Proceedings of Machine Learning Research vol 236:769–792, 2024 3rd Conference on Causal Learning and Reasoning

A causality-inspired adjusted plus-minus model for player evaluation in team sports

Caterina De Bacco CATERINA.DEBACCO@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Cyber Valley, Tuebingen, 72076, Germany

Yixin Wang
Department of Statistics, University of Michigan, Ann Arbor, MI 48109

David M. Blei
Department of Computer Science, Columbia University, New York, USA

Editors: Francesco Locatello and Vanessa Didelez

Abstract

We present a causality-inspired adjusted plus-minus model for evaluating individual players from their performance on a team. We take an explicitly causal approach to this problem, defining the value of a player to be the expected change in the score had we substituted the player for one who has zero value. (This quantity is "causal" in the sense that it is an inference about a hypothetical intervention.) We adapt recent ideas of factor modeling to handle the indirectly measured confounding in estimating player values, considering each player to be a "treatment" who contributes to the outcome of the game. We demonstrate the behavior of the model on data about soccer and basketball.

Keywords: sport analytics, causal inference, adjusted plus-minus, ridge regression, rating systems

1. Introduction

A key task in sports analytics is to evaluate the individual performance of a player on a team. Such evaluations affect scouting, trades, and predictions about a match. Intuitively, historical data about matches—who was playing and what was the outcome—should provide statistical evidence for the value of each player. But players are on the field together, and it is challenging to untangle the relative contributions of each.

The traditional approach to solving this problem is with the adjusted plus-minus model (APM). While there are many variants of the method, APM is basically a regression: the covariates are who is playing in the match, the outcome is the difference in score, and players on one team make a positive contribution while players on the other team make a negative one. Since it's a regression, APM is essentially a model for the conditional expectation of the score given who is on the field. With a fitted APM, the per-player coefficients represent the values of each.

This paper builds on APM. However, rather than model a conditional expectation, we take an explicitly causal approach to evaluating individual players. Specifically, we define the value of a player ℓ to be the expected change in the score had we substituted player ℓ for a player who has zero value. (This quantity is "causal" in the sense that it is an inference about a hypothetical intervention.) We will develop a statistical method to estimate these quantities from historical data.

The challenge to this task, as for all causal inference from non-experimental data, is the possibility of unmeasured confounding variables, variables that affect both who is on the field and

the outcome of the match. When we do not account for those variables, the conditional expectation (such as from APM) is biased away from the targeted interventional expectation. In historical sports data, for example, a coach may select the lineup based on knowledge of the opposing team's strengths and weaknesses. While the coach's decision might lead to a better score, the fitted APM will overestimate the value of the players on the field (relative to their causal value).

This paper develops a method to estimate the "causal" value of a player. The key observation is that, while many confounding factors are unmeasured, some are indirectly measured in the data. For example, traces of the (unobserved) coaching style may be present in the (observed) lineups set up by each coach. To this end, we adapt recent ideas of latent factor modelling to uncover indirectly measured confounding (a.k.a. pervasive confounding) (Abadie et al., 2015; Agrawal et al., 2021; Wang and Blei, 2019a; Frot et al., 2019). The idea is to treat the estimation problem as a causal inference problem, one where each player on the team is a "treatment" and the score of the game is the outcome. There are many players, hence many treatments. The method developed here will use an APM-like regression but include additional variables that are also derived from the historical data. We refer to it as CAPM, causality-inspired adjusted plus-minus. In more detail, CAPM first fits a latent variable model to the player lineups of each match. This model captures which players are likely to appear together on the field. CAPM then uses this model—in particular its latent variables—to calculate a substitute for (some of) the indirectly measured confounders. Finally, CAPM uses the substitute confounder to correct for bias in an outcome model, an APM of the game score that includes the latent variables. The per-player coefficients in this outcome model reveal the value of each player.

We will derive our model and efficient algorithms for each of its components. We will develop a Poisson-factorization model of lineups and use an APM model for the outcome. We will then study this method on synthetic and real data of four European professional soccer leagues from 2014 to 2016 and five NBA basketball seasons from 2014 to 2018. Compared to APM, our approach performs well in predicting score differences using the lineups of the teams in play. Further, it leads to player rankings that are more consistent with player ratings produced by human experts.

We note that there has been a fair amount of academic discussion about the ideas of using latent variable modeling to adjusting for unobserved confounding as proposed in Wang and Blei (2019a). One thread centers around the theory of the method, and in particular around the scope and relevance of the required assumptions; see D'Amour (2019); Ogburn et al. (2019); Wang and Blei (2019b); Ogburn et al. (2020), and the clarified theory in Wang and Blei (2020). This discussion reveals that the assumptions for a truly unbiased estimate of the causal value of a player are strict, and they are unlikely to hold completely. Despite this fact, in our studies we found that CAPM successfully removes some of the bias of an unadjusted estimate; see Section 4. More broadly, we discuss the limitations of CAPM in Section 3.4. Another relevant thread of discussion is Grimmer et al. (2020). It shows how classical regression (here, APM) can also be seen as targeting the causal estimate when the assumptions of CAPM hold. It further finds situations where simple regression outperforms causal adjusted algorithms, particularly where the number of datapoints (here, matches) is much larger than the number of treatments (here, players in the league). In the context of Grimmer et al. (2020), the studies of Section 4 show a different empirical result, multi-causal situations where a causality-inspired algorithm outperforms regression-based estimates.

The main contributions of this work are to quantify the impact of each player as a causal inference, and to develop an algorithm for estimating it. This is an original approach for addressing the challenging and relevant problem of how to evaluate individual players in team sports.

We evaluate the approach on basketball and soccer, performing studies with both real and synthetic datasets. In synthetic studies, we show that the algorithm accurately recovers the (synthetic) ground-truth performance of individual players. In real data analysis, we show that the algorithm more accurately correlates with external assessments of player quality, such as FIFA rankings (for soccer), and better predicts the results of held-out games.

2. Related work

Evaluating player performance in team sports is an active research area. There are various ways to evaluate players, see Santos-Fernandez et al. (2019) for a review of Bayesian methods or Terner and Franks (2021) for a review of methods in basketball. The main distinction between the methods is in which data are used to make the estimates. One type of approach focuses on individual-player data, such as event-data (e.g., number of passes or shots) (Thomas et al., 2009) or individual physical or mental characteristics (Carvalho et al., 2017; Giles et al., 2018). Another type of approach uses tracking data to consider team-level plays, such as possessions or defensive actions (Cervone et al., 2016; Le et al., 2017; Bu et al., 2019; Franks et al., 2015; Fernández et al., 2021).

This paper builds on a third type of approach, which involves regressing the lineup of the team to the outcome of the game. Compared to models of event data, an advantage of regression models is that they account for the impact of all players simultaneously. In contrast, event-data models are typically focused on a given player at a time, regardless of who else is in the field. In particular, the adjusted plus-minus (APM) algorithm is one of the main models used by professional teams and mainstream media. APM originates from seminal work in basketball (Leonhardt, 2003; Rosenbaum, 2004) and hockey (Sill, 2010; Macdonald, 2011a,b). Researchers have proposed several variants of APM (Ilardi and Barzilai, 2008; Deshpande and Jensen, 2016; Thomas et al., 2013; Gramacy et al., 2013; Nandakumar and Jensen, 2019; Hvattum, 2019), but the largest improvement in performance is due to including a ridge regression penalty (Sill, 2010). Recently, basketball practitioners have made efforts to further improve APM, by integrating information derived from tracking data Snarr (2020); Silver (2019); Medvedovsky or by using structured priors and adjustments for luck BBall Index Team.

The field of soccer analytics has only recently used methods like APM. These various models differ from each other in terms of what variables they consider in the regression (Pantuso and Hvattum, 2021; Sæbø and Hvattum, 2019); the types of priors (Matano et al., 2023); the choice of outcomes (Kharrat et al., 2020); the presence of additional information e.g. human-based ranking (Pelechrinis and Winston, 2021); how segments are wheighted (Schultze and Wellbrock, 2018). This paper further builds on APM, embedding it in a larger methodology for estimating the causal value of a player.

