UCLouvain

epl

École polytechnique de Louvain

# Predicting sports outcomes

## Investigating adaptive learning rates in the Elo rating system

Author: **Robin Rochet**
Supervisors: **Pierre-Antoine Absil, Julien Hendrickx**
Readers: **Pierre-Antoine Absil, Julien Hendrickx, Bastien Massion, Eric Piette**
Academic year 2025–2026
Master [120] in Mathematical Engineering

**UNIVERSITÉ CATHOLIQUE DE LOUVAIN**
**École polytechnique de Louvain**
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**

# Abstract

This thesis investigates the prediction of sports outcomes through rating systems. Usually, these systems use a constant parameter, denoted as $K$, to update the ratings. In a standard rating system, the $K$ parameter determines how strongly ratings react to new observations (i.e. result of a new game). Although it is generally chosen as a fixed constant, such a specification may be poorly suited to environments in which players' underlying abilities evolve over time. Such a choice also overlooks the uncertainty associated with the start of the season compared to the end of the season. The aim of this work is to improve prediction abilities of existing systems by proposing adaptive methods where this $K$ parameter is not fixed anymore, with a particular emphasis on the Elo Rating System.

Several adaptive $K$ parameter methods are therefore proposed and compared. These include optimisation-based approaches that select $K$ according to past rating trajectories, terminal rating estimates, or predictive performance, as well as analytical rules based on uncertainty, outcome variance, persistent prediction errors, empirical rating volatility, and change-point detection. The methods are evaluated against a constant-$K$ baseline, a decreasing $K$ parameter, and an oracle benchmark.

Their performance is studied through experimental simulations involving a wide range of dynamic ability profiles and through an application to real basketball data. The comparison relies on predictive accuracy, the Brier score, and a centred measure of rating estimation error. The results assess whether allowing the $K$ parameter to respond to observed information can improve both outcome prediction and the tracking of time-varying player abilities.

The experiments show that, although some scenarios are favorable to adaptive methods, their performance remains too unstable to constitute a reliable alternative to a constant-$K$ rule. Results on real basketball data reinforce this finding, suggesting that the additional flexibility offered by adaptive $K$ factors does not translate into sufficiently consistent predictive improvements to outweigh the robustness of a well-calibrated constant $K$.

# Acknowledgements

I would first like to express my sincere gratitude to my supervisors. Pr. Pierre-Antoine Absil for his guidance, support, and the time devoted to discussing the progress of this thesis. His perspective and feedback greatly contributed to improving both the methodology and the overall quality of this work. As well as Pr. Julien Hendrickx for his attentiveness, and valuable feedback throughout this master's thesis. His advice played an important role in shaping the direction of this work and in helping me address the different challenges encountered during the thesis.

I would also like to warmly thank Bastien Massion, who co-supervised the thesis, for his presence and implication. His awareness on that topic has contributed to the quality of the thesis.

I would like to extend my thanks to my friends for the many discussions, shared experiences, and encouragement throughout the year. I would especially like to thank Anthony Richelle for his help with several technical aspects at crucial moments, as well as for rereading the entire thesis and providing valuable feedback.

Finally, I would like to thank my family and those close to me for their constant support and for providing an encouraging environment throughout the completion of this work.

# Contents

# List of Figures

# List of Tables

# Notations

| Symbol | Meaning | Main use |
|---|---|---|
| $\hat{E}_{it}$ | Estimated rating of player or team $i$ before matchday $t$. | Rating framework; all methods |
| $E_{it}$ | Intrinsic rating of player or team $i$ before matchday $t$. | Simulations; rating error |
| $\hat{P}_{ijt}$ | Predicted probability that player $i$ defeats player $j$ at matchday $t$. | BTL/Elo prediction; metrics |
| $P_{ijt}$ | Intrinsic probability that player $i$ defeats player $j$ at matchday $t$. | Simulated outcomes |
| $\hat{Y}_{ijt}$ | Predicted outcome in the match between players $i$ and $j$ at matchday $t$. | Accuracy |
| $Y_{ijt}$ | Observed outcome in the match between players $i$ and $j$ at matchday $t$. | Rating updates; metrics |
| $\mathcal{Z}_{t-1}$ | Set of all information available before time $t$. | Admissibility condition |
| $D_t$ | Set of matches played at matchday $t$. | Rating framework; metrics |
| $M_t$ | Set of all matches played from time 1 to time $t$ (included). | Numerical adaptive methods |
| $M$ | Set of all observed matches; the total number of matches considered. | Objective; accuracy |
| $N$ | Number of players or teams in a league format. | Rating framework; experiments |
| $T$ | Number of matchdays in a season. | Rating framework; experiments |
| $K$ | Rating update factor not depending on player nor time. | Elo system; general notation |
| $K_i$ | Rating update factor used for player $i$, not depending on time. | Empirical adaptation rules |
| $K_t$ | Rating update factor shared by all players at matchday $t$. | Numerical adaptive methods |
| $K_{it}$ | Player-specific rating update factor at matchday $t$. | Numerical and analytical methods |
| $K_{ijt}$ | Confrontation-specific rating update factor at matchday $t$. | Numerical adaptive methods |
| $K_0$ | Initial value of the update factor. | Adaptive-method initialization |
| $K_{\min}$ | Minimum admissible value of the adaptive update factor. | Adaptive methods |

| Symbol | Meaning | Main use |
|---|---|---|
| $K_{\max}$ | Maximum admissible value of the adaptive update factor. | Adaptive methods |
| $K_{\text{const}}$ | Constant update factor used by the constant-$K$ baseline. | Baseline method |
| $K^{\star}_{\text{const}}$ | Calibrated constant update factor used in the experiments. | Preliminary calibration |
| $K^{\star}_{s,r}$ | Optimal constant-$K$ value selected for scenario $s$ and calibration replication $r$. | Preliminary calibration |
| $m_t$ | Optimal centering shift used in the centred MAE at matchday $t$. | Centred MAE |
| $d_{it}$ | Individual rating estimation error, $d_{it} = \hat{E}_{it} - E_{it}$. | Centred MAE |
| $BS$ | Overall Brier score. | Metrics |
| $BS_t$ | Brier score at matchday $t$. | Metrics |
| $k$ | Candidate scalar update factor considered in a numerical optimization. | Numerical adaptive methods |
| $\mathbf{k}$ | Candidate vector of player-specific update factors considered in a numerical optimization. | Numerical adaptive methods |
| $\hat{P}^{k}_{ij\tau}$ | Hypothetical predicted probability obtained by replaying the history using candidate factor $k$. | Historical-trajectory method |
| $\hat{P}^{k,\text{ind}}_{ij\tau}$ | Hypothetical predicted probability obtained by replaying the history with player-specific candidate factors. | Player-specific numerical methods |
| $k^{\star}_{i(t+1)}$ | Optimal candidate factor associated with player $i$ when constructing a confrontation-specific update factor for time $t+1$. | Confrontation-specific methods |
| $\hat{E}^{k}_{i(t+1)}$ | Terminal estimated rating induced by candidate factor $k$. | Terminal-rating methods |
| $e_{it}$ | Signed prediction residual for player $i$, $e_{it} = Y_{ijt} - \hat{P}_{ijt}$. | Analytical adaptive methods |
| $U_{it}$ | Uncertainty state associated with player $i$. | Uncertainty-based method |
| $\eta$ | Smoothing or memory parameter. | Analytical adaptive methods |
| $c$ | Scaling parameter. | Analytical adaptive methods |
| $D_{it}$ | Smoothed signed residual associated with player $i$. | Persistent-residual method |
| $V_{it}$ | Empirical-volatility state associated with player $i$. | Empirical-volatility method |

| Symbol | Meaning | Main use |
|---|---|---|
| $C_{it}^{+}$ | Positive cumulative detection statistic for player $i$. | Change-point method |
| $C_{it}^{-}$ | Negative cumulative detection statistic for player $i$. | Change-point method |
| $\delta$ | Tolerance parameter. | Change-point method |
| $h$ | Detection threshold. | Change-point method |
| $n_{it}$ | Number of matchdays since the last detected change for player $i$. | Change-point method |
| $\tau$ | Decay parameter controlling how quickly the update factor returns toward its lower value after a detected change. | Change-point method |
| $H_t$ | Number of matches observed up to matchday $t$, $H_t = \lvert M_t \rvert$. | Complexity analysis |
| $Q_t^{(1)}$ | Number of scalar candidate factors evaluated by the stochastic search at time $t$. | Complexity analysis |
| $Q_t^{(N)}$ | Number of $N$-dimensional candidate vectors evaluated by the stochastic search at time $t$. | Complexity analysis |
| $N_{\text{test}}$ | Number of preliminary calibration replications generated per scenario. | Preliminary calibration |
| $\Delta_L$ | Paired loss difference, $\Delta_L = L(\text{method}) - L(\text{baseline})$. | Comparative analysis |

# Introduction

## Context

This master thesis addresses the prediction of confrontation outcomes in sports, focusing on approaches that rely on a rating assigned to each player or team from the history of past results. The topic is closely related to sports betting and to the ranking systems used by sports federations; from a methodological standpoint, it involves elements of data science, statistical estimation, and optimization.

Rating systems are central tools used within sports federations such as football or chess (see Section 1.3), or in some video games. They process by assigning a score to a player—which is supposed to represent the player's strength—and updating the scores of the two players who have just faced off based on the outcome of the match. The starting point of this thesis is the intuition that, for the most part, these systems operate in a simplistic manner and could be improved.

All along this thesis, the assumption is made that every player/team actually has an intrinsic rating governing its win probability against the other members of a league/tournament. The goal is to estimate these intrinsic ratings as close as possible to the supposed real ones. This work also makes the assumption that the rating of a player/team is not fixed, it can evolve over time if player's abilities fluctuate. [Gli99] [BLSR20].

## Problem

Given a sequence of past match outcomes, we want to predict the result of a future confrontation as accurately as possible. To do so, we consider rating systems that assign a numerical score to each player or team and use a formula involving the two scores to estimate the expected outcome of their match. Designing such a system raises two competing requirements: the rating must be estimated reliably from very limited information (i.e., the cold-start problem [LKH14]), and it must remain accurate over the long run as players' skill evolves. In practice, the system should be reactive when the rating is uncertain or when a genuine change of level occurs, and stable once a player has reached a settled level. The main challenge is to find a trade-off between responsiveness and stability [VHB+26].

Figure 1 shows an illustrative example where we consider a closed system with two players playing against each other 10000 times. The straight lines represent their intrinsic level governing

the win probability distribution via a trivial probability model [Mas22] where

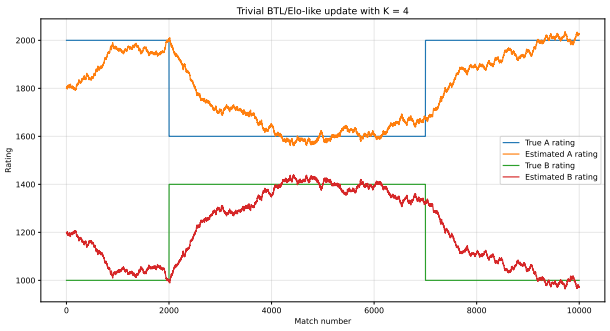
$$P_{ABt} = \mathbb{P}[A \text{ wins against } B \text{ at time } t] = \frac{\text{rating}(A)_t}{\text{rating}(A)_t + \text{rating}(B)_t}.$$

After a confrontation, both ratings are updated proportionally to a factor $K$ called the learning rate or update factor, which will be central in this thesis. More precisely :

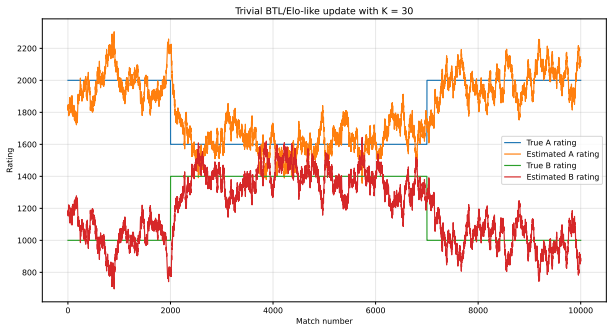
$$\text{rating(A)}_{t+1} = \text{rating(A)}_{\text{t}} + \text{K}(\text{Y}_{\text{ABt}} - \text{P}_{\text{ABt}}),$$

$$\text{rating(B)}_{t+1} = \text{rating(B)}_{\text{t}} - \text{K}(\text{Y}_{\text{ABt}} - \text{P}_{\text{ABt}}).$$

The value of $Y_{ABt}$ is the observed result of the match : $Y_{ABt} = 0$ if $A$ loses and $Y_{ABt} = 1$ if $A$ wins. On the following illustrative example, the orange and red curves represent the estimated rating, trying to fit the true rating curves.




(a) Small $K$ : low responsiveness - high stability.




(b) Big $K$ : high responsiveness - low stability.

Figure 1: Problem illustration.

In Figure 1a, we can observe a slow convergence after a change of rating has occurred, even though the ratings are stable after convergence. In contrast to that, Figure 1b shows a fast convergence after the sudden change of intrinsic rating but an excessive tendency to oscillate afterwards. The aim of this thesis is to allow the $K$ parameter to vary throughout the season to try to improve the predictability of the future matches.

# Background

Many popular rating systems, in particular the Elo Rating System presented in Section 1.2, update the rating of a player after each match by an amount proportional to the difference between the observed outcome and its predicted value, scaled by a parameter usually denoted $K$. This parameter directly controls how much a single observation can update a rating, and therefore how the system trades off responsiveness against stability. In most existing systems, $K$ is fixed or set according to simple empirical rules (see Section 1.3), independently of how informative or surprising the observed match actually was.

Related work at UCLouvain has considered the same general problem of sports-outcome prediction from pairwise results. In particular, Bastien Massion [Mas22] studied tennis match

prediction through low-rank probability and rating formulations based on the BTL model. The present work instead focuses on the sequential evolution of ratings and, more specifically, on adapting the learning rate of the Elo update.

# Research questions

**How can the predictive quality of a rating system be formally evaluated ?**
In order to design a reliable rating system, it is essential to have an objective way to determine how well a prediction algorithm performs. Part of the work will be about choosing relevant metrics to evaluate the predictive quality of different approaches.
**Can we design an adaptation rule for parameter $K$ that performs better than the usual constant-$K$ rule ?**
As explained above, existing systems typically use a fixed or empirically tuned parameter $K$. The aim of this work is to design a way of adapting $K$ to the available information so that the resulting rating system reaches a higher predictive quality – measured by the metric chosen for the first research question (Section 2.2)– than a system with a fixed $K$. Proposed methods are discussed in Section 3.

# Objectives

Formally, the goal is to find an adaptation rule $K(\cdot)$ for the update parameter that minimize a prediction error over the set $D_t$ of all observed matches at time $t$:

$$\min_{K(\cdot)} \sum_{t=1}^{T} \text{Error}\Big(\text{Prediction}_t, \text{Observation}_t\Big).$$

This objective ties together both research questions stated above: the choice of $\text{Error}(\cdot,\cdot)$ answers the first question (Section 2.2) and the optimization over $K(\cdot)$ answers the second one (Section 3).

The underlying objective is of course to approach as close as possible the supposed intrinsic ratings via this error minimization.

# Methodology

Once the metrics have been defined, candidate adaptation rules will be built in Section 3 following two main approaches :

- **Numerical solution :** In order to find an optimal sequence of values for $K$, we will try to design optimization problems that will lead to numerical candidate solutions to our main prediction problem.

- **Analytical solution :** We will also investigate adaptation rules under analytical forms that just take the observations $Y_{ijt}$ as inputs to determine its behavior all along the rating process.

Those candidates will be tested over simulated data according to a very precise protocol detailed in Section 4. A comparison of the methods' performance will then be conducted using a dataset of basketball games with a league format.

## Scopes and limitations

The empirical study in this work is restricted to one-on-one confrontations: a match involves exactly two players or teams, and only their two ratings are updated as a result. We are focusing on confrontations with two possible outcomes: win or lose. Draws are not considered in our framework.

