Using Higher-Order PLS-PM model for performance analysis in football

Mattia Cefis[1]

[1] University of Brescia (Italy). Email : mattia.cefis@unibs.it

Abstract

Nowadays, data science is applied in several area of our life, and also many applications in sports fields are growing. In this context, we are focusing on football (e.g. soccer); thanks to this work we want to evaluate and monitoring football players' performance by data provides from Electronic Arts (EA) experts and available on the famous Kaggle data science platform. For this purpose, we adopt a Higher-Order Partial Least Squares Path-Modelling (PLS-PM) approach to the sofifa Key Performance Indicators (KPIs) in order to compute a composite indicator and compare it with the well-known *Overall* index from EA Sports. Furthermore, in our project we take into account players' observed heterogeneity (i.e. role), since we often listen mass media and experts speak about differences for these features, and so we aim to verify it in a scientific way. The final goal is to underline need of a new performance index specific for each players' role, for helping policy makers of professional teams to take strategic decisions, in order to evaluate impartially players' performance.

Keywords:
Performance Index; Composite Indicators; Soccer

1. Introduction

In these last years, football, the most practised sport in the world, is moving towards a sort of data-driven revolution; analyse players' performance is becoming a strategic key for coaches and managers, in order to improve team results. We know that players' performance on the pitch has been extensively measured and described by soccer experts: in literature, very important are the detailed classification by the experts from Electronic Arts (EA), in fact they have thought performance defined by 6 composite strategic indicators, each one with specific KPIs which combined form the well-known EA/*sofifa Overall* index. But at this point the main problem is that experts' opinion is not statistically supported (Carpita, M.& Golia, S., 2020; Carpita et al., 2019) and moreover it is not very clear how they keep in consideration players' heterogeneity (i.e. the roles on the pitch). Cefis, M.& Carpita, M.(2020) showed some relevant differences in performance KPIs among players' roles. The aim of this work is firstly to replicate the *Overall* index by an innovative Higher-Order PLS-PM model and validate it, considering all players at the same way (e.g. independently from the role), then to introduce possible future improvements taking in consideration heterogeneity among players.

For this application we will use data provides from EA experts and available on the famous Kaggle data science platform from Leone (https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset, named Fifa 20 complete player dataset); in particular, we will focus on all players' stats from the top 5 European Leagues (e.g. Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). This players attributes table contains other 28 variables (e.g. KPIs), with periodic player's performance on a 0–100 scale with respect to different abilities; for our purpose we have chosen to take into account data relying the beginning of the season 2018/2019, so our dataset was composed from stats about 2662 players (note that we do not consider goalkeepers but just movement players).

As said before, for what concerns attributes' description EA/*sofifa* experts are the main recognised authority: players' performance is defined as a multidimensional entity made up of 6 latent traits (e.g. *attacking, skill, movement, power, mentality, defending*), and thanks to this work we want to support it also using statistical evidences, in order to give more solid information to football coaches and managers.

2. Methodology

For our purpose we adopted a PLS-PM (Wold, H., 1985) approach, that offer a valid alternative to the well-known covariance-based model (Joreskog, K.G., 1978). Its goal is to measure causality relation between concepts (Latent Variables or LVs, the 6 *sofifa* latent traits in our case), starting from some Manifest Variables (e.g. MVs, in our case the *sofifa* KPIs), thanks to an explorative approach: the explained variance of the endogenous LVs (outcome variables as the performance in our case) is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression (Monecke, A., Leisch, F., 2012). Another important point to underline is that PLS-PM does not require any preliminary assumptions for the data (i.e. soft-modelling technique). It estimates simultaneously two model:

- The Measurement (outer) model, that links MVs (KPIs) to their corresponding LVs (e.g. the 6 *sofifa* dimensions). Each block of MVs $\boldsymbol{X}_g, g = 1, \ldots, G = 6$ must contain at least one MV and for our case we treated this relation in a formative way (MVs are the causes of their own LV). In particular, we assumed each LV ξ_g as formed by its KPIs following a multiple regression (1), where $\boldsymbol{w}_g$ is the vector of the outer regression weights and δ_g of error terms. Finally, the vector of the outer weights for the *g-th* LV is estimated by OLS.

$$\xi_g = \boldsymbol{X}_g \, \boldsymbol{w}_g + \delta_g \tag{1}$$

- The Structural (inner) model, that divides the LVs into two groups: exogenous and endogenous. The first one does not have any predecessor in the path diagram, the rest are endogenous. For the *j-th* endogenous LV in the model, the linear equation of its own structural model is defined in (2); in particular, *R* represents the number of exogenous LVs that affect the endogenous one and β_{rj} is so called path coefficient, a sort of linkage between the *r-th* exogenous LV and the *j-th* endogenous one, where ζ_j is the error term.