The field of soccer analytics has only recently used methods like APM. These various models differ from each other by what variables they consider in the regression. Some include team or league indicators in addition to players (Pantuso and Hvattum, 2021; Sæbø and Hvattum, 2019); some use different priors for the regression parameters, e.g. Laplace, Gaussian, or domain-specific priors (Matano et al., 2023); some consider different outcomes (Kharrat et al., 2020), e.g. expected goals, win/loss or score differential; some use players' position and human-based ratings (Pelechrinis and Winston, 2021) and finally some weigh the segments differently, e.g. giving less weight to "garbage time" (Schultze and Wellbrock, 2018). This paper further builds on APM, embedding it in a larger methodology for estimating the causal value of a player.

There are a few works explicitly focusing on causal analysis in sports analytics, though none analyze lineup data as done in this paper. Gauriot and Page (2019) isolate quasi-experimental situations in soccer where the success of players is as good as random (shots hitting the post vs entering the goal taken from similar locations) to estimate the causal effect of a lucky goal on the players' evaluation. Yam and Lopez (2019) use causal reasoning to assess the unobserved benefit that teams miss in not utilizing a fourth down strategy in American football. Sandholtz and Bornn (2020) use latent Markov decision processes to estimate what could have happened if a basketball player's shot policy changes. Terner and Franks (2021) discuss the open problem of causal analysis for sports.

The key challenge to causal inference from observational data is the problem of latent confounding, unobserved variables that affect both the treatment and outcome. Here we employ the idea of factor modeling (Wang and Blei, 2019a; Abadie et al., 2015; Agrawal et al., 2021) for latent but indirectly measured confounders. Other popular approaches to latent confounding includes proxy variables (Miao et al., 2018; Kuroki and Pearl, 2014), instrumental variables (Angrist et al., 1996), and front door adjustment (Pearl, 1995). These approaches often require measuring additional covariates that satisfy specific criteria, e.g. proxy variables that are descendants of the latent confounder but do not causally affect the outcome. We focus on a setting where such covariates are not available, and these approaches do not apply.

3. Causal inference for team sports

When we evaluate a player, we aim to infer their individual contribution from the collective performance of the team. This paper manifests this individual contribution with a causal question: "How would the results of a match be different had this player been substituted for another?" Our goal is to answer this question by analyzing observational datasets of sports matches.

3.1. Player evaluation as a causal inference

We observe data from sports matches. In this paper, we will analyze data about soccer and basketball. Each dataset contains m players who comprise t teams; player ℓ is on team s_ℓ. We observe n matches (in soccer) or segments of games (in basketball); to keep the language simple, we call each a game. Each game involves two teams, indexed by $u_i \in \{1, \ldots, t\}$ and $v_i \in \{1, \ldots, t\}$, where $u_i < v_i$. For each game i, we observe $a_{i,\ell} \in \{0, 1\}$, which indicates whether player ℓ played in game i. In the soccer data, $a_{i,\ell}$ indicates whether ℓ started the match on the field; in the basketball data $a_{i,\ell}$ indicates whether player ℓ was on the court during segment i. (A segment is a portion of the game without substitutions.) Finally we observe y_i, the differential outcome when u_i plays v_i. In soccer it is the difference in goals; in basketball, it is the difference in points. A positive value of y_i favors team u_i; a negative value favors team v_i.

Our goal is to estimate the value of a player, which we treat as a causal inference. The science of causality, in general, is about trying to answer questions about a hypothetical intervention on the world (Pearl, 2009; Peters et al., 2017; Hernán and Robins, 2016; Imbens and Rubin, 2015). In this spirit, we define the value of a player in the following way. Consider the player in question ℓ, their team t_ℓ, and a special "zero-value" player $\emptyset$. We define the value of player ℓ to be the expected change in the differential score between two scenarios, one where we "force" player ℓ to play and player $\emptyset$ not to play, and one where we force player $\emptyset$ to play and player ℓ not to play. (In both, the

team of player ℓ is the first team in the differential.) We denote this quantity Λ_ℓ:

$$\Lambda_\ell := \mathbb{E}\left[Y\,;\,\mathrm{do}(a_\ell = 1, a_\emptyset = 0, u = t_\ell)\right] - \mathbb{E}\left[Y\,;\,\mathrm{do}(a_\ell = 0, a_\emptyset = 1, u = t_\ell)\right] \quad , \tag{1}$$

where u is the random variable denoting the first team, i.e. the team of player ℓ. This quantity is the difference in score-differential between the two interventions. The expectation is with respect to the distribution of games, including the opposing team and who is playing (other than player ℓ and player $\emptyset$).

As we discussed above, one of the most widely used methods for evaluating players is Adjusted Plus-Minus (APM) with ridge regression (Sill, 2010). APM assumes that each player ℓ contributes β_ℓ to the expected score for their team, and then models the score differential with a regression. Specifically, it uses lineup data—lineups and match outcomes—to estimate the players' values from a penalized regression,

$$\beta^* = \arg\min_\beta \left\{ \sum_i \left[y_i - \left(\sum_{\ell : t_\ell = u_i} \beta_\ell a_{i,\ell} - \sum_{\ell : t_\ell = v_i} \beta_\ell a_{i,\ell} \right) \right]^2 + \lambda \|\beta\|_2^2 \right\}. \tag{2}$$

When we use APM, we treat the optimizer β^*_ℓ as an estimate of the value of each player relative to a "replacement level" player, i.e., a player with $\beta = 0$ (Thomas and Ventura, 2015).

The APM value of a player differs from what we defined in Eq. 1, however, because the APM value is an observational quantity (a conditional expectation) rather than a causal one. As for all causal inference from observational data, the problem to estimating Eq. 1 is that there may be unmeasured confounding—variables that affect both who is playing in the game and the ultimate difference in score. For example, suppose a coach tends to place a particular player in easy-to-win matches to give other (better) players a rest. APM would estimate a high β^*_ℓ for this player, but the causal value Λ_ℓ in Eq. 1 may not be as high. Our goal is to develop an algorithm for estimating the value of each player that helps reduce some of the bias due to this confounding.

3.2. Using adjustment to evaluate the value of a player

Before we tackle the problem of unobserved confounders, however, we will discuss the main strategy for causal inference that we will use, backdoor adjustment (Pearl, 1993) and g-estimation (Robins, 1986). Our method for handling unobserved confounders will build on this strategy.

For now, suppose we can observe a per-match variable w that is admissible for adjustment (Pearl et al., 2009), i.e., the confounders. The game data comes from the true distribution of who is playing, the confounders, the line-ups, and the score, $\mathrm{p}(u, v, w, \mathbf{a}, y)$.

If we observe w, then we can use this data, along with backdoor adjustment, to estimate the value of a player in Eq. 1. First, define notation for the conditional expectation of the score given that team t_ℓ is playing, whether player ℓ is on the field, whether player $\emptyset$ is on the field, and the confounders w,

$$\hat{\mu}_\ell(w, a_\ell, a_\emptyset) = \mathbb{E}\left[Y \mid W = w, A_\ell = a_\ell, A_\emptyset = a_\emptyset, U = t_\ell\right] \quad , \tag{3}$$

where here we used capitalized and lowercase letters (e.g. A_ℓ and a_ℓ to distinguish explicitly the random variable from its possible value, respectively. This expectation is over the distribution of the opposing team and who else is on the field (other than ℓ and $\emptyset$). Note that while the expectations in

Eq. 1 are interventional, this expectation is conditional; it can be estimated from the observational data (where confounders w are observed).

Using these conditional expectations, we can write the value of a player as an integral over the adjustment variables w,

$$\Lambda_\ell = \int \mathrm{p}(w)\,(\mu(w,1,0) - \mu(w,0,1))\,\mathrm{d}w. \tag{4}$$

This equation is an identification formula for the value of a player, an expression of a causality quantity in terms of an observational distribution. It uses backdoor adjustment (Pearl, 1993).