Considering the probability of draws to occur or extending the approach to settings where a single observation affects more than two ratings, such as multiplayer sports, is left as a direction for future work.

## Thesis outline

This thesis is organized as follows. Chapter 1 reviews the state of the art on rating systems, especially the BTL model and the Elo Rating System that will constitute the baseline all along the experiments. Chapter 2 contains the performance metrics design, the establishment of boundaries and a description of the few methods that will constitute our benchmark to assess performances. Chapter 3 forms the core of the thesis : it describes the building of all candidate methods both intuitively and mathematically. The experimental framework is fully described in Chapter 4 covering the simulated data and the real basketball dataset. All the results are presented in Chapter 5 and Chapter 6 contains the discussion about the observed results to reach a conclusion and answer the research questions.

# Chapter 1

# State of the art

## 1.1 The BTL model

The Bradley-Terry-Luce (BTL) model [BV22] [Wik26a] is a probability model that can predict the outcome of a paired comparison. Given a pair of players $i$ and $j$ drawn from a set of players, it estimates the probability that the pairwise comparison $i > j$ turns out true. The most common way to write this comparison mathematically is :

$$\hat{P}_{ijt} = \mathrm{BTL}(\hat{E}_{it}, \hat{E}_{jt}) := \frac{e^{\hat{E}_{it}}}{e^{\hat{E}_{it}} + e^{\hat{E}_{jt}}} = \frac{1}{1 + e^{\hat{E}_{jt} - \hat{E}_{it}}}$$

Throughout this work, $\hat{P}_{ijt}$ has to be interpreted as "probability that player $i$ wins against player $j$, based on their estimated ratings $\hat{E}_{it}$ and $\hat{E}_{jt}$". It is essential to note that this probability is only governed by the difference $\hat{E}_{jt} - \hat{E}_{it}$. The BTL model has some important properties :

1. **Translation invariant :** The predicted win probability $\hat{P}_{ijt}$ remains unchanged if both $\hat{E}_{it}$ and $\hat{E}_{jt}$ are shifted by the same value, i.e,

   $$\mathrm{BTL}(\hat{E}_{it}+C, \hat{E}_{jt}+C) = \frac{e^{\hat{E}_{it}+C}}{e^{\hat{E}_{it}+C} + e^{\hat{E}_{jt}+C}} = \frac{e^{C}e^{\hat{E}_{it}}}{e^{C}e^{\hat{E}_{it}} + e^{C}e^{\hat{E}_{jt}}} = \frac{e^{\hat{E}_{it}}}{e^{\hat{E}_{it}} + e^{\hat{E}_{jt}}} = \mathrm{BTL}(\hat{E}_{it}, \hat{E}_{jt}),$$

   $\forall C \in \mathbb{R}$. Indeed, as $\hat{P}_{ijt}$ only depends on $\Delta\hat{E} = \hat{E}_{jt} - \hat{E}_{it}$, we can easily verify that $\hat{E}_{jt} + C - (\hat{E}_{it} + C) = \hat{E}_{jt} - \hat{E}_{it}$.

2. **Transitivity :** The BTL formula ensures that if $\hat{E}_{it} > \hat{E}_{jt}$ and $\hat{E}_{jt} > \hat{E}_{kt}$, then $\hat{P}_{ikt} > \frac{1}{2}$. Suppose that

   $$\hat{E}_{it} > \hat{E}_{jt} > \hat{E}_{kt}.$$

   Then

   $$\hat{E}_{it} - \hat{E}_{kt} > \max\{\hat{E}_{it} - \hat{E}_{jt}, \hat{E}_{jt} - \hat{E}_{kt}\}.$$

   Since the logistic function $x \mapsto (1 + e^{-x})^{-1}$ is strictly increasing, it follows that

   $$\hat{P}_{ikt} > \max\{\hat{P}_{ijt}, \hat{P}_{jkt}\} > \frac{1}{2}.$$

Thus, a player with a higher estimated rating is predicted to defeat every lower-rated player with probability greater than one half, and this probability increases with the rating difference.

The BTL model is not a rating system because it does not come with a rating update rule. However, this model is the starting point for most of the existing rating systems, in particular for the Elo Rating System.

## 1.2 The Elo Rating System

This rating system, designed by Arpad Elo in the 1960s, is widely used in chess and in several other disciplines. It was adopted by the World Chess Federation (FIDE) in 1970. The Elo Rating System relies on a reparameterization of the BTL formula to model win probabilities. More specifically, the exponential formulation uses base 10 instead of base $e$, a historical choice that made manual calculations more convenient through the use of base-10 logistic tables [Elo78, Elo86]. The rating difference is also scaled by a factor of 400, so that a gap of 400 Elo points corresponds to a win probability of approximately 90% for the higher-rated player [Sol22]. The resulting probabilities as a function of the rating gap are presented in Table 1.1. Unlike BTL, which is simply a probabilistic model, Elo is a rating system, so it comes with a rating update rule. The system is designed the following way[1] [GJ99] :

The win probability for player $i$ against player $j$ is given by :

$$\hat{P}_{ijt} = Elo(\hat{E}_{it}, \hat{E}_{jt}) = \frac{1}{1 + 10^{\frac{\hat{E}_{jt} - \hat{E}_{it}}{400}}}$$

Table 1.1: Elo win probability associated to the rating gap between players $i$ and $j$ (with $\Delta = \hat{E}_i - \hat{E}_j$).

| $\hat{P}_{ijt}$ | $\Delta$ | $\hat{P}_{ijt}$ | $\Delta$ | $\hat{P}_{ijt}$ | $\Delta$ |
|---|---|---|---|---|---|
| 0.00 | $-\infty \to -800$ | 0.35 | -107.54 | 0.70 | 147.19 |
| 0.05 | -511.50 | 0.40 | -70.44 | 0.75 | 190.85 |
| 0.10 | -381.70 | 0.45 | -34.86 | 0.80 | 240.82 |
| 0.15 | -301.33 | 0.50 | 0.00 | 0.85 | 301.33 |
| 0.20 | -240.82 | 0.55 | 34.86 | 0.90 | 381.70 |
| 0.25 | -190.85 | 0.60 | 70.44 | 0.95 | 511.50 |
| 0.30 | -147.19 | 0.65 | 107.54 | 1.00 | $+\infty \to 800$ |

The Elo update rule [LG21] is :

$$\hat{E}_{i(t+1)} := \hat{E}_{it} + K \left(Y_{ijt} - \hat{P}_{ijt}\right),$$

[1] In general, we consider a 800 elo gap as a 100% win probability for the higher rated player [Wik26b] (even if it is not necessarily the case in practice).

$$\hat{E}_{j(t+1)} := \hat{E}_{jt} - K\left(Y_{ijt} - \hat{P}_{ijt}\right).$$

We can therefore see that the update of the Elo rating is governed by two quantities. First, the difference $\left(Y_{ijt} - \hat{P}_{ijt}\right)$ that represents how surprising the outcome is, compared to the Elo prediction, and secondly the parameter $K$ which is our matter of interest throughout this work. We can make the reasonable assumption that, in a sport where the win rate is fairly random, all the players will converge toward the same rating. Let's say player $i$ and player $j$ are the two single players of a league of heads and tails game. At each time : $P_{ijt} = P_{jit} = 0.5$ and as $t$ increases, $\mathbb{E}[\hat{E}_{it}] = \mathbb{E}[\hat{E}_{jt}]$.

All along the following, the Elo Rating System will be our baseline to search for improvements. It is motivated by BTL model, supposing that there exists intrinsic ratings $E_{it}$'s that we want to approximate with $\hat{E}_{it}$'s. On simulated data, we will make the choice that the win or lose outcome for a player $i$ against a player $j$ at time $t$ is a Bernoulli trial with probability of success $= P_{ijt} = \text{Elo}(E_{it}, E_{jt})$.

## 1.3 Examples of empirical $K$ adaptation rules

As mentioned in the introduction, in most of existing systems, $K$ is fixed or set according to simple empirical rules. Here are two examples from famous sports federations.

# FIDE (Chess)

This is the rating system used by the FIDE [Chea] :

- $\hat{P}_{ijt} = \text{Elo}(\hat{E}_{it}, \hat{E}_{jt})$
- $\hat{E}_{i(t+1)} = \hat{E}_{it} + K_i(Y_{ijt} - \hat{P}_{ijt})$.
- $\text{K}_i = \begin{cases} 40 & \text{for the first 30 games played by player } i, \\ 20 & \text{if player } i \text{ has played more than 30 games and its Elo rating is } < 2400, \\ 10 & \text{if player } i \text{ has played more than 30 games and its Elo rating is } \geq 2400. \end{cases}$

# FIFA (Football)

The FIFA World Ranking is based on a modified Elo rating system, referred to by FIFA as the *SUM* algorithm [FIF18]. After each match, the rating of each team is updated according to

$$\hat{E}_{i(t+1)} = \hat{E}_{it} + I\left(Y_{ijt} - \hat{P}_{ijt}\right),$$

where the expected result is given by

$$\hat{P}_{ijt} = \frac{1}{1 + 10^{-(\hat{E}_{it} - \hat{E}_{jt})/600}}.$$

The observed result is encoded as

$$Y_{ijt} = \begin{cases} 1 & \text{for a win,} \\ 0.5 & \text{for a draw,} \\ 0 & \text{for a defeat.} \end{cases}$$

The parameter $I$, which plays the role of the $K$-factor in the standard Elo formulation, depends on the importance of the match:

$$I = \begin{cases} 5 & \text{friendly match outside an International Match Calendar window,} \\ 10 & \text{friendly match during an International Match Calendar window,} \\ 15 & \text{Nations League group-stage match,} \\ 25 & \text{Nations League play-off or final match,} \\ 25 & \text{qualification match for a confederation championship or FIFA World Cup,} \\ 35 & \text{confederation final competition match before the quarter-finals,} \\ 40 & \text{confederation final competition match from the quarter-finals onwards,} \\ 50 & \text{FIFA World Cup match before the quarter-finals,} \\ 60 & \text{FIFA World Cup match from the quarter-finals onwards.} \end{cases}$$

Additional rules are applied in specific situations. For matches decided by a penalty shoot-out, the winning team is assigned $Y_{ijt} = 0.75$, while the losing team is assigned $Y_{ijt} = 0.5$. Moreover, during the knock-out rounds of final competitions, a team cannot lose ranking points: whenever

$$Y_{ijt} - \hat{P}_{ijt} < 0,$$

its rating remains unchanged.

Sports use rating systems not only for predictive purposes but also to establish rankings and handle the matchmaking. In the case of chess, depending on the format of a tournament, ratings are used to perform the pairings. In the case of international or continental football tournaments, a team's FIFA ranking influences the draw for its future opponents [FIF18]. Beyond their original use as a ranking system, Elo-type ratings have also been extensively investigated as predictors of sporting outcomes, notably in football [HA10] and tennis [ACDA22].

From these two examples, we can see that the choice of $K$ is rather empirical with fixed values, meaning that the system may not be reactive enough facing a sudden change of level from a player or a team. The aim of this research work is precisely about that issue. We will investigate how an adaptive $K$ can outperform or not those fixed $K$ values in the prediction domain.

One important remark is that chess and football are out of context in our framework because those sports allow draws to occur. They are mentioned here as illustrative examples to show some choices of the $K$ parameter in practice.

## 1.4 The Glicko system

There are already models that take into account the uncertainty about a player's rating, the most famous being the Glicko rating system designed by Mark E. Glickman in 1995 [Gli95] and updated as "Glicko-2" in 2012 [Gli12]. It introduces "ratings deviation" (RD) or, in statistical terminology, a standard deviation, which measures the uncertainty in a rating (high RD's correspond to unreliable ratings). A player's RD decreases when a game has just been played but increases as the time without playing goes by.
This RD is used to build a confidence interval for the rating. Basically, Glickman defines a 95% confidence interval as :

$$CI_{95\%}(\hat{E}_{it}) = [\hat{E}_{it} - 1.96RD; \hat{E}_{it} + 1.96RD].$$

The rating update for a player after an observation depends not only on the two ratings and its RD but also on the opponent's RD. It is important to note that this new evaluation method no longer conserves the total amount of rating points in the players pool. Indeed, if player $i$ with a low $RD_i$ beats player $j$ with a high $RD_j$, then player $i$ will earn fewer points than player $j$ will lose.

This system is used by the most popular online chess platforms. Chess.com uses the Glicko-1 system and Lichess.org uses the Glicko-2 system [lic].

More generally, Glicko belongs to a broader family of rating systems in which uncertainty about a player's latent strength is explicitly taken into account. TrueSkill, for instance, adopts a Bayesian formulation in which each player's skill is represented by a probability distribution rather than by a single deterministic value, so that both the estimated level and the uncertainty surrounding it evolve as new observations become available [HMG06]. A similar probabilistic interpretation can also be given to the Elo system. Ingram [Ing21] develops a Bayesian perspective on Elo and relates its sequential updating mechanism to Kalman-filtering ideas, providing a principled framework for extending the classical model. Szczecinski and Tihon [ST23] further study online rating through a simplified Kalman-filter formulation, showing how several rating approaches can be understood within a common state-estimation perspective. Altogether, these works support the idea that the amount by which a rating should react to a new observation should not remain constant, but may instead depend on the current uncertainty surrounding the estimated skill. This principle provides additional motivation for the adaptive $K$-factor methods investigated in Chapter 3.

# Chapter 2

# Rating framework

## 2.1 Admissibility condition on $K$ and timeline of available information

Let us define the information available before time $t$ :

$$\mathcal{Z}_{t-1} = \{Y_{ij\tau} : (i,j) \in D_\tau,\ \tau < t\}$$

$\mathcal{Z}_{t-1}$ contains all observations before time t but also all the data computed from those observations, i.e, $\hat{E}_{i\tau} \quad \forall i \in \{1,2,\dots,N\}, \forall \tau < t$; $\hat{P}_{ij\tau} \quad \forall i,j \in \{1,2,\dots,N\}, i \neq j, \forall \tau < t$; $K_\tau \quad \forall \tau < t$ are also known. $D_\tau$ denotes the set of all matches played at matchday $\tau$. The admissibility condition is that $K_t$ only depends on $\mathcal{Z}_{t-1}$. The value $K_{t-1}$ will be used to compute the $\hat{E}_{it}$'s . The detailed time rule is shown in Table 2.1. The idea is that we want to have computed the $K_t$ before matchday $t$.

This is a decision motivated by the transparency that a sports federation should demonstrate regarding the potential gain or loss in rating that players can face, before they even play the match. Just as in practice, the rating systems can access the entire set of past data, not only the past ratings. The only constraint being that future results remain unknown.

