. 508–513 (2021)
6. Crocetta, C., Antonucci, L., Cataldo, R., Galasso, R., Grassia, M.G., Lauro, C.N., Marino, M.: Higher-order pls-pm approach for different types of constructs. Social Indicators Research **154**(2), 725–754 (2021)
7. Green, S.: Assessing the performance of premier leauge goalscorers. OptaPro Blog (2012). URL http://www.optasportspro.com/about/optaproblog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/
8. Gudergan, S.P., Ringle, C.M., Wende, S., Will, A.: Confirmatory tetrad analysis in pls path modeling. Journal of business research **61**(12), 1238–1249 (2008)
9. Hughes, M.D., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., Duschesne, C.: Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position (2012)
10. Jöreskog, K.G.: Structural analysis of covariance and correlation matrices. Psychometrika **43**(4), 443–477 (1978)
11. Lohmöller, J.B.: Predictive vs. structural modeling: Pls vs. ml. In: Latent variable path modeling with partial least squares, pp. 199–226. Springer (1989)
12. Sanchez, G.: Pls path modeling with r. Berkeley: Trowchez Editions **383**, 2013 (2013)
13. Shmueli, G., Ray, S., Estrada, J.M.V., Chatla, S.B.: The elephant in the room: Predictive performance of pls models. Journal of Business Research **69**(10), 4552–4564 (2016)
14. Tabet, S.M., Lambie, G.W., Jahani, S., Rasoolimanesh, S.M.: An analysis of the world health organization disability assessment schedule 2.0 measurement model using partial least squares–structural equation modeling. Assessment **27**(8), 1731–1747 (2020)
15. Wold, H.: Encyclopedia of statistical sciences. Partial least squares. Wiley, New York pp. 581–591 (1985)