The identification formula is a gateway to causal estimation. Consider a (hypothetical) dataset that contains this confounder. With this data, estimate $\mu(w, a_\ell, a_\emptyset)$, where $(a_\ell, a_\emptyset) = (1,0)$ or $(a_\ell, a_\emptyset) = (0,1)$. Then use this estimate to approximate Eq. 4, for example with Monte Carlo.

Setting aside the issue that we do not observe w_i, what do we assume when we write the causal inference of Eq. 1 with Eq. 4, and then estimate it? The main assumption is that W blocks all backdoor paths (Pearl et al., 2009) between Y and $A_\ell, A_\emptyset, U$. (Loosely, W includes all confounders.) Further, there is a positivity assumption, which requires $P((A_\ell, A_\emptyset, U) \in \mathcal{S} \mid W) > 0$ for all sets $\mathcal{S}$ such that $P(\mathcal{S}) > 0$. It says the value of the confounder W shall not limit the support of $(A_\ell, A_\emptyset, U)$. Loosely, any of the marginally possible values of $(A_\ell, A_\emptyset, U)$ shall also be possible given W.

All this said, the premise of this paper is that we do not observe the confounders W; they are often only indirectly measured. We now discuss CAPM, a method for estimating the causal value of a player in the face of (some) unmeasured confounders.

3.3. The causality-inspired adjusted plus-minus model

The key idea behind CAPM is that the inference of Eq. 1 is a *causal inference*. Each player on the field is a "treatment" who contributes to the final score, and we are interested in making causal inferences about each one. Here we appeal to the ideas of factor modeling for causal inference (Wang and Blei, 2019a; Abadie et al., 2015; Agrawal et al., 2021; Frot et al., 2019), which develops a style of algorithm, that uses data about the multiple treatments to account for some of the unobserved but indirectly measured confounding. We will develop this methodology for estimating the value of each player. Such algorithm works as in two stages. First, we construct a variable that renders the line-up conditionally independent; for example, Wang and Blei (2019a) use probabilistic factor models (Bishop, 2006) to construct such a variable, a model of which players are playing in a game. Second, we use the constructed variable from the first stage as a "substitute confounder," one that we can adjust for in estimating the value of each player in Eq. 1. This stage appeals to the estimation methods described above, but substitutes the true confounder w_i with the substitute confounder. The intuition behind these algorithms is that the factor model captures information about confounders that are not explicitly observed, but for which there is evidence in the patterns of who is playing. Returning to the example of the coach's strategy, hints of the strategy might be revealed in patterns of how different players co-occur on the field. An offensive strategy might prefer one set of players; a defense strategy might prefer another.

Loosely, the theory of this approach says that if the constructed variable can render the line-up conditionally independent, then it has the potential to capture some indirectly measured "multi-cause" confounders, i.e., the unmeasured confounders that affect many players. But this theory relies on some idealized assumptions. In practice, the hope is that a good probabilistic factor model

will capture *some* of the unmeasured confounding, and that using its constructed variable in a downstream inference will reduce the bias from uncorrected estimates.

Before describing the algorithm, however, we state a key causal assumption that is required: we assume no single-player confounding.[1] In other words, the algorithm can never control for unobserved confounding factors that causally affect one and only of the players. As an example, suppose that one player is (temporarily) injured. This injury will affect both whether the player is on the field and the score of the match; it is a single-player confounder. The algorithm will not be able to control for such confounders.

3.3.1. A PROBABILISTIC FACTOR MODEL OF WHO IS PLAYING IN A GAME

The first stage of CAPM requires a factor model of who is playing in the game. The *line-up factor model* is based on Poisson factorization (Gopalan et al., 2015), a Bayesian variation of non-negative matrix factorization (Lee and Seung, 1999; Cemgil, 2009). The model has K components. Each player ℓ is associated with a K-vector of non-negative latent features θ_ℓ; each team u is associated with a K-vector of non-negative latent features q_u. Whether a player appears in the game depends on the affinity between the player's features and the opposing team's features. More specifically, the line-up factor model is described by the following generative process:

- Generate latent features for each player ℓ and team u,

$$q_u \sim \text{gamma}_K(-) \qquad u = 1, \ldots, t \tag{5}$$
$$\theta_\ell \sim \text{gamma}_K(-) \qquad \ell = 1, \ldots, m. \tag{6}$$

- For game i, where u_i plays v_i, draw the line-up as

$$a_{i,\ell} \sim \text{Poisson}(\theta_\ell \cdot q_v) \qquad \text{if } s_\ell = u_i \tag{7}$$
$$a_{i,\ell} \sim \text{Poisson}(\theta_\ell \cdot q_u) \qquad \text{if } s_\ell = v_i \tag{8}$$
$$a_{i,\ell} = 0 \qquad \text{if } s_\ell \neq u_i \text{ and } s_\ell \neq v_i \tag{9}$$

Note this model does not involve the outcome of the game, i.e., the difference in score. Rather, it is a model of the line-ups alone. The latent variables θ_ℓ and q_u capture regularities in how players are assigned to the field the game is played.

As described above, we use the line-up factor model to estimate a "substitute confounder." First we condition on the line-ups from each game and estimate the posterior distribution of the latent variables $\text{p}(\theta_{1:m}, q_{1:t} \mid \mathbf{a}_{1:n})$. Here we approximate the posterior with variational inference (Blei et al., 2017; Gopalan et al., 2015). Variational inference for this model scales to the datasets we will study; see Appendix A for algorithmic details. Following the framework of a causal adjusted algorithm, we will use the approximate posterior to construct a substitute confounder for each game. Specifically, we concatenate the posterior expectations of the latent features of which teams are playing, $\hat{z}_i = (\hat{q}_{u_i}, \hat{q}_{v_i})$, where $\hat{q}_u \approx \mathbb{E}\left[Q_u \mid \mathbf{A} = \mathbf{a}\right]$. If the model provides a good fit to the data of line-ups then this variable will render the $a_{i,\ell}$ conditionally independent; see Appendix A.

1. Readers who are familiar with Wang and Blei (2019a) know that for some inferences, it makes another assumption, which is called "pinpointability." (Loosely, it means that the unmeasured confounders are an unknown but deterministic function of the player line-up.) Pinpointability is a strict assumption, but can be relaxed if the practitioner is willing to make other (parametric) assumptions; see the discussion in Wang and Blei (2019a, 2020).

There are several other details to fitting the factor model of line-ups. In variational inference, algorithmic convergence is determined by monitoring changes in log-likelihood. The latent dimension K is found using 5-fold cross-validation, considering held-out log-likelihood as a measure of model fitness. Finally, we evaluate the goodness-of-fit of the factor model by performing predictive checks, similar to those in Wang and Blei (2019a); see Appendix A for details.

3.3.2. THE OUTCOME MODEL

We discussed the lineup factor model for how players are sent to the field. Now we use this model to estimate Eq. 1, the value of each player. We appeal to the identification formula of Eq. 4, but with data that does not contain the confounders w_i. To use Eq. 4 with substitute confounders, we need to define the conditional expectation

$$\mu(z_i, a_\ell, a_\emptyset) = \mathbb{E}\left[Y \mid A_\ell = a_\ell, A_\emptyset = a_\emptyset, Z_i = z_i\right]. \tag{10}$$

We will define this expectation through an *outcome model*, a model of the conditional expectation of the difference in score given the line-ups of both teams and the substitute confounder. The simplest outcome model uses a difference of regressions,

$$\begin{aligned}\mathbb{E}\left[Y \mid \mathbf{A} = \mathbf{a}_i, Z_i = z_i\right] = \\ \left(\sum_{\ell=1}^{m} \mathbb{1}\left(t_\ell = u_i\right) a_{i,\ell}\, \beta_\ell + \gamma \cdot q_{v_i}\right) - \left(\sum_{\ell=1}^{m} \mathbb{1}\left(t_\ell = v_i\right) a_{i,\ell}\, \beta_\ell + \gamma \cdot q_{u_i}\right).\end{aligned} \tag{11}$$

This model contains per-player coefficients β_ℓ and per-team confounder coefficients γ_u. Eq. 11 models the expected score of each team as a sum of the contributions of each player on the field along with a term for the substitute confounders. It then subtracts the expected score for team v from the expected score for team u. (Note it is a descriptive model of the conditional expectation of the score differential; it does not make linear causal assumptions.)