## 2.2 Metrics

We want to evaluate the quality of a rating system with some tools. Remember that our primary goal is to assign the players a rating that reflects as accurately as possible the win probability distribution among a set of players. To do so, we can consider several approaches :

### 2.2.1 Accuracy

The first approach is to minimize the prediction error i.e.,

$$\min_K \sum_{t=1}^{T} \sum_{(i,j)\in D_t} \mathcal{L}(\hat{Y}_{ijt}, Y_{ijt})$$

| Step | Available information and computations |
|---|---|
| | **Match** $t-1$ |
| Before the match | The ratings $\hat{E}_{i(t-1)}$ and $\hat{E}_{j(t-1)}$ are available. The prediction is $$\hat{P}_{ij(t-1)} = \text{Elo}\left(\hat{E}_{i(t-1)}, \hat{E}_{j(t-1)}\right).$$ |
| After the match | Once $Y_{ij(t-1)}$ is observed, the rating of player $i$ is updated using $K_{i(t-1)}$: $$\hat{E}_{it} = \hat{E}_{i(t-1)} + K_{i(t-1)}\left(Y_{ij(t-1)} - \hat{P}_{ij(t-1)}\right).$$ The new factor $K_{it}$ is then computed. |
| | **Match** $t$ |
| Before the match | The ratings $\hat{E}_{it}$ and $\hat{E}_{jt}$ are available. The prediction is $$\hat{P}_{ijt} = \text{Elo}\left(\hat{E}_{it}, \hat{E}_{jt}\right).$$ |
| After the match | Once $Y_{ijt}$ is observed, the rating of player $i$ is updated using $K_{it}$: $$\hat{E}_{i(t+1)} = \hat{E}_{it} + K_{it}\left(Y_{ijt} - \hat{P}_{ijt}\right).$$ The new factor $K_{i(t+1)}$ is then computed. |
| | **Match** $t+1$ |

Table 2.1: Timeline of the information available and the computations performed around matches $t-1$, $t$, and $t+1$.

where $\hat{Y}_{ijt}$ is the result of a rounding function applied to $\hat{P}_{ij}$ (in this work, this function is chosen to be the basic rounding operator, motivated by the result of [Mas22] to round upwards when win probability is $> \frac{1}{2}$ and round downwards if win probability is $< \frac{1}{2}$). Among the possible loss functions that we can consider, a natural choice is the zero-norm (i.e. counting the number of times when $Y_{ijt} \neq \hat{Y}_{ijt}$) and get a percentage of correct guesses. The argument for considering this approach is contextual : this thesis is about sports outcomes prediction. This means that losing your bet causes the same loss, no matter the margin of the lost bet. (e.g., assuming you place a bet on "team $i$ wins", you will lose your bet either way, whether $i$ loses the game or draws it. Optimizing over this criterion corresponds to minimize the number of wrong predictions. This corresponds to the definition of accuracy. Formally :

$$\text{Accuracy} = 1 - \frac{1}{M} \sum_{t=1}^{T} \sum_{(i,j) \in \mathcal{D}_t} ||\hat{Y}_{ijt} - Y_{ijt}||_0,$$

with $M = \sum_{t=1}^{T} |D_t|$, and the zero-norm operator :

$$||\hat{Y}_{ijt} - Y_{ijt}||_0 = \begin{cases} 0 & \text{if } \hat{Y}_{ijt} = Y_{ijt}, \\ 1 & \text{if } \hat{Y}_{ijt} \neq Y_{ijt}. \end{cases}$$

The main benefit of this metric is that it is easy to interpret : an accuracy of 0.7355 means that the system predicts the correct outcome 73.55% of the time. One weakness of this approach is that the accuracy does not make the distinction between a good prediction that would have been rounded to 1 with a $\hat{P}_{ijt}$ of 0.51 or 0.99. Using only this metric can lead to lack of information, this is why we will use the accuracy as a secondary criterion.

### 2.2.2 MAE

Second approach is to minimize the error between the estimated rating and the real rating of a player i.e. we want to

$$\min_{K} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathcal{L}(E_{it}, \hat{E}_{it}).$$

A pertinent choice is to use the 1-norm, corresponding to the mean absolute error (MAE). Optimizing according to the MAE criterion corresponds to minimize the absolute difference between the estimated rating of every player and their intrinsic[1] rating. Formally :

$$\text{MAE} = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left| E_{it} - \hat{E}_{it} \right|,$$

Unfortunately, we cannot use the MAE under this form in our context, because of the property of the Elo rating system to be invariant under translation. Indeed, as long as the rating system predicts the correct value of the rating gap between two players, it will provide

[1]This approach uses the assumption that an actual rating exist for each team, corresponding to their deterministic probability density function

the right probability $\hat{P}_{ijt}$ even if $\hat{E}_i$ and $\hat{E}_j$ are both far from the actual values of $E_i$ and $E_j$. Let us define a centred version :

$$\mathrm{MAE}_{\mathrm{centred}}(t) = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{E}_{it} - E_{it} - m_t\right|, \qquad m_t \in \mathbb{R}.$$

The optimal shift $m_t$ that ensures the centering is given by a median of the individual errors

$$d_{it} = \hat{E}_{it} - E_{it}.$$

$$\mathrm{MAE}_{\mathrm{centred}}(t) = \frac{1}{N}\sum_{i=1}^{N}\left|d_{it} - \mathrm{median}(d_{\cdot t})\right|.$$

[2] Let us now impose :

$$\mathrm{MAE} = \mathrm{MAE}_{\mathrm{centred}} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{MAE}_{\mathrm{centred}}(t)$$

To formally interpret that metric, observing a MAE value of 130 means that a rating is estimated with an error of 130 Elo points from the intrinsic rating, on average.

Actually, this approach would not be applicable in practical situations because such intrinsic ratings would be unknown. However, it will be useful to test some $K$ adaptation methods over simulated data, where we know the actual intrinsic rating of the players governing the win probability distributions. This will be used to help building a $K$ rule that will be tested over true data.

### 2.2.3 Brier Score

Third approach is to measure how close we get from the actual probability distribution. Let us proceed by trying to minimize a loss function involving the predicted probability and the actual outcome i.e.,

$$\min_K \sum_{t=1}^{T}\sum_{(i,j)\in D_t}\mathcal{L}(Y_{ijt}, \hat{P}_{ijt}).$$

The idea is to minimize "how surprised we are by the observations". Minimizing the squared difference between the predicted probability and the actual outcome corresponds to the definition of the Brier Score [Bri50] :

$$BS = \frac{1}{M}\sum_{t=1}^{T} BS_t, \quad \text{where} \quad BS_t = \sum_{(i,j)\in D_t}(Y_{ijt} - \hat{P}_{ijt})^2$$

This metric is harder to interpret than the two others. To provide a sense of scale : a Brier score of 0.01 means that $|Y_{ijt} - \hat{P}_{ijt}| = 0.1$ because the difference is squared. $BS = 0.25$ means that $|Y_{ijt} - \hat{P}_{ijt}| = 0.5$ on average. The strength of this metric is that it only relies on

[2] $\mathrm{median}(d_{\cdot t}) = \mathrm{median}(d_{1t}, d_{2t}, \ldots, d_{Nt})$

the estimated probabilities and observations, meaning that the intrinsic ratings do not have to be known to compute this metric. This will be used as the main metric to reach possible conclusions [GR07].

## 2.3 Bounds over the metrics

In order to assess performance, in addition to the metrics, we want to have bounds on what we can expect to reach as a peak performance from a $K$ update rule.

### 2.3.1 Lower bound on accuracy

Mathematically, a suitable lower bound on the accuracy is to choose 0.5, corresponding to a naive predictor always giving a random binary prediction. Considering that we want to outperform the classical Elo constant-$K$ rule, we could use the accuracy achieved by the constant $K$ as a minimal requirement for new adaptation rules.

### 2.3.2 Upper bound on accuracy

Of course, the 100% accuracy is the highest value for an upper bound, corresponding to never failing a prediction. In fact, 100% accuracy would mean that one could predict every outcome, without probabilistic aspect. But we want to have a more realistic peak performance that we can aim to reach. To do so, this upper bound will be the accuracy that "we would have had if we knew the intrinsic ratings before betting". Note that we cannot have such a bound on real data because we cannot be sure if this intrinsic rating governing the win probability distribution does exist or not. This is why such an upper bound will be used on simulated data to facilitate the design of the methods of $K$ parameter adaptation. Under the assumed probabilistic model, the optimal prediction in terms of expected accuracy consists in systematically selecting the player whose true winning probability exceeds $1/2$; a formal derivation of this result is given by [Mas22]. Accordingly, the upper benchmark considered here is the accuracy achieved by an Oracle method that knows the intrinsic ratings before each match. (cf section 2.4.3).

### 2.3.3 Bounds on MAE

The theoretical lower bound of the MAE is zero. This value is achieved by the Oracle predictor, since it has access to the true ratings $E_i$. As the MAE has no finite theoretical upper bound, we use the MAE obtained with the classical constant-$K$ method as a practical reference upper bound, under the assumption that the proposed adaptive methods should aim to improve upon this baseline.

### 2.3.4 Bounds on Brier Score

Again, the best reachable theoretical Brier Score is zero but even the Oracle predictor will have a strictly positive Brier Score because the produced $\hat{P}_{ijt} = P_{ijt}$ will not always be equal to 0 or 1, which will produce nonzero error. The value of the Brier Score achieved by the Oracle method will correspond to our lower bound on the Brier Score metric. For each observation, the squared Brier loss is bounded between 0 and 1. Consequently, since $BS_t$ is the sum of the squared errors over all matches played at time $t$, it satisfies $0 \leq BS_t \leq |D_t|$, it follows that $0 \leq BS \leq 1$. Therefore, 1 constitutes the theoretical upper bound of the overall Brier Score.

## 2.4 Baseline Methods

The proposed adaptive methods will be compared with three baseline procedures in order to get comparative evaluations on their ability to perform according to the metrics defined in section 2.2.

### 2.4.1 Constant $K$.

The first and main baseline is the standard Elo model with a fixed $K$ parameter :

$$K_t = K_{\text{const}}.$$

The same value is therefore used for every player and at every matchday. In the experiments, $K_{\text{const}}$ is calibrated on independent simulations by selecting the value that performs best in terms of Brier score (more in Chapter 4). This baseline represents a conventional, non-adaptive Elo system and determines whether the additional complexity of an adaptive rule leads to a genuine improvement.

### 2.4.2 Square-root decreasing $K$.

The second baseline uses an observation-independent adaptation parameter :

$$K_t = \frac{K_0}{\sqrt{1+t}}.$$

Here $t$ denotes the number of matches previously played by each player. The learning rate is initially large, allowing ratings to adjust rapidly, and then progressively decreases as more information becomes available. This baseline evaluates whether improvements can be obtained from a simple deterministic decrease in $K$, without using prediction errors or any other data-dependent adaptation mechanism.

### 2.4.3 Oracle ratings

The oracle baseline computes match probabilities directly from the true intrinsic ratings used to generate the simulated outcomes:

$$\hat{P}_{ijt}^{\text{oracle}} = P_{ijt} = \frac{1}{1 + 10^{-(E_{it} - E_{jt})/400}}.$$

It does not estimate or update ratings and is therefore unavailable in practice. Its purpose is to provide an ideal informational benchmark, corresponding to predictions made with perfect knowledge of the underlying strengths. This is the Oracle method used to evaluate our bounds on the metrics.

# Chapter 3

# Adaptive $K$-Factor Methods

## 3.1 Numerical Methods

The methods presented in this section all follow the pattern of Algorithm 1. They are all split into three sub-methods : for shared parameter among all players at time $t$ (denoted $K_t$), for player-specific update parameter at time $t$ (denoted $K_{it}$) and for confrontation-specific update parameter at time $t$ (denoted $K_{ijt}$). The initial $K$-factors are set as

$$K_1 = K_0, \qquad K_{i1} = K_0, \qquad K_{ij1} = K_0,$$

for every team $i$ and every pair $(i, j)$. After the outcomes at time $t$ have been observed, the adaptive procedure computes the factors to be used at time $t + 1$. Consequently,

$$K_{t+1}, \qquad K_{i(t+1)}, \qquad K_{ij(t+1)}$$

may depend only on the information contained in $M_t$. The corresponding hypothetical probabilities are denoted by $\hat{P}^k_{ij\tau}$, and $\hat{P}^{\mathbf{k},\text{ind}}_{ij\tau}$.

### 3.1.1 Best historical trajectory

This method selects the value of $K$ that would have produced the largest cumulative classification accuracy over the complete historical rating trajectory up to time $t$. Formally, at each matchday $t$, the method selects the value of $K$ that would have lead to the best accuracy on period $\tau \in [1, t]$.

#### Shared factor $K_{t+1}$

For a single common factor, define[1]

$$K_{t+1} \leftarrow \arg \max_{k \in [K_{\min}, K_{\max}]} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \mathbf{1}\left\{\mathbf{1}\left\{\hat{P}^k_{ij\tau} \geq \frac{1}{2}\right\} = Y_{ij\tau}\right\}.$$

[1] $K_{min} = 0$ and $K_{max} = 200$

**Algorithm 1:** Sequential Elo-like prediction procedure with adaptive $K$

**Input:** Number of teams $N$, time horizon $T$, match sets $\mathcal{D}_t$, initial rating $E_0$, initial factor $K_0$, and outcomes $Y_{ijt}$ revealed sequentially

**Output:** Ratings $\hat{E}_{it}$, predicted probabilities $\hat{P}_{ijt}$ predicted outcomes $\hat{Y}_{ijt}$, and factors $K_t$.

**for** $i = 1, \ldots, N$ **do**
  $\hat{E}_{i1} \leftarrow E_0$;
$K_1 \leftarrow K_0$;
**for** $t = 1, \ldots, T$ **do**
  `// Pre-match prediction using the information available at time` $t$
  **foreach** $(i, j) \in \mathcal{D}_t$ **do**

$$\hat{P}_{ijt} \leftarrow \mathrm{Elo}(\hat{E}_{it}, \hat{E}_{jt}) = \frac{1}{1 + 10^{-(\hat{E}_{it} - \hat{E}_{jt})/400}}$$

$$\hat{Y}_{ijt} \leftarrow \mathbf{1}\{\hat{P}_{ijt} \geq 1/2\}$$

  Observe $Y_{ijt}$ for all $(i, j) \in \mathcal{D}_t$;
  `// Initialize the next-period ratings`
  **for** $i = 1, \ldots, N$ **do**
    $\hat{E}_{i(t+1)} \leftarrow \hat{E}_{it}$;
  `// Post-match rating update using the admissible factor` $K_t$
  **foreach** $(i, j) \in \mathcal{D}_t$ **do**

$$\hat{E}_{i(t+1)} \leftarrow \hat{E}_{i(t+1)} + K_t \left(Y_{ijt} - \hat{P}_{ijt}\right)$$

$$\hat{E}_{j(t+1)} \leftarrow \hat{E}_{j(t+1)} + K_t \left(Y_{jit} - \hat{P}_{jit}\right)$$

  `// Compute the factor to be used at time` $t + 1$
  **if** $t < T$ **then**

$$K_{t+1} \leftarrow \text{adaptive-}K \text{ rule based on } \left\{Y_{ij\tau} : (i, j) \in \mathcal{D}_\tau,\ \tau \leq t\right\}.$$

Here, $\hat{P}^k_{ij\tau}$ = hypothetical $\hat{P}_{ij\tau}(\hat{E}_{i\tau}, \hat{E}_{j\tau})$ if we use $k$ all along past matchdays. This is, the probability obtained by replaying the complete historical trajectory with the same factor $k$ used for every rating update.

### Individual factor $K_{i(t+1)}$

For team-specific factors,

$$K_{i(t+1)} \leftarrow \left[\arg \max_{\mathbf{k} \in [K_{\min}, K_{\max}]^N} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \mathbf{1}\left\{\mathbf{1}\left\{\hat{P}^{\mathbf{k},\mathrm{ind}}_{ij\tau} \geq \frac{1}{2}\right\} = Y_{ij\tau}\right\}\right]_i.$$

The probability $\hat{P}^{\mathbf{k},\mathrm{ind}}_{ij\tau} = \mathrm{Elo}\left(\hat{E}_{i\tau}, \hat{E}_{j\tau}\right)$ is obtained by replaying the historical trajectory while updating team $i$ with $k_i$ and team $j$ with $k_j$ all along.

**Confrontation factor $K_{ij(t+1)}$**

For confrontation-specific updates, let[2]

$$\mathbf{k}^{\star}_{i(t+1)} \leftarrow \left[ \arg \max_{\mathbf{k} \in [K_{\min}, K_{\max}]^N} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \mathbf{1}\left\{ \mathbf{1}\left\{ \hat{P}^{\mathbf{k},\text{ind}}_{ij\tau} \geq \frac{1}{2} \right\} = Y_{ij\tau} \right\} \right]_i .$$

The factor assigned to the confrontation between teams $i$ and $j$ is then[3]

$$K_{ij(t+1)} \leftarrow \frac{\mathbf{k}^{\star}_{i(t+1)} + \mathbf{k}^{\star}_{j(t+1)}}{2}.$$

### 3.1.2 Best terminal ratings induced by $K$

This method selects the value of $K$ whose induced terminal ratings best classify all matches observed up to time $t$.