$$\xi_j = \beta_0 + \sum_{r=1}^{R} \beta_{rj} \, \xi_r + \zeta_j \tag{2}$$

In this study the High-Order or Hierarchical model is adopted (Sanchèz, 2013), so that LVs that represent superior levels of abstraction can be included. In particular, a third-order model is used (Fig. 1). In fact, for the purpose it has been assumed, after consulting with some soccer-experts, players' performance was viewed as extra-latent construct of higher (third) order, formed from two extra LVs (second order constructs), as *Off_phase* (the phase of ball possession) and *Def_phase* (the phase without ball possession). It has been assumed that the initial 6 *sofifa* LVs (first order constructs) contribute to the second-order LVs in the following way: all first order LVs except *defending* shape the *Off_phase*, while all LVs except *attacking* contribute to *Def_phase*. Since the second and the third order constructs are without any MVs, literature suggested an interesting technique in order to modelling this framework: a two-step or patch approach (Sanchèz, 2013). In the first step of this approach, the Principal

Component Analysis (PCA) is used to obtain the scores of the lower-order LVs (the first principal component - I PC - of each one), and in the second step the standard PLS-PM use these PCs as MVs for the endogenous LVs. In particular, as MVs of *Off_phase* it has been adopted the I PCs of *attacking, skill, movement power and mentality* (named from off1 to off5 in Fig. 2), while for *Def_phase* the I PCs of *movement, power, mentality, defending and skill* (named from def1 to def5 in Fig. 2). Finally, I PCs of *Off_phase* and *Def_phase* are used as MVs for *Performance*. For this work it has been used the R software packages *csem* (Mehmetoglu, M., Venturini, S., 2020) and *seminr* (Shmueli et al., 2016) for model plots; we provided a bootstrap validation for the full model in order to see path significance. In the next section we will share our results and a brief discussion.

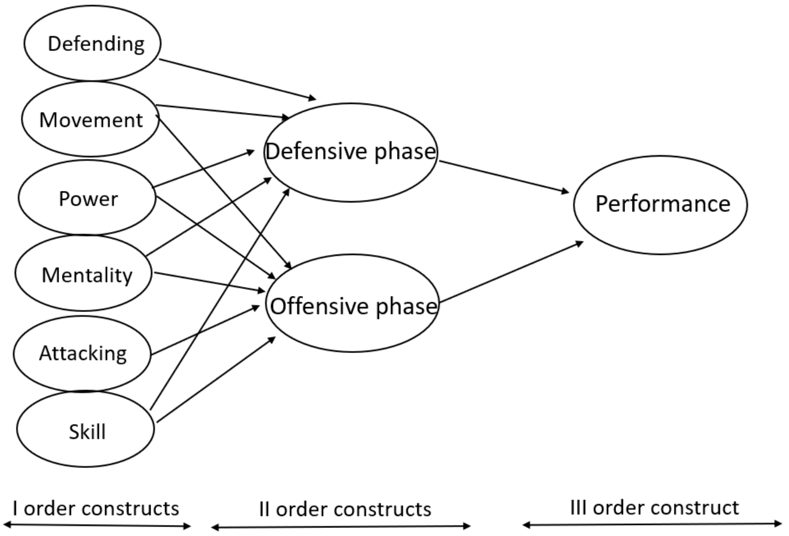

Fig. 1: The third-order inner model for evaluating players' performance

3. Result

In Fig. 2 parameter estimates and their statistical significance of the full model are showed, with circle represent LVs and rectangles represent MVs. Directions of the arrows for the outer model from MVs to LVs represent the formative framework adopted, while the thickness is proportional to the strength of their effects. Above each arrow of the outer model are located loadings (i.e. λ) between each MV and the corresponding LV. For the inner model, estimated beta coefficients above each arrows among LVs are showed. Asterisks next to each estimate represents its statistical significance (after 1000 bootstrap resampling); dotted arrows mean negative values for the corresponding parameters.

We can see immediately how in Fig. 2 *Off_phase* has a stronger impact on the performance than *Def_phase* (0.57 vs 0.43, both significative); it is interesting noting also how all LVs are significative ($\alpha = 1\%$, ***, see Fig. 2) except *skill*, that is less significative ($\alpha = 10\%$, just one *, see Fig. 2) for the *Def_phase*, considering the inner model. Furthermore, some MVs have negative loadings, making difficult interpretation for some LVs (e.g. *defending*). So, we decided to modify the inner model and create a nested model, removing the low-significative path-coefficient named above (*skill* for *Def_phase*) and compare it with the first one using the information criteria AIC and BIC (Sharma et al., 2019).