Alternatively, we can use a "reconstructed causes" term in place of the substitute confounder (Wang and Blei, 2019a). The reconstructed causes are the conditional expectation of the lineup indicators given the substitute confounder,

$$\hat{a}_{i\ell} := \mathbb{E}\left[a_{i\ell} \mid z_i\right] = \sum_k z_{iuk}\theta_{\ell k}. \tag{12}$$

A linear outcome model with reconstructed causes is,

$$\begin{aligned}&\mathbb{E}\left[Y \mid \mathbf{A} = \mathbf{a}_i, Z_i = z_i\right] \\ &\quad = \left(\sum_\ell \mathbb{1}\left[t_\ell = u_i\right]\left(\beta_\ell\, a_{i\ell} + \gamma_\ell\, \hat{a}_{i\ell}\right)\right) - \left(\sum_\ell \mathbb{1}\left[t_\ell = v_i\right]\left(\beta_\ell\, a_{i\ell} + \gamma_\ell\, \hat{a}_{i\ell}\right)\right) \quad .\end{aligned} \tag{13}$$

This model contains per-player coefficients β_ℓ and per-reconstructed-player coefficients γ_ℓ. Wang and Blei (2019a) found that an outcome model using reconstructed causes often outperforms the simpler one that directly uses the substitute confounder. We study both in Section 4.

How do we estimate the outcome models? Consider a dataset of line-ups and scores $\{(u_i, v_i, \mathbf{a}_i, y_i)\}$. First estimate substitute confounders, as described in Section 3.3.1, and create an extended dataset

Algorithm 1 CAPM (causality-inspired adjusted plus-minus)

Require: lineups and score differentials $\{(\mathbf{a}_i, y_i)\}$

1: Approximate the posterior of the line-up factorization model, Eqs. 5 to 9.
2: Use the posterior to estimate substitute confounders for each game $\hat{z}_i = (\hat{q}_{u_i}, \hat{q}_{v_i})$.
3: Fit the outcome model to the extended dataset $\{(\mathbf{a}_i, y_i, \hat{q}_{u_i}, \hat{q}_{v_i})\}$, Eq. 11.
4: **Return**: The value (causal contribution) of each player: $\{\beta_\ell\}$.

$\{u_i, v_i, \hat{q}_{u_i}, \hat{q}_{v_i}, \mathbf{a}_i, y_i\}$. Then estimate the outcome model in Eq. 11 with maximum likelihood,

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg\max \sum_{i=1}^{n} \log p(y_i \mid \mathbf{a}_i, \hat{q}_{u_i}, \hat{q}_{z_i}\,;\, \boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{14}$$

where the response is Gaussian with mean in Eq. 11 or Eq. 13. In practice, we use regularized regression, such as with a ridge penalty. Finally, with the fitted outcome model, use a causality-inspired adjusted model to estimate the value of each player from Eq. 1. Specifically, we use the g-estimation strategy of Section 3.2, but where we replace the observed confounders with the substitute confounders. When we estimate $\mu(z_i, a_\ell, a_\emptyset)$ with the linear outcome model of Eq. 11, the result is to approximate the value of each player by their fitted regression coefficient,

$$\Lambda_\ell \approx \hat{\beta}_\ell. \tag{15}$$

See Appendix B for the derivation of this fact and see Algorithm 1 for the full algorithm.

3.4. Limitations of the causality-inspired adjusted plus-minus

We have defined the value of a player as a causal inference (Eq. 1) and described a method for estimating it. We use the factor model of Section 3.3.1 to identify co-occurrence patterns in how players are deployed to the field. We then use those patterns to estimate substitute confounders, variables that help us adjust for some of the confounders that are not explicitly measured.

The theory of factor modeling for indirectly measured confounding says that if the factor model of the line-up can indirectly measure all the unmeasured confounders, then the causal inference will be unbiased (Wang and Blei, 2019a, 2020). However, there are many assumptions and caveats to such an idealization. There are many complexities to the data which the models that comprise CAPM cannot capture.

At the outset, CAPM can only hope to capture unmeasured multi-cause confounders, variables that affect multiple players and the outcome. There may be many single-cause confounders in the data, e.g., individual sports injuries is an immediate example, that no factor model will be able to capture because they do not induce dependence among the players on the field.

Furthermore, the simple factor model will necessarily be an imperfect model of the line-up. Many covariates can affect the coaching decisions around the lineup. As examples, consider the status of the team as home or away, the weather on the day of the game, and other individualized factors. Even those factors that affect multiple players, as is required for a multi-cause confounders, may not be absorbed by the simple Poisson factorization described in Section 3.3.1. The performance our model may also be compromised when the posterior distributions of the substitute confounders are not concentrated. In this case, the posterior mean estimate of the substitute confounder

would be a poor approximation of the posterior distribution, which can induce bias in downstream causal estimation.

Another limitation is that the problem setting studied here may suffer from collinearity in the lineups. In particular, players who almost always appear in the same lineups will have individual contributions that may not be identifiable. We note that this problem is intrinsic to lineup data and any analysis based on regression, and it was the original motivation for applying ridge regression (Sill, 2010). We also note that while ridge regression will result in more stable estimates, it arbitrarily allocates credit approximately equally among the co-occurring players, which may not reflect actual player contributions. To appropriately handle collinear players, we can consider integrating other types of data such as tracking or play-by-play events, or using generalizations of factor models to hypergraphs, that better handle interactions involving more than two players (Contisciani et al., 2022; Ruggeri et al., 2023) (But these solutions require more complex models and finer-grained data, which goes behind the scope of this work.)

Finally, there is a time-series element to the line-up that the models above do not account for. Such an element is particularly prominent in the basketball data, which involves sequential segments of games, but also in the soccer data when considering the trajectory of a season of games. Innovations on this procedure could account for the sequential nature of the data.

In short, the factor model cannot provide a foolproof solution to adjusting for unmeasured confounding. However, we can use it to absorb some of it. In our empirical studies we study how effective it is.

4. Empirical studies

We evaluate CAPM with soccer and basketball data, both real and synthetic. We find that CAPM predicts the outcomes of the match better than the APM model with ridge regression (Eq. 2). Further, CAPM ranks players in a way that is more consistent with rankings produced by human experts.

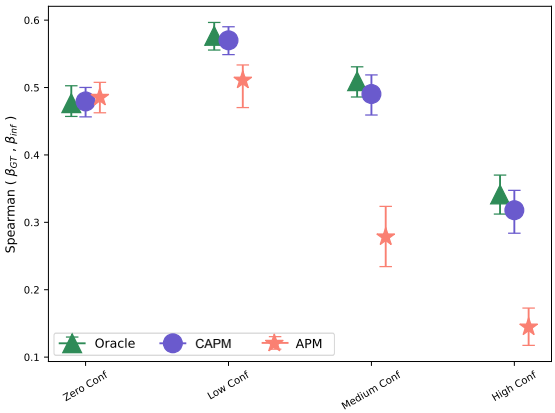


Figure 1: Synthetic data: ability to recover the ground-truth β. The causality-inspired adjusted plus-minus (CAPM) produces player rankings more consistent with the ground truth rankings than the standard APM for various confounding levels. We show the Spearman correlations between ground truth rankings β_{GT} versus the inferred ones β_{inf} for various confounding levels from low (leftmost) to high (rightmost), here higher is better. Each datapoint is the median with bars being the quartiles over 100 simulations.

We first study performance on synthetic data of simulated soccer outcomes. In this simulation, we know the true causal effect of each player. We synthesize confounded data and then evaluate the accuracy of different estimation methods. We generate data with various degrees of confounding. Data are generated in two steps: first, we generate confounders and lineups, sampling from a factor model of Section 3.3.1. (We chose hyperparameters by fitting to real soccer data.) Then we generate outcomes as in Section 3.3.2, for various levels of confounding (controlled by fixing β and γ, the

per-player coefficient and per-team confounder coefficient, respectively). We generate lineups of 11 players for 380 matches, as in a soccer league of 20 teams and about 500 players. We simulate the situation where one player per team has a much higher impact on the game than the others. See Appendix C for details about how the data is generated. From this data, we estimate the value of each player with different methods, specifically APM and CAPM. We evaluate performance in terms of closeness-to-truth of the estimates to the true causal estimates. In addition, we consider an "oracle" model which uses the true unobserved confounders and g-estimation for an unbiased estimate of the causal value of each player. (All details are in Appendix C.) Note that in these synthetic examples there is no evident violation of the assumptions described in section 3.4. In the study of real data below this may not be the case.