Unlike the previous method, the probability associated with a past match is not the probability that was available when that match occurred. Instead, every past match is evaluated using the terminal ratings obtained after replaying the complete history $M_t$, and the choice of $K$ is just the one that lead to those ratings.

**Shared factor $K_{t+1}$**

For a common factor,

$$K_{t+1} \leftarrow \arg \max_{k \in [K_{\min}, K_{\max}]} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \mathbf{1}\left\{ \mathbf{1}\left\{ \hat{P}^{k}_{ij(t+1)} \geq \frac{1}{2} \right\} = Y_{ij\tau} \right\},$$

Here, $\hat{P}^{k}_{ij(t+1)} = \text{Elo}\left(\hat{E}^{k}_{i(t+1)}, \hat{E}^{k}_{j(t+1)}\right)$ and the ratings $\hat{E}^{k}_{i(t+1)}$ and $\hat{E}^{k}_{j(t+1)}$ are the terminal ratings obtained after replaying all matches in $M_t$ with the common factor $k$.

**Individual factor $K_{i(t+1)}$**

For team-specific factors,

$$K_{i(t+1)} \leftarrow \left[ \arg \max_{\mathbf{k} \in [K_{\min}, K_{\max}]^N} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \mathbf{1}\left\{ \mathbf{1}\left\{ \hat{P}^{\mathbf{k},\text{ind}}_{ij(t+1)} \geq \frac{1}{2} \right\} = Y_{ij\tau} \right\} \right]_i ,$$

Here, $\hat{P}^{\mathbf{k},\text{ind}}_{ij(t+1)} = \text{Elo}\left(\hat{E}^{\mathbf{k},\text{ind}}_{i(t+1)}, \hat{E}^{\mathbf{k},\text{ind}}_{j(t+1)}\right)$ and these terminal ratings are obtained by replaying $M_t$, with team $i$ updated using $k_i$ and team $j$ updated using $k_j$.

[2]This is the same formula as used for Individual factor $K_{i(t+1)}$ right above.

[3]I made the simple choice of using the arithmetic mean to avoid $K_i = 0$ or $K_j = 0$ to be absorbent.

**Confrontation factor $K_{ij(t+1)}$**

Let

$$\mathbf{k}^{\star}_{i(t+1)} \leftarrow \left[ \arg \max_{\mathbf{k} \in [K_{\min}, K_{\max}]^N} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \mathbf{1} \left\{ \mathbf{1} \left\{ \hat{P}^{\mathbf{k},\text{ind}}_{ij(t+1)} \geq \frac{1}{2} \right\} = Y_{ij\tau} \right\} \right]_i ,$$

The factor assigned to the confrontation between teams $i$ and $j$ is

$$K_{ij(t+1)} \leftarrow \frac{\mathbf{k}^{\star}_{i(t+1)} + \mathbf{k}^{\star}_{j(t+1)}}{2}.$$

### 3.1.3 Best Brier score induced by $K$

The previous method selects the value of $K$ that maximizes the classification accuracy induced by the terminal ratings. However, classification accuracy only determines whether the predicted probability lies on the correct side of the threshold $1/2$. The present method instead selects the value of $K$ whose induced terminal ratings minimize the average Brier score over all matches observed up to time $t$.

**Shared factor $K_{t+1}$**

For a common factor,

$$K_{t+1} \leftarrow \arg \min_{k \in [K_{\min}, K_{\max}]} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \left( Y_{ij\tau} - \hat{P}^{k}_{ij(t+1)} \right)^2 ,$$

where

$$\hat{P}^{k}_{ij(t+1)} = \text{Elo} \left( \hat{E}^{k}_{i(t+1)}, \hat{E}^{k}_{j(t+1)} \right).$$

**Individual factor $K_{i(t+1)}$**

For team-specific factors,

$$K_{i(t+1)} \leftarrow \left[ \arg \min_{\mathbf{k} \in [K_{\min}, K_{\max}]^N} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \left( Y_{ij\tau} - \hat{P}^{\mathbf{k},\text{ind}}_{ij(t+1)} \right)^2 \right]_i ,$$

where

$$\hat{P}^{\mathbf{k},\text{ind}}_{ij(t+1)} = \text{Elo} \left( \hat{E}^{\mathbf{k},\text{ind}}_{i(t+1)}, \hat{E}^{\mathbf{k},\text{ind}}_{j(t+1)} \right).$$

**Confrontation factor $K_{ij(t+1)}$**

Let

$$\mathbf{k}^{\star}_{i(t+1)} \leftarrow \left[ \arg \min_{\mathbf{k} \in [K_{\min}, K_{\max}]^N} \frac{1}{|M_t|} \sum_{(i,j,\tau) \in M_t} \left( Y_{ij\tau} - \hat{P}^{\mathbf{k},\text{ind}}_{ij(t+1)} \right)^2 \right]_i .$$

The confrontation factor used for teams $i$ and $j$ is

$$K_{ij(t+1)} \leftarrow \frac{\mathbf{k}^{\star}_{i(t+1)} + \mathbf{k}^{\star}_{j(t+1)}}{2}.$$

### 3.1.4 Search algorithm

Those 9 sub-methods would require a huge amount of computing time with a classical grid search to test all $\mathbf{k} \in [K_{\min} = 0, K_{\max} = 200]^N$. This is why I will use a stochastic grid search that first searches for best points among a small list and then performs a local search around the best candidates found. The detailed process is explained in Section 4.3.3.

## 3.2 Analytical methods

In this section, the proposed methods do not solve an optimization problem, they attempt to find an expression whose behavior is determined using the observations $Y_{ijt}$ as input at time t. $\hat{P}_{ijt}$ is still defined as $\text{Elo}(\hat{E}_{it}, \hat{E}_{jt})$. The signed prediction residual of player $i$ is defined as

$$e_{it} = Y_{ijt} - \hat{P}_{ijt}.$$

For player $j$, the residual is

$$e_{jt} = (1 - Y_{ijt}) - (1 - \hat{P}_{ijt}) = -e_{it}.$$

These residuals are positive when a player outperforms the Elo prediction and negative when a player underperforms it.

For modelling and interpretability reasons, the following methods all compute a player-specific factor $K_{it}$. The cases $K_t$ and $K_{ijt}$ are no more considered for the five methods of this section. At this stage, it is necessary to point out that player-specific $K$ parameters no longer conserve the total amount of Elo points. This issue is addressed by re-centering the complete rating vector at every time step so that

$$\frac{1}{N}\sum_{i=1}^{N} \hat{E}_{it} = 1200, \quad t = 1, \ldots, T.$$

The methods are configured with a set of parameters. The parameter selection process is detailed in Section 4.3.2.

### 3.2.1 Uncertainty-based adaptive $K$

For each player $i$, we maintain an uncertainty level regarding its rating, denoted $U_{it}$. The uncertainty level is updated according to

$$U_{i(t+1)} = (1 - \eta)U_{it} + \eta|Y_{ijt} - \hat{P}_{ijt}|,$$

The value of $U$ is bounded between 0 and 1. At each observation, we consider that the match played by player $i$ at time $t$ reduces the uncertainty about its rating, according to a quantity $(1 - \eta)$. This amount of uncertainty is then replaced by the unpredictability of the outcome via $\eta|Y_{ijt} - \hat{P}_{ijt}|$. The adaptive $K$-factor is then defined as

$$K_{it}^{\text{unc}} = K_{\min} + (K_{\max} - K_{\min})\frac{U_{it}}{U_{it} + c}.$$

The $K$ parameter therefore reaches $K_{\min}$ when $U_{it} = 0$ and increases monotonically with $U_{it}$; since $U_{it} \leq 1$, its largest attainable value is $K_{\min} + (K_{\max} - K_{\min})/(1 + c)$. The assumption is that $U_{it}$ measures how uncertain the model currently is about player $i$'s rating. If recent predictions involving this player have produced large errors, then $U_{it}$ increases and the model uses a larger $K$. The rating therefore becomes more reactive when the player is difficult to predict.

### 3.2.2 Bernoulli-variance adaptive $K$

As the assumption is made that the win or lose outcome for a player $i$ against a player $j$ at time $t$ is a Bernoulli trial with probability of success $= \hat{P}_{ijt}$, the Bernoulli variance of the match outcome is

$$\text{Var}_{Bernoulli}(\hat{P}_{ijt}) = \hat{P}_{ijt}(1 - \hat{P}_{ijt}).$$

This quantity is bounded between 0 and 0.25, and is maximized when $\hat{P}_{ijt} = 0.5$. To normalize this quantity between 0 and 1, we use

$$\text{Normalized-Var}_{Bernoulli}(\hat{P}_{ijt}) = 4\hat{P}_{ijt}(1 - \hat{P}_{ijt}).$$

The Bernoulli-variance adaptive $K$-factor is therefore

$$K_{it}^{\text{Bern}} = K_{\min} + (K_{\max} - K_{\min})4\hat{P}_{ijt}(1 - \hat{P}_{ijt}).$$

The behavior is similar to the Uncertainty-based method : $K_{it}$ reaches its maximum when the Bernoulli variance normalized form is 1, and its minimum when it is $= 0$. This model assumes that the match contains more information when its outcome is uncertain. When $\hat{P}_{ijt} \approx 0.5$, the match is hard to predict and the Bernoulli variance is maximal, so $K$ is large. When $\hat{P}_{ijt}$ is close to 0 or 1, the outcome is almost expected and the update is smaller. The use of the variance is motivated by the functioning of the Glicko system [Gli95] that takes into account the variance of a player's rating as a tool.

### 3.2.3 Persistent residual adaptive $K$

We first define a smoothed signed residual

$$D_{it} = (1 - \eta)D_{i(t-1)} + \eta e_{i(t-1)},$$

where $\eta \in (0, 1)$ controls the memory of the process. The adaptive $K$-factor is then defined as

$$K_{it}^{\text{res}} = K_{\min} + (K_{\max} - K_{\min})\frac{|D_{it}|}{|D_{it}| + c},$$

with $c > 0$.

This model assumes that if a player repeatedly performs better or worse than predicted, then the current rating is probably biased. In that case, the model should increase $K$ in order to correct the rating quicker. If the signed errors cancel out over time, then the deviations are more likely to be noise, and $K$ remains small.

### 3.2.4 Empirical volatility adaptive $K$

Instead of focusing on the direction of the residual, one can focus on the recent variability of the residuals. Let

$$V_{it} = (1-\eta)V_{i(t-1)} + \eta\left(e_{i(t-1)} - D_{i(t-1)}\right)^2,$$

where $D_{it}$ is the smoothed signed residual. The adaptive $K$-factor is

$$K_{i,t}^{\text{vol}} = K_{\min} + (K_{\max} - K_{\min})\frac{V_{it}}{V_{it}+c}.$$

This model assumes that a player whose recent performances are unstable should have a more reactive rating. A high empirical volatility means that the player's level is harder to estimate, so the Elo system should keep a larger learning rate.

### 3.2.5 Detection adaptive $K$

To detect sudden changes in player strength, we introduce two cumulative sums:

$$C_{it}^{+} = \max\left(0, C_{i(t-1)}^{+} + e_{i(t-1)} - \delta\right),$$

$$C_{it}^{-} = \max\left(0, C_{i(t-1)}^{-} - e_{i(t-1)} - \delta\right),$$

where $\delta > 0$ is a tolerance parameter. A change is detected when

$$\max(C_{it}^{+}, C_{it}^{-}) > h,$$

where $h > 0$ is a detection threshold. After a detected change, the $K$-factor is temporarily increased:

$$K_{it}^{\text{det}} = K_{\min} + (K_{\max} - K_{\min})\exp\left(-\frac{n_{it}}{\tau}\right),$$

where $n_{it}$ is the number of matchdays since the last detected change for player $i$, and $\tau > 0$ controls how quickly $K$ returns to its low value. This proposed method is inspired from [Pag54].

This model makes the assumption that, when a sequence of prediction errors points consistently in the same direction, this may indicate a regime change. The model should then temporarily learn faster. This method works similarly to an alarm that pushes $K_{it}$ to $K_{max}$

when it detects a change and smoothly returns to $K_{min}$ as nothing is detected. Before the first detection, $K_{it}^{\text{det}} = K_{\min}$.

## 3.3 Summary of the candidate methods

Table 3.1 summarizes the 14 adaptive methods introduced in this chapter. For the numerical methods, the reported complexity corresponds to one computation of the factor at time $t+1$. Let

$$H_t := |\mathcal{M}_t|$$

denote the number of matches observed up to time $t$. Moreover, let $Q_t^{(1)}$ denote the number of scalar candidate factors evaluated by the stochastic search, and let $Q_t^{(N)}$ denote the number of $N$-dimensional candidate vectors evaluated by the search.

For the analytical methods, the complexity is reported per player and per new observation. Consequently, each analytical rule has a total complexity of $\mathcal{O}(|D_t|)$ over an entire matchday.

Table 3.1: Summary of the numerical and analytical adaptive $K$-factor methods.

| **Method** | **Label** | **Parameters** | **Complexity** |
|---|---|---|---|
| **Numerical optimization methods** | | | |
| Best historical trajectory, shared factor $K_t$ | *Traj* $K_t$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings, detailed in Section 4.3.3 | $\mathcal{O}\left(Q_t^{(1)} H_t\right)$ |
| Best historical trajectory, player-specific factors $K_{it}$ | *Traj* $K_{it}$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(N)} H_t\right)$ |
| Best historical trajectory, confrontation-specific factors $K_{ijt}$ | *Traj* $K_{ijt}$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(N)} H_t\right)$ |
| Best terminal ratings induced by $K$, shared factor $K_t$ | *Term* $K_t$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(1)} H_t\right)$ |
| Best terminal ratings induced by $K$, player-specific factors $K_{it}$ | *Term* $K_{it}$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(N)} H_t\right)$ |
| Best terminal ratings induced by $K$, confrontation-specific factors $K_{ijt}$ | *Term* $K_{ijt}$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(N)} H_t\right)$ |

Continued on next page

Table 3.1 – continued from previous page

| **Method** | **Label** | **Parameters** | **Complexity** |
|---|---|---|---|
| Best Brier score induced by $K$, shared factor $K_t$ | *Brier* $K_t$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(1)} H_t\right)$ |
| Best Brier score induced by $K$, player-specific factors $K_{it}$ | *Brier* $K_{it}$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(N)} H_t\right)$ |
| Best Brier score induced by $K$, confrontation-specific factors $K_{ijt}$ | *Brier* $K_{ijt}$ | $K_0$, $K_{\min}$, $K_{\max}$; stochastic-search settings | $\mathcal{O}\left(Q_t^{(N)} H_t\right)$ |
| **Analytical adaptation methods** | | | |
| Uncertainty-based adaptive $K$ | *Uncertainty* $K$ | $K_{\min}$, $K_{\max}$, $\eta$, $c$; $U_{i1} = 0.10$ | $\mathcal{O}(1)$ |
| Bernoulli-variance adaptive $K$ | *Bernoulli* $K$ | $K_{\min}$, $K_{\max}$ | $\mathcal{O}(1)$ |
| Persistent-residual adaptive $K$ | *Persistent* $K$ | $K_{\min}$, $K_{\max}$, $\eta$, $c$; $D_{i1} = 0$ | $\mathcal{O}(1)$ |
| Empirical-volatility adaptive $K$ | *Volatility* $K$ | $K_{\min}$, $K_{\max}$, $\eta$, $c$; $D_{i1} = 0$, $V_{i1} = 0.02$ | $\mathcal{O}(1)$ |
| Detection-based adaptive $K$ | *Change-point* $K$ | $K_{\min}$, $K_{\max}$, $\delta$, $h$, $\tau$; $C_{i1}^{+} = C_{i1}^{-} = 0$; no change has initially been detected. After a threshold crossing, set $C_{i,t+1}^{+} = C_{i,t+1}^{-} = 0$ and $n_{i,t+1} = 0$ | $\mathcal{O}(1)$ |

# Chapter 4

# Experimental Framework

## 4.1 Simulated league format

The simulated experiments are conducted in closed leagues with an even number of players. At each matchday, every player takes part in exactly one match. This format, called a round-robin tournament, is used in most sports leagues, notably in certain chess tournaments [Cheb]. A single round-robin cycle is generated with the circle method and therefore contains $N - 1$ matchdays and $N/2$ matches per matchday. The same cycle is then repeated, with the order of the two players reversed on alternating cycles. This reversal has no probabilistic effect because no home advantage is used in the simulations, but it avoids systematically placing the same player in the first position of the binary comparison.