Tab. 1: comparison between inner models (full vs nested) using information criteria

Model	AIC	BIC
Full: *Off_phase*	-12670.6	-12629.4
Nested: *Off_phase*	-12676.5	-12691.2
Full: *Def_phase*	-10320.5	-10279.3
Nested: *Def_phase*	-9824.8	-9795.4

From Tab. 1 it is evident how the nested model is preferred for the *Def_phase* framework (both AIC and BIC lower in absolute value for the nested model) while for the *Off_phase* are quite similar. Furthermore, Tab. 2 shows the linear correlation coefficient between performance scores of the PLS-PM models with the EA *Overall* Performance Index.

Tab. 2: Correlation between performance indicators of the PLS-PM models with the EA *Overall*

PLS-PM Model	Corr.
Full	0.64
Nested	0.65

As criterion validity, performance indicator of the nested (restricted) model has a bit higher correlation with the EA *Overall*. Then we have replicated bootstrap validation for the nested model, and now all inner model path are significative, but some different outer model loadings are near to zero or negative yet.

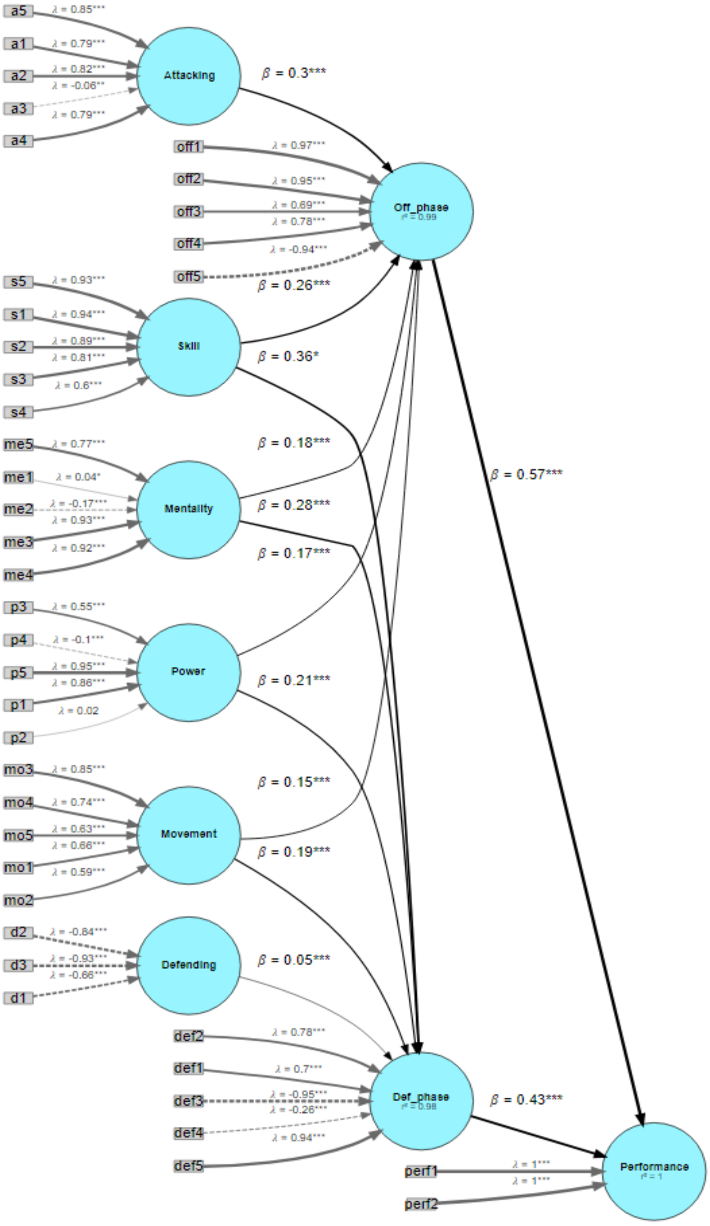

Fig. 2: The output of the full model considering all 2662 players

4. Discussion and Conclusion

As summary, we tried to replicate and build an innovative players' performance model and two others interesting indicators as *Off_phase* and *Def_phase*, thanks to a third-order PLS-PM approach. We built two models, a full (i.e. 5 *sofifa* LVs connected to *Def_phase* and *Off_phase* both) and a nested one (i.e. without no significative path): as result, the nested is a bit better than the full one, but it shows instability in outer estimates too; we think it is due by considering all players in our analysis. As advice, for future improvements, it should be interesting to consider observed heterogeneity (e.g. players' roles on the pitch) and replicate the nested model for each role, in order to stabilize outer estimates. For roles' classification on the pitch, it should be useful to consider specific roles (Huges, M. et al, 2012) and not classical three ones (i.e. defenders, midfielders and forwards). In this sense, Fig. 3 shows a possible starting point for future research, comparing EA *Overall* performance versus the performance indicator of the PLS-PM nested model (i.e. standardized values), suggesting a pattern by role. The final goal remains to customize specific performance index, also considering the *Off_phase* and *Def_phase* models for helping scouting and staff of a soccer team.