The results show that CAPM better recovers the ground truth values in all regimes except when there is little confounding; see Figure 1. As confounding increases, so does the improvement of CAPM's performance over APM. The Appendix contains further results.

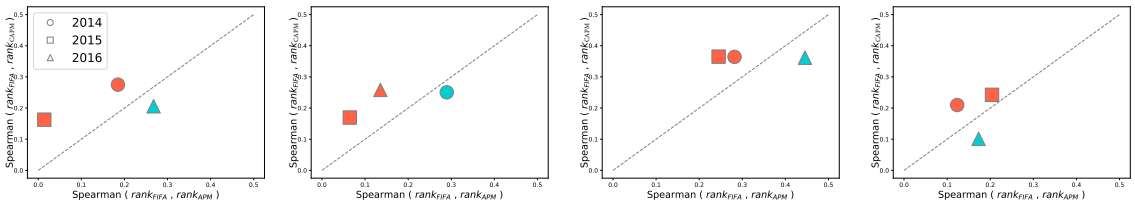
4.1. Real soccer and basketball data

We next study CAPM with soccer and basketball data. For soccer, we use an open-source dataset.[2] It contains the starting lineups and match outcomes for four professional men's league from 2014 to 2016; see Table 1. The datasets do not contain information about player substitutions during a match. We assume that a player contributed to the outcome if he was in the starting lineup, and we ignore the contribution of players that entered during the game. For basketball, we use proprietary NBA datasets containing lineups and scores from 2014 to 2018.[3] In contrast to the soccer dataset, the basketball data does contain detailed information about player substitutions. Thus we split the games into segments with constant lineups, generating a new segment at each substitution. As for synthetic data, we estimate the value of each player with APM ridge regression and with CAPM.

Ranking players by estimated value. We present the differences between the causal and ridge regression model by assessing the ridge regression's coefficients β_{CAPM} and β_{APM} obtained with and without the confounder (respectively), and rank the players by their decreasing values. While there is no ground truth for the value of a player, we can compare the rankings provided by CAPM with soccer ratings given by the FIFA video game[4]. The FIFA rankings are extracted from the subjective evaluations of over 9000 data-reviewers as scouts, coaches, and season-ticket holders into ratings for over 18,000 players (Lindberg, 2016). We calculate the Spearman correlation to measure how well a set of algorithmic rankings agrees with the expert FIFA rankings. Figure 2 compares the rankings provided by CAPM with those provided by the APM model. This metric allows to compare the rank of the data points, rather than the actual values of their scores. It is an appropriate metric in situation like ours where scores calculated using different methods have very different ranges. For the majority of the datasets, CAPM rankings have higher correlation to the expert FIFA rankings.

Using APM and CAPM as a predictive model. We next evaluate the two models as predictors of the outcome of matches. We quantitatively measure their abilities on both regular and *distributional shift* test sets. A test set with distributional shift is one that comes from a different distribution from the training set; a more accurate causal model should be robust to such changes. In particular,

2. https://www.kaggle.com/efezinoerome/analyzing-soccer-data/data
3. https://www.bigdataball.com/
4. www.sofifa.com



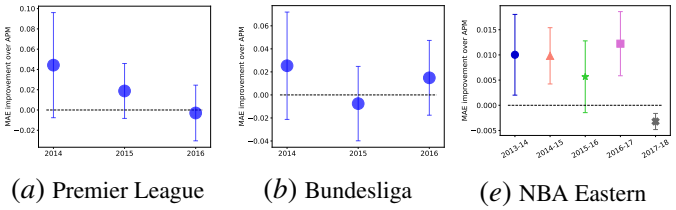
(*a*) Premier League (*b*) Bundesliga (*c*) Serie A (*d*) La Liga

Figure 2: On 2/3 of the datasets, the causality-inspired adjusted plus-muns (CAPM) leads to player rankings more consistent with the FIFA rankings by human experts. (Above the diagonal line is better.) Each dot represents the results on a dataset (4 European soccer leagues and 3 seasons); it shows the Spearman correlation between FIFA rankings ($rank_{\text{FIFA}}$) versus CAPM ($rank_{CAPM}$) and APM ($rank_{APM}$) on this dataset. Rankings are calculated using the inferred β and sorting them, so that the player with highest β is the first in the ranking. Red markers denote datapoints where $rank_{CAPM}$ has higher Spearman than $rank_{APM}$.

we construct the shifted test set based on the match date, selecting matches that happen at the very end of the season where the lineups are often different from the typical lineup used during the bulk of the season.[5] In contrast, a regular test set contains matches that are selected uniformly at random across time. With both shifted and unshifted test sets, we compare CAPM with reconstructed causes (Eq. 13) to the APM model. For each match, we evaluate the Mean Absolute Error (MAE) between the predicted outcome and the observed outcome, averaging over all the matches in the test set.

CAPM consistently outperforms APM, across leagues and seasons in both sports, for the majority of unshifted test sets (Figure 3) and all of the shifted test sets (Table 3). This improvement is confirmed when comparing trial-by-trial, as reported in Table 2. CAPM improves the prediction of scores and its performance is stable to shift in distribution.

Figure 3: *CAPM vs APM model performance.* We show the distribution of differences in MAE for the predicted score difference of the CAPM compared to APM. Error bars indicate standard deviations and markers show mean, corresponding to 100 independent trials of cross-validation; in each trial 20% of the matches where held-out (test set). Points above the horizontal line at 0 correspond to better performance for the CAPM.



(*a*) Premier League (*b*) Bundesliga (*e*) NBA Eastern

(*c*) Serie A (*d*) La Liga (*f*) NBA Western

5. Discussion

We developed CAPM, a causality-inspired approach to individual player evaluation in sports matches. CAPM first fits a Poisson factorization to the team lineups to construct a substitute confounder; then

5. The reason is that, at the end of the season, the relative standing of the teams will determine the fight to reach playoffs, win the title, or avoid relegation; thus the pressure to win kicks in more urgently than during the rest of the season.

it uses that substitute in an expanded APM of the match outcome. Across simulated and real matches of different European soccer and NBA basketball leagues, CAPM makes better predictions than regularized APM, a widely-used method. It produces player rankings that are more consistent with human expert rankings. Its main output is a coefficient that indicates the value of a player. This coefficient accounts for "hidden" confounders, those for which there is evidence in the statistics of the lineups of the players. By comparing this estimate with that obtained from standard models as APM on the same input data, scouts can find potential overlooked players and better evaluate players' contributions in a game.

CAPM can be easily adapted to serve the specific needs of practitioners. For example, we can model other team-level outcome variables, such as the number of goals, shots, or assists, and we can also involve a multivariate outcome variable. In a multivariate analysis, we can include defensive outcome variables as well, such as the number of rebounds or tackles. More generally, CAPM would benefit from possible extensions when provided with additional data as expected goal or assists, and could be customized to focus on evaluating particular skills.

While the data we analyzed only contained lineups and match outcomes, a real-world application of CAPM should also condition on known confounding variables, such as who is the home team or the injury status of individual players. Conditioning on known confounders relieves the factor model of needing to capture them, and further improves the bias of CAPM's estimates. In a similar vein, the Poisson factorization is a simple model of lineups. CAPM provides an application for good modeling of pre-game lineups, particularly those that use the domain knowledge of the sport at hand. Finally, the simple algorithm presented here does not account for the sequential nature of the data. The basketball data, in particular, contains segments within games. Developing causal adjusted algorithms that model such a time series is a fruitful avenue for future work.