The only two parameters that have remained constant so far ($N$ = number of teams in a league and $T$ = the number of matchdays in a season) will be duplicated to figure out if their modification has a significant impact. Two league sizes are considered. The small league contains $N = 4$ players and the large league contains $N = 20$ players. Different league lengths are also considered : for the four players case, a short season is simulated over 14 round-robin cycles, corresponding to 42 matchdays and 84 matches, and a long season is simulated over 26 round-robin cycles, corresponding to 78 matchdays and 156 matches. For the twenty players case, a short season is simulated over 2 round-robin cycles, corresponding to 38 matchdays and 380 matches, and a long season is simulated over $T4$ round-robin cycles, corresponding to 76 matchdays and 760 matches. The two experiments therefore have almost the same temporal length. Those settings are summarized in Table 4.1. Their main difference is the number of opponents and the amount of information that must be distributed across the rating system : a player repeatedly faces only three opponents in the four-player league, whereas a player faces 19 distinct opponents in the twenty-player league. This design isolates the effect of league size and interaction complexity.

The initial estimated rating is 1200[1] for every player. With $E_{it}$ denoting the intrinsic rating of player $i$ at matchday $t$, for $t = 0, \ldots, T$. After constructing a scenario, the complete rating

[1]Let us recall that it could be any value because the only quantity that governs the predicted win probability is the rating gap between two players.

Table 4.1: Implemented simulated-league formats.

| Feature | Four-player league | Twenty-player league |
|---|---|---|
| Number of players | 4 | 20 |
| Round-robin cycles | 14 or 26 | 2 or 4 |
| Matchdays per cycle | 3 | 19 |
| Total matchdays | 42 or 78 | 38 or 76 |
| Matches per matchday | 2 | 10 |
| Total matches per season | 84 or 156 | 380 or 760 |
| Scenarios | 30 | 30 |
| Replications per scenario | 100 | 100 |

vector is re-centred at every time step so that

$$\frac{1}{N}\sum_{i=1}^{N} E_{it} = 1200,$$

and every component is clipped to $[700, 1700]$. This normalization removes arbitrary common-level movements and preserves the relative-strength interpretation of the Elo model.

For a match between players $i$ and $j$ at matchday $t$, the outcome is sampled from a Bernoulli distribution whose success probability is given by the Elo system formula applied to their intrinsic ratings. Each method is evaluated on exactly the same simulated schedule and outcomes, which makes every comparison paired.

The stress test contains 30 explicit scenarios. They cover stationary rating gaps, smooth linear drifts, abrupt shocks occurring at different moments of the season, random walks with several volatility levels, periodic oscillations, and heterogeneous mixed-profile leagues. Each scenario is independently replicated 100 times, producing 3000 simulated seasons for each league size. These profiles are intended to cover both settings in which a stable learning rate should be sufficient and settings in which rapid adaptation may be necessary. The complete definition of every scenario, including the player-level profiles used in the mixed cases, is provided in Appendix A.

## 4.2 Euroleague basketball dataset

The real-data experiment uses a file containing 3832 recorded basketball matches and 33 raw variables covering the period from 3 November 2003 to 13 December 2019 [Gla]. The available variables include the date, home and away team names, the winner, final scores, quarter-by-quarter scores, half-time aggregates, overtime scores, and several score-gap variables.

Only five raw fields are required by the rating experiment: the date (`DATE`), the home and away team identifiers (`HT` and `AT`), and the final home and away scores (`HS` and `AS`). The final score observed between home team $i$ and away team $j$ at matchday $t$ is converted into the binary response :

$$Y_{ijt} = \mathbf{1}\{\texttt{HS} > \texttt{AS}\}.$$

The explicit winner column is discarded because it is exactly recoverable from the final scores. Quarter, half, overtime, and score-gap variables are also discarded because the investigated Elo rating methods predict only the match winner. Using variables observed during or after the match would additionally create information leakage, while using the winning margin would change the modelling problem from binary outcome prediction to margin prediction.

A season is assigned according to an August boundary: matches from August to December belong to the season beginning in that calendar year, whereas matches from January to July belong to the season ending in the preceding year. Since the file does not provide a reliable common matchday identifier, matchdays are reconstructed chronologically. A new matchday is opened whenever a team would otherwise appear twice in the same group.

A season is retained when it contains at least 25 reconstructed matchdays and at least 150 matches. The number of teams is not constrained. 13 seasons satisfy the two completeness conditions : all seasons are kept except the 2003–2004, 2007–2008, and 2008–2009 seasons because they contain fewer than 25 reconstructed matchdays, while 2019–2020 is excluded because the file contains only the beginning of that season.

For every retained season, all teams start from a rating of 1500, and no rating information is transferred from one season to the next. A fixed home advantage of 65 Elo rating points is included in every prediction. In contrast with the balanced simulated leagues, the Euroleague seasons vary in team count, number of matches, and schedule structure. They therefore provide a substantially less controlled robustness test. A detailed description of the raw variables, the preprocessing operations, and the retained-season metadata is given in Appendix B.

## 4.3 Preliminary calibration

Before the main experiment, an independent preliminary sequence, is generated for parameter calibration. This sequence uses the same 30 scenario definitions and the same league formats as the main experiment, but it relies on a distinct seed range. Consequently, no simulated season used for calibration is reused in the final evaluation. For each scenario, $N_{\text{test}} = 10$ independent preliminary seasons are generated, giving 300 calibration seasons for each league size.

The purpose of this separation is to avoid selecting parameters directly on the test replications. The preliminary sequence provides a common and reproducible basis for choosing the constant-$K$ benchmark and the hyperparameters of the analytical rules. All candidate configurations are evaluated on the same preliminary seasons, so their calibration comparisons are paired in the same way as the final comparisons.

### 4.3.1 Constant $K$ parameter

The constant-$K$ benchmark is not selected manually. On every preliminary season, the following candidate set is evaluated:

$$K_{\text{const}} = \{2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60\}.$$

For each candidate, a complete constant-$K$ Elo season is replayed and its Brier score is computed. The candidate attaining the smallest Brier score is retained for that season. When several values are tied up to numerical precision, their arithmetic mean is used. The universal constant used in the main experiment is then the mean of the selected season-specific values over all 30 scenarios and all ten preliminary replications:

$$K^{\star}_{\text{const}} = \frac{1}{30 N_{\text{test}}} \sum_{s=1}^{30} \sum_{r=1}^{N_{\text{test}}} K^{\star}_{s,r}.$$

This procedure gives the baseline access to a broad range of stationary and non-stationary environments instead of favouring a single scenario. It also makes the comparison more demanding for the adaptive methods, since they are compared with a globally calibrated constant rather than with an arbitrary conventional value. The calibration is performed separately for the four-player and twenty-player experiments. In the Euroleague scripts, the $K^{\star}_{\text{const}}$ value retained will be the one computed for the twenty players simulated short league.

### 4.3.2 Choice of parameters for analytical methods

The five analytical methods are calibrated on the same preliminary seasons used for the constant benchmark. The common candidate bounds are

$$(K_{\min}, K_{\max}) \in \{(0,40), (5,40), (10,60), (10,80), (20,60), (20,80), (0,100), (10,120), (20,120)\},$$

and the smoothing parameter is selected from

$$\eta \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.33\}.$$

The remaining grids are method-specific:

- uncertainty-based rule : $c \in \{0.10, 0.20, 0.30, 0.40\}$, giving $9 \times 6 \times 4 = 216$ configurations;
- Bernoulli-variance rule : only $(K_{\min}, K_{\max})$ is calibrated, giving 9 configurations;
- persistent-residual rule : $c \in \{0.02, 0.05, 0.10, 0.20\}$, giving 216 configurations;
- empirical-volatility rule : $c \in \{0.02, 0.05, 0.10, 0.20\}$ , giving 216 configurations;
- change-point rule : $\delta \in \{0.05, 0.10, 0.15\}, h \in \{0.75, 1.25, 2.00\}, \tau \in \{1, 2, 4, 6, 8\}$, giving $9 \times 3 \times 3 \times 5 = 405$ configurations.

The total calibration grid therefore contains 1062 analytical configurations. The initial state values are fixed rather than tuned: $U_{i1} = 0.10$ for the uncertainty state, $D_{i1} = 0$ for the persistent-residual and volatility states, and $V_{i1} = 0.02$ for the initial empirical variance.

For each candidate and scenario, the Brier score, accuracy, and centred MAE are first averaged over the ten preliminary replications. Candidates are then ranked separately within

each scenario according to mean Brier score and mean centred MAE, both in ascending order. Their primary score is the average of these two ranks. This primary rank is averaged across all 30 scenarios, and the mean accuracy rank is used as the secondary criterion. Remaining ties are resolved by the mean Brier score, mean centred MAE, and candidate identifier. The selected configuration is therefore the one that performs consistently across scenario families rather than the one that dominates a single favorable case.

For the Euroleague analytical experiment over the Euroleague dataset, the retained fixed configurations will be the ones retained by the twenty players simulated short league.

### 4.3.3 Stochastic grid search for numerical methods

For the numerical methods, every component of the optimized $K$ vector is restricted to the integer interval

$$K_{\min} = 0, \qquad K_{\max} = 200.$$

The lower bound includes the possibility of temporarily freezing a rating, whereas the upper bound is sufficiently large to represent very aggressive adaptation while preventing unbounded updates. The interval also contains the complete constant-$K$ calibration grid and the conventional range in which the benchmark methods operate.

An exhaustive search is feasible for the scalar structure $K_t$, which contains only 201 possibilities, but it becomes prohibitive for the player-specific and confrontation structures. The corresponding spaces contain $201^4 = 1{,}632{,}240{,}801$ vectors in the four-player case and $201^{20} \approx 1.16 \times 10^{46}$ vectors in the twenty-player case. The numerical scripts therefore use a dimension-aware stochastic integer search.

The search starts from the previously selected vector, several fixed reference vectors based on 0, 10, 20, 30, the calibrated constant $K$, 60, 100, and 200, and structured increasing, decreasing, and low–high vectors in dimensions greater than one. Random vectors are then added over the full interval. When a previous solution is available, additional candidates are sampled in neighbourhoods of radii 10, 25, and 50. After evaluating this initial population, the best candidates are retained as elites. Successive local rounds perturb each elite and test positive and negative coordinate moves with a radius that decreases over the search. Duplicate vectors are cached and evaluated only once.

The exact settings are summarized in Table 4.2.

Table 4.2: Stochastic search parameters for the scalar and vector settings.

| Setting | Scalar search | Four-player vector | Twenty-player vector |
|---|---|---|---|
| Random candidates | 70 | 70 | 35 |
| Number of elites | 8 | 6 | 5 |
| Local rounds | 12 | 15 | 8 |
| Random perturbations per elite | 2 | 2 | 1 |
| Local radius | $60 \to 3$ | $60 \to 5$ | $60 \to 5$ |

The optimized value is recomputed every three matchdays in the four-player experiment and every four matchdays in the twenty-player experiment; between two optimization dates, the previous value is reused. The four-player objective can use up to the previous 80 matchdays, which is effectively the complete history of its 78-matchday season. The twenty-player objective uses a rolling window limited to the previous 40 matchdays. This restriction reduces replay cost and allows the selected $K$ to respond more strongly to recent changes.

For the vector structures, the implemented search evaluates on the order of $10^3$ distinct candidates per optimization rather than the complete Cartesian product. Relative to exhaustive enumeration, this represents a reduction of more than $1.6 \times 10^6$ in candidate count for a four-dimensional vector and approximately $6.7 \times 10^{42}$ for a twenty-dimensional vector. These ratios quantify the reproducible computational saving.

## 4.4 Comparative analysis

Every adaptive method is compared with the constant-$K$, oracle, and square-root-decay baselines on paired observations. In the simulated experiment, a paired difference is computed for each scenario and replication. For losses, the stored differences are

$$\Delta L = L(\text{method}) - L(\text{baseline}),$$

so negative values indicate an improvement. Accuracy is naturally stored as the method accuracy minus the baseline accuracy. In the heatmaps, its sign is reversed only for display purposes, so that negative cells consistently indicate that the method is better for every metric.

Two complementary graphical summaries are used. Delta heatmaps display the mean paired performance difference for every method–scenario combination, while win-rate heatmaps report the proportion of replications in which the method outperforms the baseline. The first representation measures the average magnitude of an improvement or deterioration; the second indicates whether that behaviour is stable across independently generated seasons. Boxplots show the full empirical distribution of paired deltas across the 100 replications and reveal asymmetry, dispersion, and occasional extreme failures that may be hidden by a mean value. Unless all-scenario plotting is explicitly activated, representative stationary, linear, shock, random-walk, and oscillating scenarios are used for these boxplots.

The main simulated-data comparisons use the Brier score, cumulative accuracy, and centred MAE. On the Euroleague data, only the Brier score and accuracy are available because the intrinsic team ratings are unobserved and a rating error such as centred MAE cannot be computed. Each retained season acts as one paired real-data replication. The real-data heatmaps place seasons on the horizontal axis, and the boxplots summarize the distribution of season-level differences. Means, standard deviations, win rates, and normal-approximation 95% confidence intervals are saved with the graphical outputs.

## 4.5 Exploratory analysis Procedure

Aggregate rankings indicate whether a method performs well, but they do not explain how its $K$ values produce that result. A second diagnostic stage therefore identifies, for each adaptive method, one favourable and one unfavourable scenario relative to the constant-$K$ benchmark. Scenarios are ranked primarily by the average of their Brier-delta rank and centred-MAE-delta rank. Accuracy is used as a secondary ranking criterion. The favourable case minimizes the two loss deltas and favours a positive accuracy difference, whereas the unfavourable case reverses these preferences.

For every selected scenario, player-level trajectories are averaged pointwise over all 100 replications. Each diagnostic graph displays the mean intrinsic rating, the mean rating estimated by the calibrated constant-$K$ method, the mean rating estimated by the adaptive method, the constant benchmark value, and the mean adaptive $K$ assigned to the player. Ratings are plotted on the left axis and $K$ values on the right axis. The four-player and twenty-player scripts inspect all players in their respective leagues.

These graphs put several mechanisms under the spotlight: the delay with which an estimated rating follows a drift or shock, the extent to which a large $K$ produces useful responsiveness or excessive noise, the persistence of an adaptation after the underlying change has disappeared, and the differences between players following heterogeneous profiles. Inspecting both a best and a worst case prevents the interpretation from being based only on favourable examples and helps connect the aggregate metric differences to the actual evolution of ratings and learning rates.

## 4.6 Expectations

Before observing the results, the oracle is expected to provide the strongest benchmark in the simulations because it predicts directly from the intrinsic ratings. The calibrated constant-$K$ rule should be competitive on average because it is selected across all scenario families, but it cannot adapt its responsiveness to the current level of non-stationarity. The square-root rule should be stable in stationary environments, yet its decreasing learning rate may make it slow to recover from late shocks or sustained drifts. Among large panel of scenarios, we can expect that the square-root rule underperforms the constant one.

The analytical methods are expected to differ according to the type of change they measure. The uncertainty and Bernoulli-variance rules should allocate larger updates to uncertain comparisons, while the persistent-residual and change-point rules should be especially useful after systematic forecast errors or abrupt structural breaks. The empirical-volatility rule should be well suited to random walks and oscillating strengths. These benefits may be offset in stationary scenarios if the method interprets ordinary Bernoulli noise as evidence that a larger $K$ is required.

The numerical methods should benefit from directly optimizing an explicit historical objective. The scalar structure $K_t$ is expected to be the most stable and cheapest to optimize. Player-specific values $K_{it}$ and confrontation values $K_{ijt}$ provide more flexibility and should be most

useful in the mixed scenarios, where different players evolve according to different mechanisms. Their larger search dimension may nevertheless increase variability and make the selected vector more sensitive to the finite historical window.