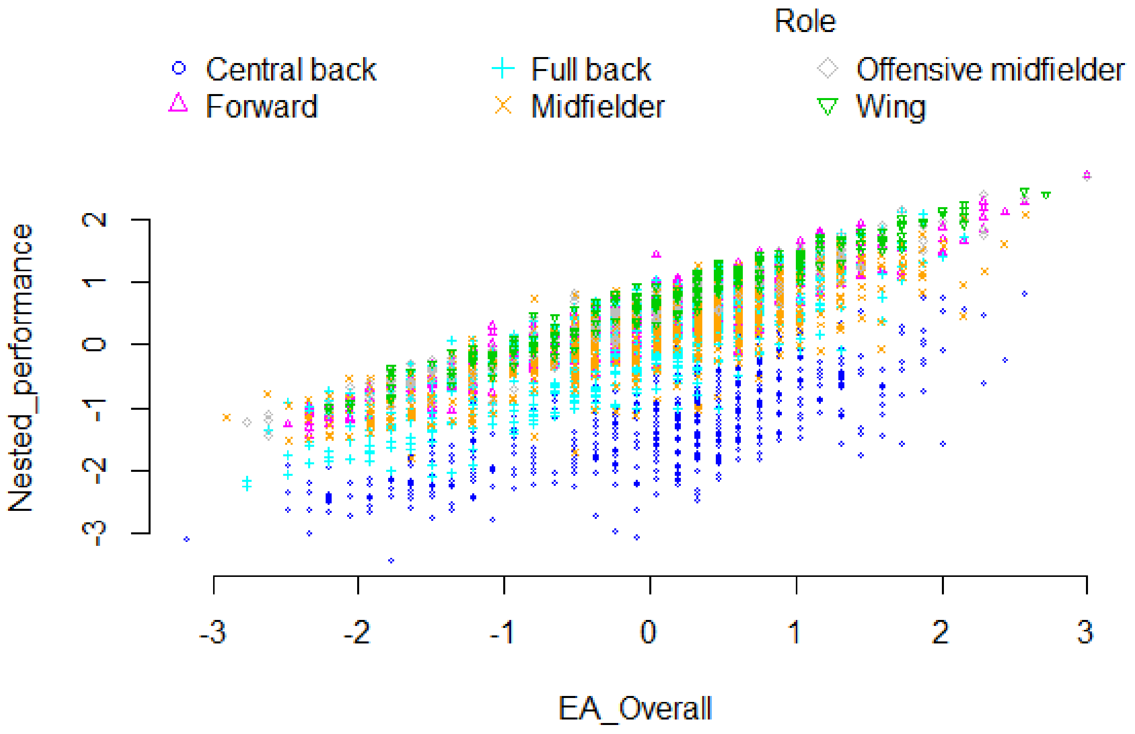

Fig. 3: EA Overall vs PLS-PM performance of the nested model by roles

References

1. Carpita, M. & Golia, S. (2020). Discovering associations between players' performance indicators and matches' results in the European Soccer Leagues. Journal of Applied Statistics: 1-16.
2. Carpita, M., Ciavolino, E. & Pasca, P. (2019). Exploring and modelling team performances of the Kaggle European Soccer database. Statistical Modelling 19.1: 74-101.
3. Cefis, M. & Carpita, M. (2020). Football Analytics: Performance analysis differentiate by role. In book of abstracts: p. 22.
4. Hughes, M.D., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A. & Duschesne, C. (2012). Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position.
5. Joreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. Psychometrika 43.4: 443-477.
6. Mehmetoglu, M. & Venturini, S. (2020). Structural equation modelling with partial least squares using Stata and R. CRC Press.
7. Monecke, A. & Leisch, F. (2012). semPLS: structural equation modeling using partial least squares. Journal of Statistical Software, 48 (3), 1-32.
8. Sharma, P.N., Shmueli, G., Sarstedt, M., Danks, N. & Ray, S. (2019). Prediction-oriented model selection in partial least squares path modelling. Decision Sciences.
9. Sanchez, G. (2013). PLS path modeling with R. Berkeley: Trowchez Editions 383.
10. Shmueli, G., Ray, S., Estrada, J.M.V. & Chatla, S.B. (2016): The elephant in the room: Predictive performance of PLS models. In: Journal of Business Research 69.10, pp. 4552–4564.
11. Wold, H. (1985). Encyclopedia of statistical sciences. Partial least squares.Wiley, New York: 581-591.