Acknowledgments

This work has been partially supported by the Office of Naval Research under grant number N00014-23-1-2590 and the National Science Foundation under grant number CHE-2231174 and DMS-2310831.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. The decamfounder: Non-linear causal discovery in the presence of hidden variables. *arXiv preprint arXiv:2102.07921*, 2021.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

BBall Index Team. Lebron: The man, the myth, the metric? Accessed on 17.02.2022.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Fan Bu, Sonia Xu, Katherine Heller, and Alexander Volfovsky. Smogs: Social network metrics of game success. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2406–2414. PMLR, 2019.

Humberto Moreira Carvalho, José A Lekue, Susana M Gil, and Iraia Bidaurrazaga-Letona. Pubertal development of body size and soccer-specific functional capacities in adolescent players. *Research in Sports Medicine*, 25(4):421–436, 2017.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorization models. *Computational Intelligence and Neuroscience*, 2009:785152, 2009.

Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry. A multiresolution stochastic process model for predicting basketball possession outcomes. *Journal of the American Statistical Association*, 111(514):585–599, 2016.

Martina Contisciani, Federico Battiston, and Caterina De Bacco. Inference of hyperedges and overlapping communities in hypergraphs. *Nature communications*, 13(1):7229, 2022.

Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*, 2019.

Sameer K Deshpande and Shane T Jensen. Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12(2):51–72, 2016.

Javier Fernández, Luke Bornn, and Daniel Cervone. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, 110(6):1389–1427, 2021.

Alexander Franks, Andrew Miller, Luke Bornn, Kirk Goldsberry, et al. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9(1): 94–121, 2015.

Benjamin Frot, Preetam Nandy, and Marloes H Maathuis. Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3):459–487, 2019.

Romain Gauriot and Lionel Page. Fooled by performance randomness: Overrewarding luck. *Review of Economics and Statistics*, 101(4):658–666, 2019.

Brandon Giles, Paul SR Goods, Dylan R Warner, Dale Quain, Peter Peeling, Kagan J Ducker, Brian Dawson, and Daniel F Gucciardi. Mental toughness and behavioural perseverance: A conceptual replication and extension. *Journal of science and medicine in sport*, 21(6):640–645, 2018.

P. Gopalan, J. Hofman, and D. Blei. Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*, pages 326–335, 2015.

Robert B Gramacy, Shane T Jensen, and Matt Taddy. Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, 9(1):97–111, 2013.

J. Grimmer, D. Knox, and B. Stewart. Naïve regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *arXiv 2007.12702*, 2020.

Miguel A Hernán and James M Robins. *Causal inference*. CRC Boca Raton, FL, 2016.

Lars Magnus Hvattum. A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, 18(1):1–23, 2019.

S Ilardi and A Barzilai. Adjusted plus-minus ratings: New and improved for 2007-2008. http://www.82games.com/ilardi2.htm, 2008.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Tarak Kharrat, Ian G McHale, and Javier López Peña. Plus–minus player ratings for soccer. *European Journal of Operational Research*, 283(2):726–736, 2020.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Hoang M Le, Peter Carr, Yisong Yue, and Patrick Lucey. Data-driven ghosting using deep imitation learning. In *MIT Sloan Sports Analytics Conference*, 2017.

D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

D. Leonhardt. Pro basketball; mavericks' new math may be an added edge. https://www.nytimes.com/2003/04/27/sports/pro-basketball-mavericks-new-math-may-be-an-added-edge.html, 2003.

A. Lindberg. FIFA 17's player ratings system blends advanced stats and subjective scouting. http://www.espn.com/soccer/blog/espn-fc-united/68/post/2959703/, 2016.

B. Macdonald. An improved adjusted plus-minus statistic for NHL players. In *Proceedings of the MIT Sloan Sports Analytics Conference*, volume 3, 2011a.

B. Macdonald. A regression-based adjusted plus-minus statistic for NHL players. *Journal of Quantitative Analysis in Sports*, 7(3), 2011b.

Francesca Matano, Lee Richardson, Taylor Pospisil, Collin A Politsch, and Jining Qin. Augmenting adjusted plus-minus in soccer with fifa ratings. *Journal of Quantitative Analysis in Sports*, 19(1): 43–49, 2023.

Kostya Medvedovsky. Daily adjusted and regressed kalman optimized projections — darko. Accessed on 17.02.2022.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Namita Nandakumar and Shane T Jensen. Historical perspectives and current directions in hockey analytics. *Annual Review of Statistics and Its Application*, 2019.

E. Ogburn, I. Shpitser, and E. Tchetgen Tchetgen. Comment on "Blessings of multiple causes". *Journal of the American Statistical Association*, 114(528):1611–1615, 2019.

E. Ogburn, I. Shpitser, and E. Tchetgen Tchetgen. Counterexamples to "The blessings of multiple causes" by wang and blei. *arXiv:2020.001*, 2020.

Giovanni Pantuso and Lars Magnus Hvattum. Maximizing performance with an eye on the finances: a chance-constrained model for football transfer market decisions. *TOP*, 29(2):583–611, 2021.

Judea Pearl. [bayesian analysis in expert systems]: Comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Konstantinos Pelechrinis and Wayne Winston. A skellam regression model for quantifying positional value in soccer. *Journal of Quantitative Analysis in Sports*, 2021.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

D. T. Rosenbaum. Measuring how NBA players help their teams win, 2004. URL http://www.82games.com/comm30.htm.

Nicolò Ruggeri, Martina Contisciani, Federico Battiston, and Caterina De Bacco. Community detection in large hypergraphs. *Science Advances*, 9(28):eadg9159, 2023.

Olav Drivenes Sæbø and Lars Magnus Hvattum. Modelling the financial contribution of soccer players to their clubs. *Journal of Sports Analytics*, 5(1):23–34, 2019.

Nathan Sandholtz and Luke Bornn. Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball. *The Annals of Applied Statistics*, 14(3): 1122–1145, 2020.

Edgar Santos-Fernandez, Paul Wu, and Kerrie L Mengersen. Bayesian statistics meets sports: a comprehensive review. *Journal of Quantitative Analysis in Sports*, 15(4):289–312, 2019.

Aaron Schein, John Paisley, David M Blei, and Hanna Wallach. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1045–1054, 2015.

Steven R Schultze and Christian-Mathias Wellbrock. A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics*, 4(2):121–131, 2018.

J. Sill. Improved NBA adjusted+/-using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT Sloan Sports Analytics Conference*, 2010.

Nate Silver. Introducing raptor, our new metric for the modern NBA. https://fivethirtyeight.com/features/introducing-raptor-our-new-metric-for-the-modern-nba/, 2019. Accessed on 17.02.2022.

Taylor Snarr. NBA player metric comparison. https://dunksandthrees.com/blog/metric-comparison, 2020. Accessed on 17.02.2022.

Zachary Terner and Alexander Franks. Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, 8:1–23, 2021.

A. C. Thomas and S. L. Ventura. The road to war. http://blog.war-on-ice.com/index.html%3Fp=429.html, 2015.

AC Thomas, Samuel L Ventura, Shane T Jensen, and Stephen Ma. Competing process hazard function models for player ratings in ice hockey. *The Annals of Applied Statistics*, pages 1497–1524, 2013.

Camille Thomas, Gilbert Fellingham, and Pat Vehrs. Development of a notational analysis system for selected soccer skills of a women's college team. *Measurement in Physical Education and Exercise Science*, 13(2):108–121, 2009.

Y. Wang and D. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019a.

Y. Wang and D. Blei. Towards clarifying the theory of the deconfounder. *arXiv preprint arXiv:2003.04948*, 2020.

Yixin Wang and David Blei. The blessings of multiple causes: Rejoinder. *Journal of the American Statistical Association*, 114(528):1616–1619, 2019b.

Derrick R Yam and Michael J Lopez. What was lost? a causal estimate of fourth down behavior in the national football league. *Journal of Sports Analytics*, 5(3):153–167, 2019.