Finally, adaptation should be easier to identify in the four-player league because each pair meets frequently and each player's behaviour is observed repeatedly against the same opponents. The twenty-player league provides a harder estimation problem: information is spread over more opponents, player-specific parameter vectors are larger, and fewer repeated confrontations are available. A method that remains competitive in both settings can therefore be regarded as more robust than one whose gains depend on the highly repetitive structure of the small league. All Python codes used in this thesis are available in this GitHub repository [rob26].

# Chapter 5

# Results

## 5.1 Baseline methods and associated Bounds

The performances achieved by the baseline methods are reported in Table 5.1. The first observation that comes up is the influence of $N$ and $T$. A smaller number of players seems to tend to an easier prediction frame. Where a bigger season duration seems to slightly improve predictions. Those results first confirm that the calibrated constant-$K$ rule constitutes the most relevant practical benchmark. In all four configurations, it outperforms the square-root decreasing rule simultaneously in MAE, Brier score, and cumulative accuracy. The deterioration caused by the decreasing rule is particularly visible for the rating error: progressively reducing the learning rate prevents the system from remaining sufficiently reactive when the intrinsic strengths evolve during the season. This indicates that this naive method is not sufficient in the non-stationary environments considered here. However, the Shrinking $K$ sqrt method will be evaluated on real data in Section 5.5 for testing purpose.

The comparison with the oracle also provides an indication of the remaining room for improvement. The oracle improves the Brier score over constant $K$ by approximately 9.263% to 12.191% (absolute differences of 0.021 to 0.027), while the corresponding accuracy gap ranges from about 0.049 to 0.067. Consequently, the constant-$K$ system already captures a substantial part of the available predictive information, leaving a relatively limited margin for an adaptive procedure to exploit.

As stated in Section 2.3, we will consider here the performances achieved by Oracle method as upper bounds on the expected performances for experimental methods, while performances achieved by the constant-$K$ rule will be considered as lower bounds, i.e., minimal requirements for the other methods to be considered better predictive methods. As a reminder : a Brier score of 0.01 corresponds to a 0.1 gap between $Y_{ijt}$ and $\hat{P}_{ijt}$ which is squared to 0.01. Regarding accuracy, an improvement of 0.01 means that we predict 1% more often the true outcome. This metric is linear, we therefore cannot expect better than a 4.9% to 6.7% improvement for the number of correct predictions.

Table 5.1: Performance of the baseline methods across the four simulated-league configurations.

(a) $N = 4$ players, $T = 42$ matchdays.

| Method | $\text{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Oracle intrinsic ratings | 0.000000 | 0.193600 | 0.695889 |
| Constant $K$ | 99.836113 | 0.220479 | 0.631639 |
| Shrinking $K$ sqrt | 142.705229 | 0.230324 | 0.621044 |

(b) $N = 4$ players, $T = 78$ matchdays.

| Method | $\text{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Oracle intrinsic ratings | 0.000000 | 0.191478 | 0.699979 |
| Constant $K$ | 82.168738 | 0.212824 | 0.650509 |
| Shrinking $K$ sqrt | 136.462916 | 0.224673 | 0.638243 |

(c) $N = 20$ players, $T = 38$ matchdays.

| Method | $\text{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Oracle intrinsic ratings | 0.000000 | 0.209166 | 0.651305 |
| Constant $K$ | 110.656498 | 0.235766 | 0.583867 |
| Shrinking $K$ sqrt | 139.422978 | 0.241097 | 0.573353 |

(d) $N = 20$ players, $T = 76$ matchdays.

| Method | $\text{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Oracle intrinsic ratings | 0.000000 | 0.207612 | 0.654810 |
| Constant $K$ | 93.233199 | 0.228806 | 0.603521 |
| Shrinking $K$ sqrt | 135.268201 | 0.236744 | 0.591455 |

## 5.2 Numerical methods simulations

The numerical methods are now evaluated against the constant-$K$ baseline over the 30 simulated scenarios. Table 5.2 reports the numerical results of the experiment and Figures 5.1 to 5.3 show the mean paired differences for the Brier score, centred MAE, and accuracy, respectively. Since the signs of the displayed differences are chosen such that negative values systematically correspond to an improvement, blue cells indicate favorable cases for the adaptive rule whereas red cells indicate a deterioration.

### 5.2.1 Results

The aggregated results from Table 5.2 show that none of the 9 numerical adaptive methods outperforms the constant-$K$ baseline simultaneously on the three evaluation criteria. More importantly from a prediction perspective, every numerical method obtains a higher mean Brier score and a lower mean cumulative accuracy than constant $K$ in each of the four league configurations.

The ranking nevertheless differs when MAE is considered. In several configurations, some numerical procedures estimate the intrinsic ratings more closely than constant $K$. For example, in the long four-player experiment, the trajectory methods with $K_t$ and $K_{ijt}$ improve the MAE by 1.031% and 0.388%, respectively (81.32 and 81.85, compared with 82.17 for constant $K$). In the long twenty-player experiment, the Brier-terminal method with $K_t$ improves the MAE by 2.674% (90.74 compared with 93.23 for constant $K$). These results already suggest that improving the reconstruction of the latent ratings does not necessarily imply an improvement in outcome prediction.

Among the numerical procedures, no single parametrization dominates for every metric. The trajectory-based approaches tend to be comparatively favourable for rating reconstruction, whereas the Brier-terminal $K_t$ method often obtains some of the strongest accuracy values within the numerical family. However, these differences remain internal to a family whose predictive performance is, on average, below that of the calibrated constant baseline. More visualisation is available via boxplots in Appendix C.

Table 5.2: Performance of the constant-$K$ baseline and the 9 numerical adaptive $K$-factor methods across the four simulated-league configurations.

(a) $N = 4$ players, $T = 42$ matchdays.

| Method | $\mathbf{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Constant $K$ | 99.836113 | 0.220479 | 0.631639 |
| Trajectory – $K_t$ | **96.802826** | 0.233739 | 0.609024 |
| Trajectory – $K_{it}$ | 105.915074 | 0.234368 | 0.600599 |
| Trajectory – $K_{ijt}$ | **97.295210** | 0.231572 | 0.610357 |
| Terminal – $K_t$ | 105.005127 | 0.233752 | 0.617302 |
| Terminal – $K_{it}$ | 111.219212 | 0.240456 | 0.606988 |
| Terminal – $K_{ijt}$ | 104.899592 | 0.235637 | 0.615663 |
| Brier terminal – $K_t$ | 103.620007 | 0.232102 | 0.620385 |
| Brier terminal – $K_{it}$ | 107.385783 | 0.238795 | 0.611520 |
| Brier terminal – $K_{ijt}$ | 104.477759 | 0.235286 | 0.616599 |

(b) $N = 4$ players, $T = 78$ matchdays.

| Method | $\mathbf{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Constant $K$ | 82.168738 | 0.212824 | 0.650509 |
| Trajectory – $K_t$ | **81.321233** | 0.223976 | 0.632846 |
| Trajectory – $K_{it}$ | 88.542219 | 0.225035 | 0.625100 |
| Trajectory – $K_{ijt}$ | **81.850224** | 0.222637 | 0.633645 |
| Terminal – $K_t$ | 102.311957 | 0.225555 | 0.633870 |
| Terminal – $K_{it}$ | 106.028183 | 0.233304 | 0.621932 |
| Terminal – $K_{ijt}$ | 101.864974 | 0.227303 | 0.632500 |
| Brier terminal – $K_t$ | 101.452292 | 0.223133 | 0.639017 |
| Brier terminal – $K_{it}$ | 101.124205 | 0.230817 | 0.628622 |
| Brier terminal – $K_{ijt}$ | 101.226851 | 0.226922 | 0.633534 |

(c) $N = 20$ players, $T = 38$ matchdays.

| Method | $\mathbf{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Constant $K$ | 110.656498 | 0.235766 | 0.583867 |
| Trajectory – $K_t$ | **110.502156** | 0.246539 | 0.574915 |
| Trajectory – $K_{it}$ | 125.235760 | 0.245652 | 0.567085 |
| Trajectory – $K_{ijt}$ | 114.183777 | 0.243287 | 0.574991 |
| Terminal – $K_t$ | 113.199780 | 0.252602 | 0.572372 |
| Terminal – $K_{it}$ | 117.444765 | 0.250244 | 0.570092 |
| Terminal – $K_{ijt}$ | 111.606545 | 0.248602 | 0.573857 |
| Brier terminal – $K_t$ | **106.269525** | 0.250513 | 0.577182 |
| Brier terminal – $K_{it}$ | **110.228830** | 0.252553 | 0.572754 |
| Brier terminal – $K_{ijt}$ | **108.790118** | 0.252201 | 0.575202 |

(d) $N = 20$ players, $T = 76$ matchdays.

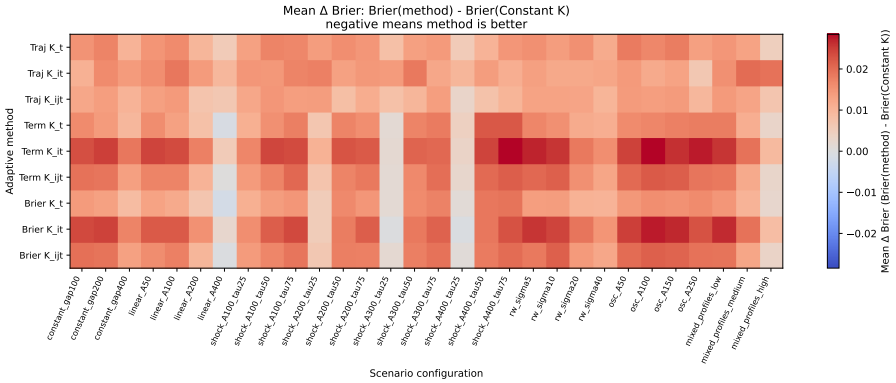
| Method | $\mathbf{MAE}_c$ | Brier | Accuracy |
|---|---|---|---|
| Constant $K$ | 93.233199 | 0.228806 | 0.603521 |
| Trajectory – $K_t$ | **92.340731** | 0.240758 | 0.593933 |
| Trajectory – $K_{it}$ | 105.100788 | 0.239921 | 0.585036 |
| Trajectory – $K_{ijt}$ | 94.730891 | 0.237679 | 0.593650 |
| Terminal – $K_t$ | 99.547382 | 0.242894 | 0.590290 |
| Terminal – $K_{it}$ | 102.982161 | 0.243645 | 0.584889 |
| Terminal – $K_{ijt}$ | 96.694814 | 0.241085 | 0.590691 |
| Brier terminal – $K_t$ | **90.739769** | 0.239032 | 0.597389 |
| Brier terminal – $K_{it}$ | 96.201948 | 0.242146 | 0.590878 |
| Brier terminal – $K_{ijt}$ | 93.772813 | 0.241651 | 0.593359 |

### 5.2.2 Focus on Brier score

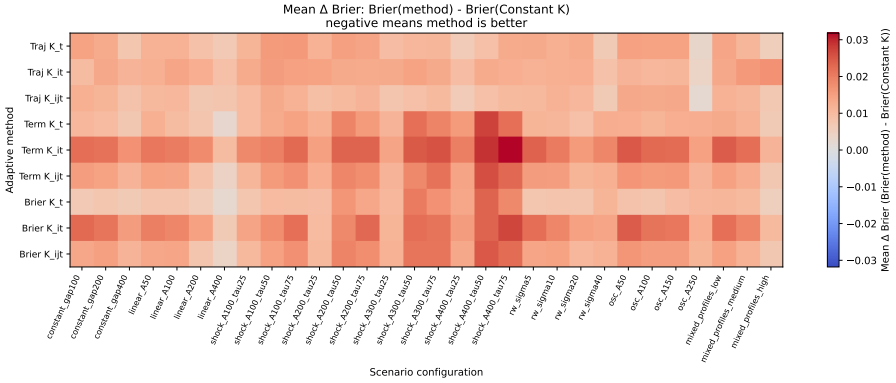
Figure 5.1 provides the clearest negative result for the numerical family. The heatmaps are predominantly positive, meaning that the adaptive procedures usually produce a larger Brier score than the constant-$K$ benchmark. This observation is consistent across both league sizes and both season lengths. The deterioration is particularly visible for several terminal and Brier-terminal variants, while the trajectory methods generally remain closer to the baseline.

A small number of negative cells appear in particular non-stationary scenarios, showing that the numerical adaptation can occasionally be beneficial. These improvements are nevertheless isolated rather than systematic across neighboring scenarios, methods, or league configurations. Especially, scenarios `linear_A400, shock_A200_tau25, shock_A300_tau25` and `shock_A400_tau25` visually stand out as the most favorable across all numerical methods for Brier score criterion. The resulting conclusion at this stage is therefore not that the numerical optimization is incapable of finding useful values of $K$, but rather that the values that optimize the retrospective objective do not generalize reliably to the next observations.

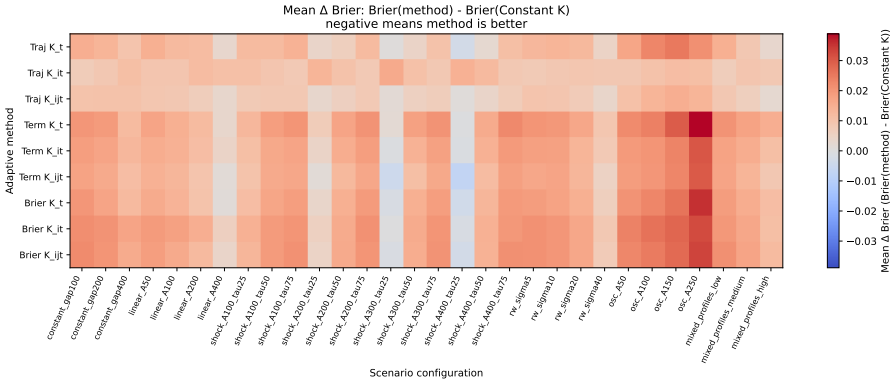
This result is especially important for the Brier-terminal family. Although its optimization criterion is directly related to probabilistic prediction, optimizing historical Brier performance does not guarantee improved future Brier performance. The procedure appears sufficiently flexible to adapt to realized past outcomes, but part of this adaptation is likely to capture realization-specific noise rather than persistent changes in the latent strengths.



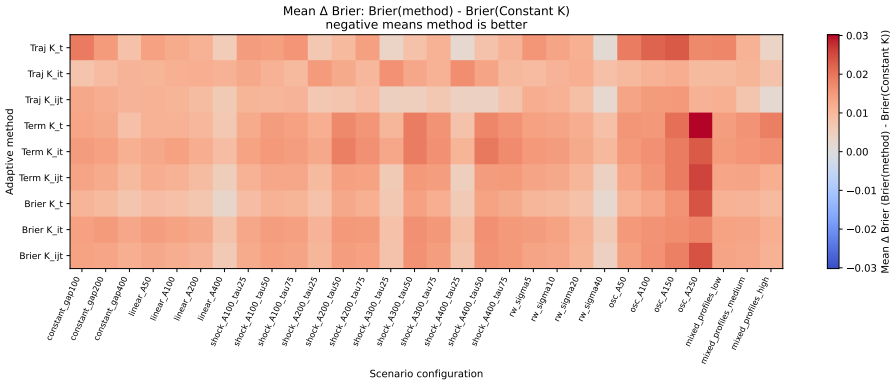

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.




(c) $N = 20$ players, $T = 38$ matchdays.




(d) $N = 20$ players, $T = 76$ matchdays.

Figure 5.1: Brier score performance of the numerical methods relative to the constant-$K$ baseline.

### 5.2.3 Focus on MAE

The centred-MAE heatmaps in Figure 5.2 lead to a more nuanced assessment. Contrary to the Brier score, several clearly negative regions appear, particularly in the four scenarios mentionned in Section 5.2.2, plus `rw_sigma40` and `mixed_profiles_high` that appear slightly favorable as well. The trajectory-based methods benefit most visibly from these situations, sometimes improving by 40 points the mean estimated rating, which is significant. When the intrinsic ratings undergo a sufficiently large displacement, sufficiently early in the season, a dynamically selected $K$ can allow the estimated ratings to catch up more rapidly than under a fixed learning rate.

The improvement is, however, highly scenario-dependent. Stationary and mildly varying environments generally provide little advantage to the constant rule, and several methods show a substantial positive MAE difference. The terminal methods in particular can display large deterioration despite occasionally reacting successfully to abrupt changes. Increasing the league size or the length of the season does not remove this pattern: adaptation can improve tracking in specific regimes, but none of the presented numerical rule provides a uniformly superior estimate of the intrinsic ratings.

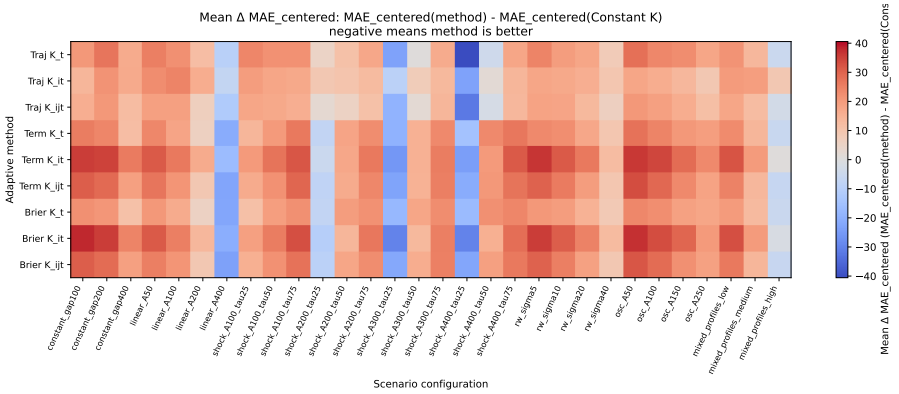
These results illustrate the responsiveness–stability trade-off motivating this work. Large or rapidly changing values of $K$ can be useful immediately after a genuine change in strength, but the same flexibility becomes a drawback when the players profiles are more stable.

### 5.2.4 Focus on accuracy

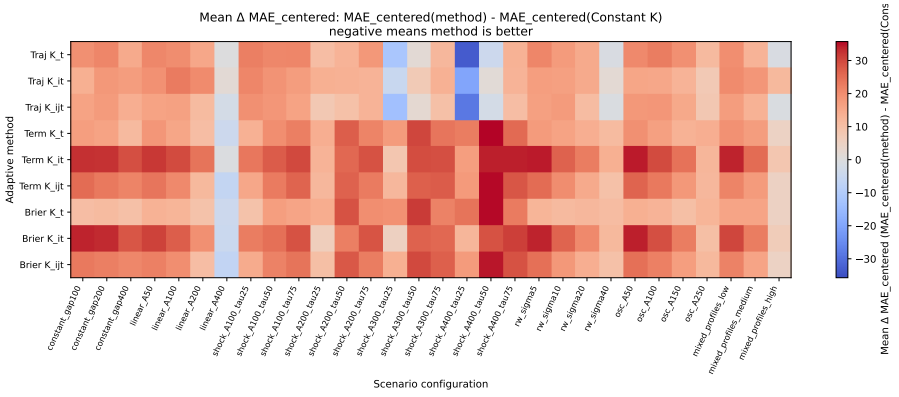
The accuracy results in Figure 5.3 broadly confirm the conclusions drawn from the Brier score. Most method–scenario combinations remain on the unfavorable side of the constant-$K$ benchmark. Only isolated improvements are visible in scenario `osc_A250` for Trajectory methods.

Accuracy is also less sensitive than the Brier score to moderate changes in the predicted probabilities. Two methods can therefore generate noticeably different probability estimates while retaining the same predicted winner, which partly explains why many cells remain close to zero. Nevertheless, the aggregate values in Table 5.2 leave no ambiguity: every numerical adaptive method has a lower overall accuracy than constant $K$.

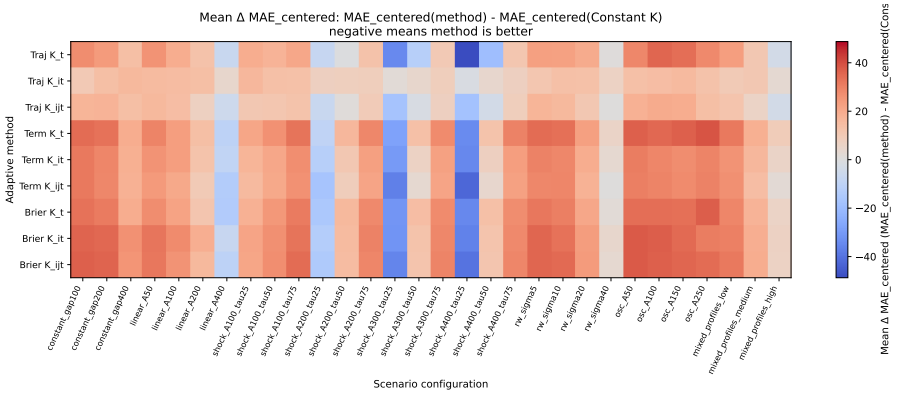
The proposed numerical procedures therefore fail to provide evidence of a robust predictive advantage. Their main positive result concerns their ability to improve intrinsic-rating tracking in selected dynamic environments rather than their ability to consistently predict additional match outcomes correctly.



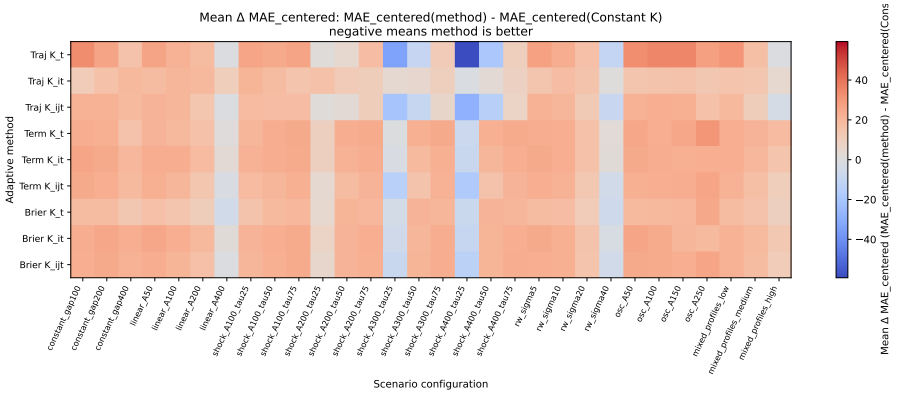

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

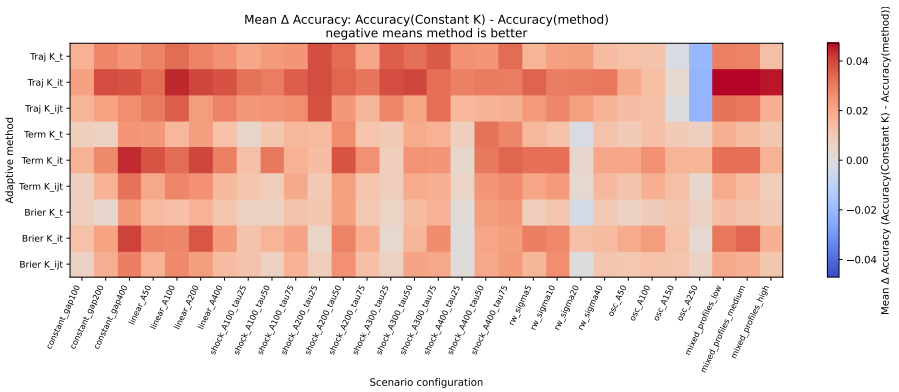



(c) $N = 20$ players, $T = 38$ matchdays.

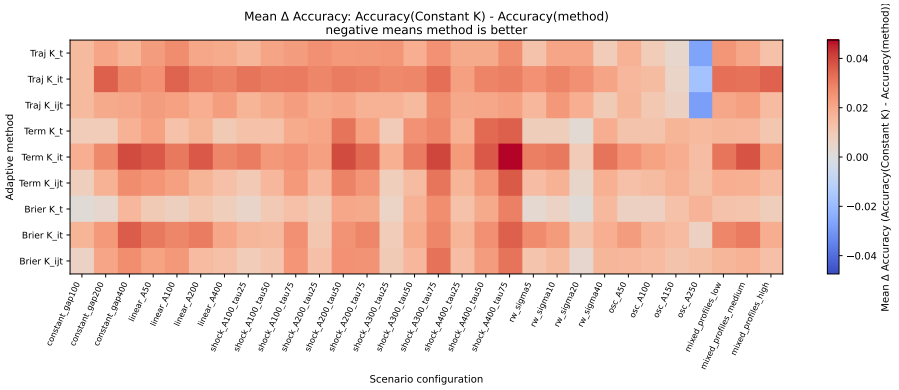



(d) $N = 20$ players, $T = 76$ matchdays.

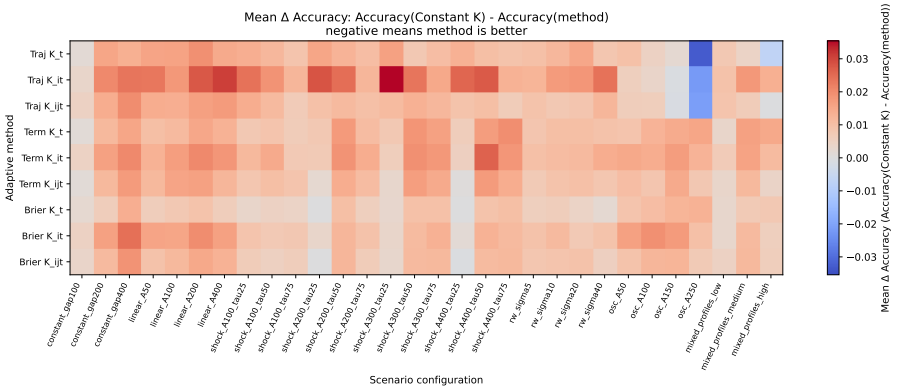
Figure 5.2: MAE performance of the numerical methods relative to the constant-$K$ baseline.



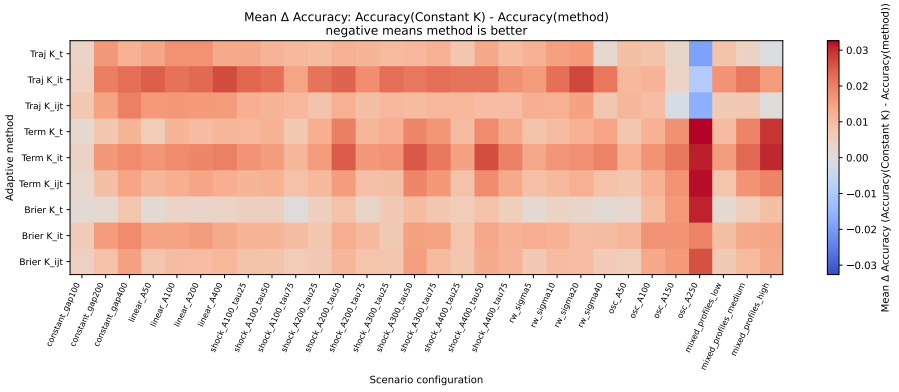

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.




(c) $N = 20$ players, $T = 38$ matchdays.




(d) $N = 20$ players, $T = 76$ matchdays.

Figure 5.3: Accuracy performance of the numerical methods relative to the constant-$K$ baseline.

## 5.3 Analytical methods simulations

The analytical methods produce a substantially different picture. Their adaptation rules are constrained by explicit mechanisms rather than by repeatedly solving a high-dimensional retrospective optimization problem. As a result, their behavior is less flexible but also appears considerably more stable. The retained parameters for each league format are summarized in Table 5.3 and the corresponding numerical results are provided in Table 5.4.

Figures 5.4 to 5.6 show that the differences with constant $K$ are considerably smaller in magnitude than for the numerical methods. At the same time, negative regions occur in structured groups of scenarios, most notably among the abrupt-shock configurations. This indicates that some analytical mechanisms respond to identifiable forms of non-stationarity rather than merely to individual realizations of the simulated outcomes.

### 5.3.1 Parameters retained and results

The selected parameters reported in Table 5.3 exhibit a noticeable degree of stability across the four experimental settings. Small smoothing parameters are generally preferred: $\eta = 0.05$ is retained for most uncertainty, persistent-residual, and empirical-volatility configurations. The volatility coefficient is consistently selected as $c = 0.20$, while the change-point rule always retains $K_{\min} = 20$, $K_{\max} = 60$, and $\tau = 8$. The main exceptions occur in the shortest four-player setting, where less data are available and the persistent-residual rule selects a larger smoothing parameter.

This relative stability is encouraging because the calibration is performed independently for each league configuration. It suggests that the preferred analytical dynamics are not entirely specific to a single season length or league size. In particular, the repeated selection of relatively small smoothing rates indicates that the best-performing rules do not react strongly to a single residual, but accumulate evidence before substantially modifying the update factor.

Table 5.4 also shows that the analytical methods are more competitive with constant $K$ than the numerical family. The uncertainty and empirical-volatility methods repeatedly reduce centred MAE, while small Brier-score and accuracy gains are observed in some configurations. The empirical-volatility rule is particularly competitive in the twenty-player experiments, improving the Brier score by 0.260% in the short season (0.235152 against 0.235766 for constant $K$), and by 0.107% in the long season (0.228562 against 0.228806), with a slightly better accuracy as well. These gains are poorly significant, but unlike the numerical results they demonstrate that adaptive behavior can improve predictive performance under some aggregate simulated settings. To illustrate this, corresponding boxplots are available in Appendix C.

Table 5.3: Selected parameters of the five analytical adaptive $K$-factor methods across the four simulated-league configurations.

(a) $N = 4$ players, $T = 42$ matchdays.

| **Method** | $\mathbf{K}_{\min}$ | $\mathbf{K}_{\max}$ | $\boldsymbol{\eta}$ | **c** | $\boldsymbol{\delta}$ | **h** | $\boldsymbol{\tau}$ |
|---|---|---|---|---|---|---|---|
| Uncertainty-based $K$ | 10 | 80 | 0.05 | 0.30 | – | – | – |
| Bernoulli variance $K$ | 20 | 60 | – | – | – | – | – |
| Persistent residual $K$ | 20 | 60 | 0.20 | 0.10 | – | – | – |
| Empirical volatility $K$ | 20 | 80 | 0.05 | 0.20 | – | – | – |
| Change-point $K$ | 20 | 60 | – | – | 0.15 | 1.25 | 8 |

(b) $N = 4$ players, $T = 78$ matchdays.

| **Method** | $\mathbf{K}_{\min}$ | $\mathbf{K}_{\max}$ | $\boldsymbol{\eta}$ | **c** | $\boldsymbol{\delta}$ | **h** | $\boldsymbol{\tau}$ |
|---|---|---|---|---|---|---|---|
| Uncertainty-based $K$ | 10 | 80 | 0.05 | 0.30 | – | – | – |
| Bernoulli variance $K$ | 5 | 40 | – | – | – | – | – |
| Persistent residual $K$ | 20 | 60 | 0.10 | 0.05 | – | – | – |
| Empirical volatility $K$ | 20 | 80 | 0.05 | 0.20 | – | – | – |
| Change-point $K$ | 20 | 60 | – | – | 0.10 | 2.00 | 8 |

(c) $N = 20$ players, $T = 38$ matchdays.

| **Method** | $\mathbf{K}_{\min}$ | $\mathbf{K}_{\max}$ | $\boldsymbol{\eta}$ | **c** | $\boldsymbol{\delta}$ | **h** | $\boldsymbol{\tau}$ |
|---|---|---|---|---|---|---|---|
| Uncertainty-based $K$ | 0 | 100 | 0.05 | 0.40 | – | – | – |
| Bernoulli variance $K$ | 5 | 40 | – | – | – | – | – |
| Persistent residual $K$ | 20 | 60 | 0.05 | 0.10 | – | – | – |
| Empirical volatility $K$ | 0 | 100 | 0.05 | 0.20 | – | – | – |
| Change-point $K$ | 20 | 60 | – | – | 0.05 | 2.00 | 8 |

(d) $N = 20$ players, $T = 76$ matchdays.

| **Method** | $\mathbf{K}_{\min}$ | $\mathbf{K}_{\max}$ | $\boldsymbol{\eta}$ | **c** | $\boldsymbol{\delta}$ | **h** | $\boldsymbol{\tau}$ |
|---|---|---|---|---|---|---|---|
| Uncertainty-based $K$ | 10 | 80 | 0.05 | 0.40 | – | – | – |
| Bernoulli variance $K$ | 5 | 40 | – | – | – | – | – |
| Persistent residual $K$ | 20 | 60 | 0.05 | 0.10 | – | – | – |
| Empirical volatility $K$ | 10 | 60 | 0.05 | 0.20 | – | – | – |
| Change-point $K$ | 20 | 60 | – | – | 0.10 | 2.00 | 8 |

Table 5.4: Performance of the constant-$K$ baseline and the five analytical adaptive $K$-factor methods across the four simulated-league configurations.