Appendix A. The factor model: Variational inference and model checking

A.1. Variational inference of the factor model

We give the detailed coordinate-ascent variational inference updates for the parameters of our model (Blei et al., 2017; Gopalan et al., 2015). The latent variables for the model are the players factors θ_ℓ and the team factors q_v, both are K-dimensional. First, we posit the *mean-field* variational family over the latent variables:

$$h(\theta, q) = \prod_{\ell,k} h(\theta_{\ell k}|\gamma_{\ell k}^{shape}, \gamma_{\ell k}^{rte}) \prod_{v,k} h(q_{vk}|\lambda_{vk}^{shape}, \lambda_{vk}^{rte}) \quad , \tag{16}$$

where all the $h(\cdot)$ are Gamma distributions. After specifying the family, we fit the variational parameters $\nu = \{\gamma, \lambda\}$ to minimize the KL divergence to the posterior, and then use the corresponding variational distribution $h(\cdot|\nu^*)$ as its proxy. We optimize the parameters with a coordinate ascent algorithm, see details in Algorithm 2. Finally, the optimal parameters are obtained by using the geometric expectations over the posteriors, as explained in Schein et al. (2015):

$$\theta_{\ell k} = \frac{\exp\left\{\Psi(\gamma_{\ell k}^{shape})\right\}}{\gamma_{\ell k}^{rte}} \tag{17}$$

$$q_{vk} = \frac{\exp\left\{\Psi(\lambda_{vk}^{shape})\right\}}{\lambda_{vk}^{rte}} \quad . \tag{18}$$

A.2. Selecting K and model checking

The model takes in input the number of factors K. This can be selected using cross-validation, where we hide part of the dataset (test set), the lineups of a subset of the matches. Then we train the parameters on the set of matches that where not hidden (training set) and calculates performance on the test set. Here as performance metric we use held-out log-likelihood and split train and test sets using 5-fold cross validation. This experiment is repeated for several values of K, the best K is selected as the one that gives highest performance averaging over the test sets.

The goodness-of-fit of the factor model is performed using posterior predictive checks, one form of model checking, similar to those in Wang and Blei (2019a). In detail, we first create a matrix of lineup held out data a^{heldout} by randomly hiding matches from the dataset. Then, we create a replicated dataset a^{rep} where for each match we draw s samples of the two lineups from the posterior predictive. This is approximated by sampling from $p(a_{i\ell}^{\text{rep}}|\bar{\theta}_\ell, \bar{q}_u)$ for $s_\ell = u_i$ (and similarly using $\bar{q}_v$ for $s_\ell = v_i$), where $\bar{\theta}_\ell$ and $\bar{q}_u$ are the posterior mean calculated as in Eqs. 17-18 from the observed dataset. Finally, we calculate the log-likelihood $\mathcal{L}(a^{\text{rep}})$ and $\mathcal{L}(a^{\text{heldout}})$ on the replicated and on the held out dataset, respectively. The posterior predictive p-value is $p(\mathcal{L}(a^{\text{rep}}) > \mathcal{L}(a^{\text{heldout}}))$. It can be estimated empirically by sampling m different replicated datasets and producing the ratio $\frac{m_R}{m}$ where m_R is the number of datasets in which $\mathcal{L}(a^{\text{rep}}) > \mathcal{L}(a^{\text{heldout}})$. We obtain p-values close to 0.5, suggesting that the model explains the replicated data as well as it explains the heldout data.

A.3. Conditional independence implied by the factor model

Suppose the check above returns that the factor model fits the data well. It implies that the factor model can generate replicated data that is indistinguishable with the observed data in terms of log-likelihood. Roughly, it implies that

$$p(\mathbf{a}_{1:n}) = \int \prod_{i,\ell} p(a_{i,\ell} \mid q_{1:t}, \theta_{1:m}) p(\theta_{1:m}) p(q_{1:t}) \mathrm{d}\theta_{1:m} \mathrm{d}\mathrm{d}q_{1:t}), \tag{19}$$

$$\approx \prod_{i,\ell} p(a_{i,\ell} \mid \hat{q}_{1:t}\,;\, \hat{\theta}^*_{1:m}) \tag{20}$$

The second equation is due to the posteriors of $\theta_{1:m}$ and $q_{1:t}$ being close to a point mass $p(\theta_{1:m}, q_{1:t} \mid \mathbf{a}_{1:n}) \approx \delta_{\hat{\theta}^*_{1:m}} \delta_{\hat{q}_{1:t}}$. Eq. 20 implies that the latents $\hat{q}_{1:t}$ render the matches $\mathbf{a}_{1:n}$ conditionally independent.

Algorithm 2 causality-inspired adjusted plus-minus (CAPM) factor model: CAVI algorithm

Data: lineups $\{a_i\}_i$,number of factors K, hyper-prior α, β.

Result: parameters $\theta = \left[\theta_{\ell,k}\right]$, $q = [q_u]$.

For all players and teams, initialize the team parameters $\gamma_\ell^{shape}, \gamma_\ell^{rte}$ and team parameters $\lambda_u^{shape}, \lambda_u^{rte}$ at random.

while *convergence not satisfied* **do**

For each player/team such that $a_{i,\ell} > 0$ and $s_\ell \neq v_i$, update the multinomial

$$\phi_{\ell,v_i} \propto \exp\left\{\Psi(\gamma_{\ell k}^{shape}) - \log \gamma_{\ell k}^{rte} + \Psi(\lambda_{v_i k}^{shape}) - \log \lambda_{v_i k}^{rte}\right\}$$

For each player update the shape and rate parameters of her posterior

$$\gamma_{\ell k}^{shape} = \alpha + \sum_{i|s_\ell \neq v_i} a_{i,\ell}\phi_{\ell,v_i k} \tag{21}$$

$$\gamma_{\ell k}^{rte} = \beta + \sum_{i|s_\ell \neq v_i} \frac{\lambda_{v_i k}^{shape}}{\lambda_{v_i k}^{rte}} \tag{22}$$

For each team update the shape and rate parameters of its posterior

$$\lambda_{vk}^{shape} = \alpha + \sum_{i|v_i = v, \ell|s_\ell \neq v} a_{i,\ell}\phi_{\ell,vk} \tag{23}$$

$$\lambda_{vk}^{rte} = \beta + \sum_{i|v_i = v, \ell|s_\ell \neq v} \frac{\gamma_{\ell k}^{shape}}{\gamma_{\ell k}^{rte}} \tag{24}$$

end

Calculate posterior estimates:

$$\theta_{\ell k} = \frac{\exp\left\{\Psi(\gamma_{\ell k}^{shape})\right\}}{\gamma_{\ell k}^{rte}} \tag{25}$$

$$q_{vk} = \frac{\exp\left\{\Psi(\lambda_{vk}^{shape})\right\}}{\lambda_{vk}^{rte}} \tag{26}$$

Appendix B. Causal interpretation of β

To see how the estimate in Eq. 15 is obtained, we flesh out how to use the substitute confounders and an outcome model for the causal inference of Section 3.2. First define a conditional expectation, similar to Eq. 3 but with substitute confounders,

$$\mu_\ell(\hat{q}, a_\ell, a_\emptyset) = \mathbb{E}\left[Y \mid \hat{Q} = \hat{q}, A_\ell = a_\ell, A_\emptyset = a_\emptyset, U = t_\ell\right]. \tag{27}$$

Again, the expectation is over the joint distribution of the other team v and which players (other than ℓ and $\emptyset$) are on the field.

Next write it with an iterated expectation,

$$\hat{\mu}_\ell(\hat{q}, a_\ell, a_\emptyset) = \mathbb{E}_{\mathbf{A}_{-\ell}, V}\left[\mathbb{E}\left[Y \mid \hat{Q} = \hat{q}, A_\ell = a_\ell, A_\emptyset = a_\emptyset, U = t_\ell, V, \mathbf{A}_{-\ell}\right]\right]. \tag{28}$$

Now we use the outcome model. Approximate the inner expectation with the fitted linear regression of Eq. 11,

$$\hat{\mu}_\ell(\hat{z}, a_\ell, a_\emptyset) \approx \hat{\beta}_\ell + \hat{\gamma} \cdot \hat{q}_{u_{t_\ell}} + \mathbb{E}_{\mathbf{A}_{-\ell}, V}\left[\hat{\beta}_{-\ell} \cdot \mathbf{A}_{-\ell} - \hat{\gamma} \cdot \hat{q}_V\right]. \tag{29}$$

Note, by definition, the 0-value player has $\beta_\emptyset = 0$. Finally, plug this estimate into the identification formula of Eq. 4 to obtain that $\Lambda_\ell \approx \hat{\beta}_\ell$.

Appendix C. Synthetic data generation

We generated synthetic data using a causal model with various degrees of confounding. We considered the soccer case, i.e. 11 players per lineup, and chose parameters such that the generated outcomes resemble the observed ones from real datasets.

Data are generated in two steps. First we generate Gamma-distributed unobserved confounders z as in Section 3.3.1. We use the inferred values obtained fitting to real data. We do this to consider realistic values. Then we generate synthetic lineups of 380 matches, as in a soccer league of 20 teams and about 500 players. We sample each lineup using a distribution that depends on the unobserved confounder as follows. We first consider the factor model distribution as in the main manuscript to assign scores to each player and game:

$$\mathbb{E}\left[a_{i,\ell}\right] = \theta_\ell \cdot q_v \qquad \text{if } s_\ell = u_i \tag{30}$$

$$\mathbb{E}\left[a_{i,\ell}\right] = \theta_\ell \cdot q_u \qquad \text{if } s_\ell = v_i \tag{31}$$

$$\mathbb{E}\left[a_{i,\ell}\right] = 0 \qquad \text{if } s_\ell \neq u_i \text{ and } s_\ell \neq v_i \quad . \tag{32}$$

Then we sample without replacement from a list of all eligible players a lineup of 22 players, with sampling probabilities proportional to these scores.

We then build reconstructed causes using the unobserved confounder as in Section 3.3.1 and then simulate outcomes for each match that are also dependent on the confounder using:

$$\begin{aligned} y_i &\sim \mathcal{N}(\delta_i, \sigma^2) \\ \delta_i &= \left(\sum_{\ell=1}^{m} \mathbb{1}(t_\ell = u_i)(a_{i,\ell} - \hat{a}^U_{i,\ell})\beta_\ell\right) - \left(\sum_{\ell=1}^{m} \mathbb{1}(t_\ell = v_i)(a_{i,\ell} - \hat{a}^V_{i,\ell})\beta_\ell\right), \end{aligned} \tag{33}$$

where $\sigma = 0.1$. To make the outcome discrete, we round the real values to integer.

We consider various regimes of confounding by varying the ground truth β and γ. These are both extracted from a Gaussian distribution but with different parameters. In addition, we select one player per team who has an expected higher β and call this set of overall 20 players as "Top players".

Specifically, the parameters are generated using:

$$\begin{aligned} \beta_\ell &\sim \mathcal{N}(0.0, \sigma) \quad \text{if } \ell \text{ is not Top Player} && (34)\\ \beta_\ell &\sim \mathcal{N}(0.1, \sigma_{TP}) \quad \text{if } \ell \text{ is Top Player} && (35)\\ \gamma_k &\sim \mathcal{N}(0.0, \sigma_\gamma) \quad . && (36)\\ & && (37) \end{aligned}$$

Here are the details of the different regimes:

Zero confounding: $\sigma = 0.1$, $\sigma_{TP} = 0.1$, $\sigma_\gamma = 0.01$.

Low confounding: $\sigma = 0.2$, $\sigma_{TP} = 0.2$, $\sigma_\gamma = 0.2$.

Medium confounding: $\sigma = 0.2$, $\sigma_{TP} = 0.2$, $\sigma_\gamma = 0.5$.

High confounding: $\sigma = 0.1$, $\sigma_{TP} = 0.1$, $\sigma_\gamma = 0.5$.

Appendix D. Synthetic and real data extra results

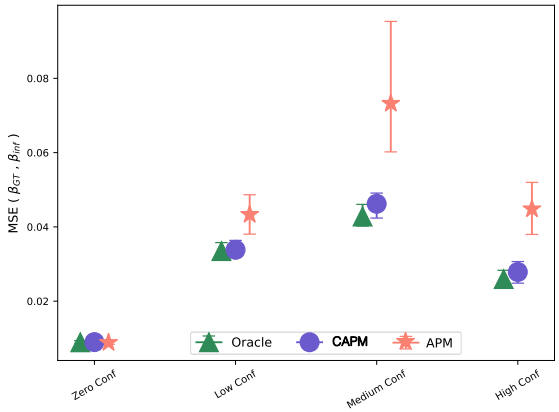


Figure 4: Synthetic data: ability to recover the ground-truth β. CAPM produces player rankings more consistent with the ground truth rankings than the standard APM for various confounding levels. We show the MSE between inferred and the ground truth β, here lower is better. Each datapoint is the median with bars being the quartiles over 100 simulations.

Table 1: Basketball and soccer datasets used in the empirical study. The number of players that play at least one game varies across seasons and leagues, from $M = 417$ for the German Bundesliga to $M = 515$ for the Italian Serie A (similar numbers for the NBA). The number of NBA games slightly varies depending on the season and playoffs. We consider 3 seasons from 2014 to 2016 for the soccer dataset and 5 seasons from 2014 to 2018 for the NBA dataset.

League	Country	N_{teams}	N_{games}
English Premier League	England	20	380
Italian Serie A	Italy	20	380
Spanish La Liga	Spain	20	380
German Bundesliga	Germany	18	306
NBA	USA	30	$\sim$ 1300

Table 2: CAPM vs APM model performance. In the table we compare the performance of the APM model (MAE_{APM}) and CAPM (MAE_{CAPM}), and present the percentage of trials (over 100) where the CAPM does better (lower values of the errors). In the columns MAE_{APM} and MAE_{CAPM} we report the average MAE over 100 trials.

Sport	League	Season	MAE_{APM}	MAE_{CAPM}	% MAE_{APM} > MAE_{CAPM}
Soccer	English Premier League	2014	1.36	1.31	**0.86**
		2015	1.24	1.22	**0.75**
		2016	1.25	1.25	0.44
	German Bundesliga	2014	1.46	1.43	**0.77**
		2015	1.26	1.27	0.36
		2016	1.37	1.36	**0.65**
	Italian Serie A	2014	1.26	1.21	**0.95**
		2015	1.18	1.18	**0.54**
		2016	1.22	1.2	**0.81**
	Spanish La Liga	2014	1.38	1.33	**0.9**
		2015	1.27	1.21	**0.93**
		2016	1.38	1.32	**0.9**
Basketball	Eastern	2014	2.284	2.274	**0.79**
		2015	2.276	2.266	**0.94**
		2016	2.314	2.309	**0.69**
		2017	2.373	2.361	**0.9**
		2018	2.357	2.36	0.02
	Western	2014	2.312	2.3	**0.95**
		2015	2.255	2.246	**0.97**
		2016	2.318	2.317	0.45
		2017	2.323	2.318	**0.78**
		2018	2.373	2.369	0.4

Table 3: CAPM vs APM model performance, *shifted* test set. In the table we compare the MAE for the predicted score difference of the CAPM compared to APM for a shifted test set where we hide the last 20% of the matches in a season.

Sport	League	Season	MAE_{APM}	MAE_{CAPM}
Soccer	English Premier League	2014	1.501	**1.46**
		2015	**1.369**	1.39
		2016	**1.357**	1.369
	German Bundesliga	2014	1.343	**1.309**
		2015	1.065	**1.044**
		2016	1.384	**1.37**
	Italian Serie A	2014	1.25	**1.212**
		2015	1.279	**1.269**
		2016	1.459	**1.392**
	Spanish La Liga	2014	1.302	**1.282**
		2015	1.344	**1.28**
		2016	1.447	**1.334**
Basketball	NBA Eastern	2014	2.352	**2.339**
		2015	2.353	**2.344**
		2016	2.39	**2.373**
		2017	2.403	**2.39**
		2018	**2.311**	2.314
	NBA Western	2014	2.336	**2.328**
		2015	2.283	**2.269**
		2016	**2.332**	2.335
		2017	2.381	**2.375**
		2018	**2.342**	2.349