(a) $N = 4$ players, $T = 42$ matchdays.

| **Method** | $\mathbf{MAE}_c$ | **Brier** | **Accuracy** |
|---|---|---|---|
| Constant $K$ | 99.836113 | 0.220479 | 0.631639 |
| Uncertainty-based $K$ | **92.891506** | **0.220137** | **0.631817** |
| Bernoulli variance $K$ | **92.197951** | 0.221635 | **0.631790** |
| Persistent residual $K$ | **96.761992** | 0.220701 | 0.630960 |
| Empirical volatility $K$ | **93.333322** | **0.220160** | **0.631770** |
| Change-point $K$ | 100.423012 | 0.221278 | 0.630115 |

(b) $N = 4$ players, $T = 78$ matchdays.

| **Method** | $\mathbf{MAE}_c$ | **Brier** | **Accuracy** |
|---|---|---|---|
| Constant $K$ | 82.168738 | 0.212824 | 0.650509 |
| Uncertainty-based $K$ | **75.247057** | 0.213323 | 0.649603 |
| Bernoulli variance $K$ | 84.870243 | 0.213332 | 0.650434 |
| Persistent residual $K$ | **76.426412** | 0.213133 | 0.649645 |
| Empirical volatility $K$ | **75.554989** | 0.213375 | 0.649718 |
| Change-point $K$ | 84.535840 | 0.213443 | 0.649536 |

(c) $N = 20$ players, $T = 38$ matchdays.

| **Method** | $\mathbf{MAE}_c$ | **Brier** | **Accuracy** |
|---|---|---|---|
| Constant $K$ | 110.656498 | 0.235766 | 0.583867 |
| Uncertainty-based $K$ | **103.041867** | **0.235277** | **0.585226** |
| Bernoulli variance $K$ | **108.582959** | 0.236293 | **0.583968** |
| Persistent residual $K$ | **108.676843** | **0.235752** | 0.583846 |
| Empirical volatility $K$ | **103.846706** | **0.235152** | **0.585473** |
| Change-point $K$ | **108.442728** | 0.235844 | 0.583264 |

(d) $N = 20$ players, $T = 76$ matchdays.

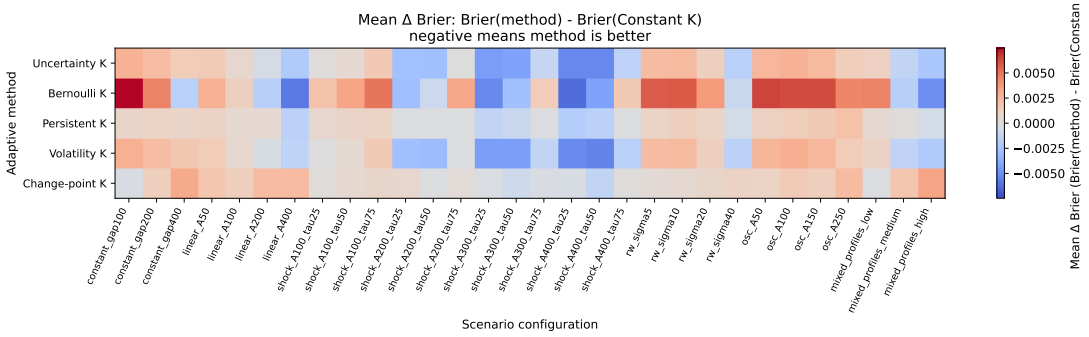
| **Method** | $\mathbf{MAE}_c$ | **Brier** | **Accuracy** |
|---|---|---|---|
| Constant $K$ | 93.233199 | 0.228806 | 0.603521 |
| Uncertainty-based $K$ | **86.562425** | 0.229300 | **0.603716** |
| Bernoulli variance $K$ | **92.694164** | 0.229773 | **0.603614** |
| Persistent residual $K$ | **89.331067** | 0.228838 | 0.603257 |
| Empirical volatility $K$ | **90.177646** | **0.228562** | **0.603736** |
| Change-point $K$ | **92.899964** | 0.229136 | 0.602272 |

### 5.3.2 Focus on Brier score

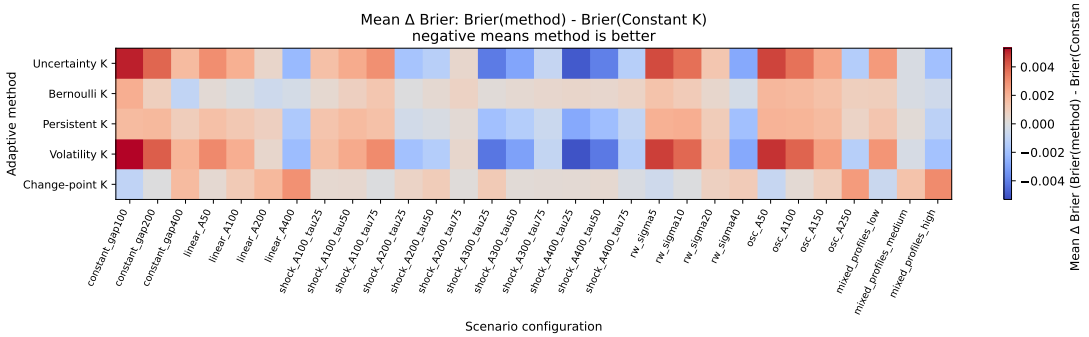
Figure 5.4 reveals a strong dependence on the type of rating dynamics. In the stationary and weakly varying scenarios, most analytical adaptations provide little advantage and can slightly deteriorate the Brier score. In contrast, a substantial set of shock scenarios produces negative differences. The effect becomes particularly visible for the uncertainty-based and empirical-volatility rules, whose mechanisms are naturally able to increase the learning rate when recent observations become inconsistent with the current rating estimates.

The empirical-volatility method displays some of the strongest and most persistent negative regions around medium- and high-amplitude shocks. This is consistent with its construction: an abrupt displacement of intrinsic strength generates a sequence of atypical residuals and an increase in empirical residual variability, which temporarily raises $K$ and accelerates rating adjustment. The uncertainty method displays a similar, although not identical, response.

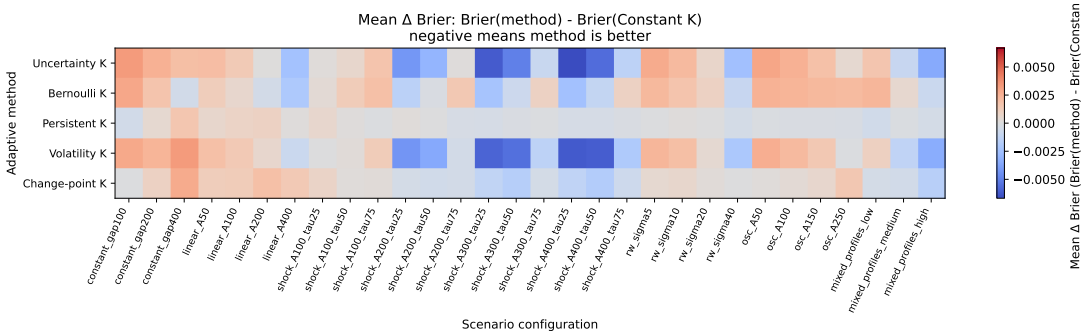
The Bernoulli-variance and change-point rules are generally more conservative. Their mean performance remains close to the baseline across many configurations, which limits both large improvements and large deterioration. Overall, the heatmaps indicate that analytical adaptation is most useful when the data-generating process contains a sufficiently strong and identifiable change, whereas the calibrated constant rule remains difficult to improve upon in comparatively stable environments.



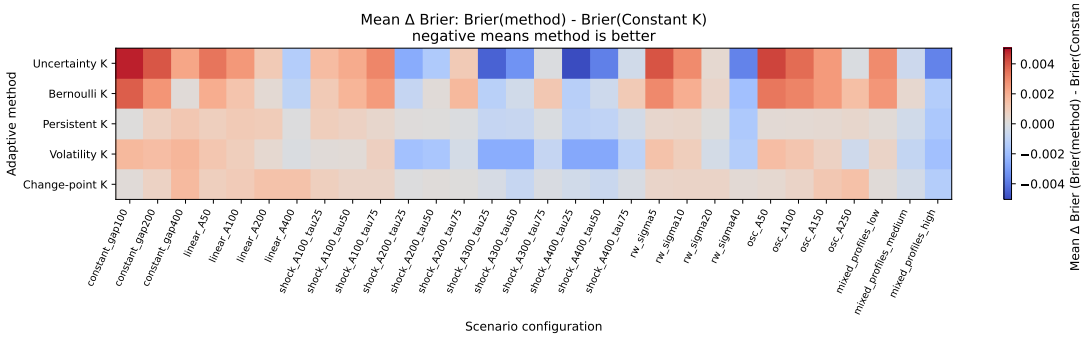

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.




(c) $N = 20$ players, $T = 38$ matchdays.




(d) $N = 20$ players, $T = 76$ matchdays.

Figure 5.4: Brier score performance of the analytical methods relative to the constant-$K$ baseline.

### 5.3.3 Focus on MAE

The MAE results provide the strongest evidence in favor of the analytical methods. Figure 5.5 contains clearly more negative regions than the corresponding numerical prediction heatmaps, with the largest gains again concentrated around abrupt changes in intrinsic strength. The uncertainty, Bernoulli, persistent-residual, and empirical-volatility methods are able to reduce the delay with which the estimated rating follows the new underlying level in most of league formats.

This observation is also reflected in the aggregate values in Table 5.4. For the two 20-player formats, every analytical method obtains a lower MAE than constant $K$, while the uncertainty-based and empirical-volatility rules reduce it significantly in both cases. In the long four-player setting, the uncertainty, persistent-residual, and volatility methods reduce the error by approximately 6.988%–8.424% (from 82.17 to approximately 75–76).

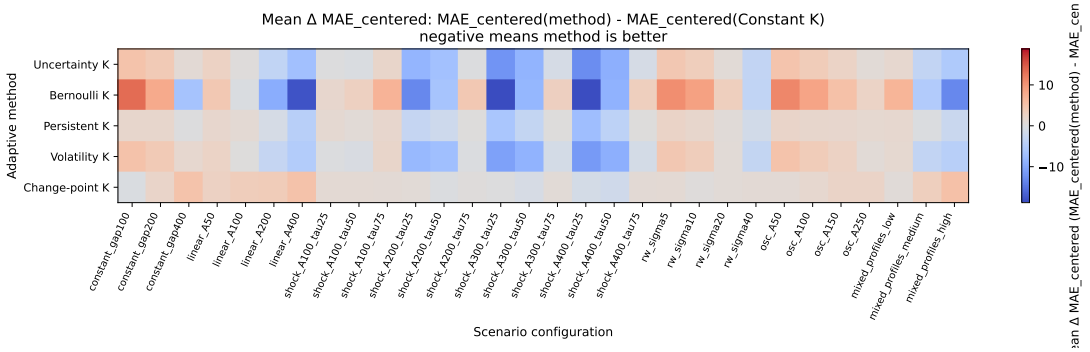
The gain in rating reconstruction is therefore considerably more convincing than the gain in direct predictive performance. Adaptation can help the Elo process follow a moving latent state, but the stochastic nature of individual match outcomes limits how directly this improved reconstruction can be converted into better Brier scores or accuracy.

### 5.3.4 Focus on accuracy

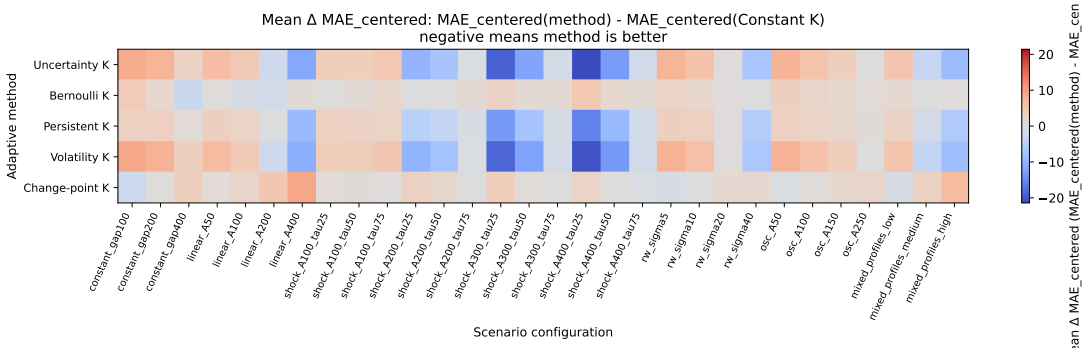
Accuracy differences in Figure 5.6 are generally small, varying by less than 1% most of the time. This is expected because accuracy depends only on which side of 0.5 the estimated probability lies. A substantial improvement in the estimated probability may therefore leave the predicted winner unchanged.

Consequently, the biggest part of the scenario dependence observed for the other metrics has vanished. Several shock and strongly heterogeneous configurations benefit from the more reactive analytical rules, whereas the stationary regimes provide little reason to depart from constant $K$. At the aggregate level, the uncertainty and empirical-volatility methods slightly exceed constant-$K$ accuracy in both twenty-player settings, but this improvement is not significant enough.

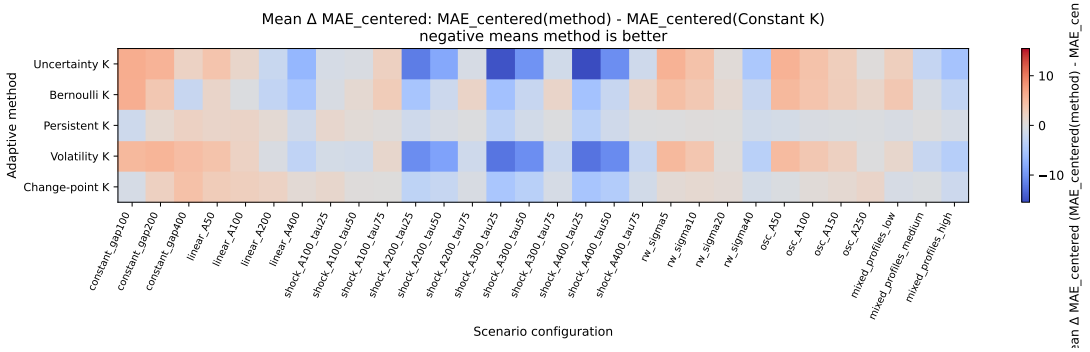
Consequently, the analytical methods should not be interpreted as producing a large gain in the number of correctly classified matches. Their principal advantage is a more accurate and responsive reconstruction of the evolving strengths, accompanied by small probabilistic improvements in some sufficiently non-stationary environments.



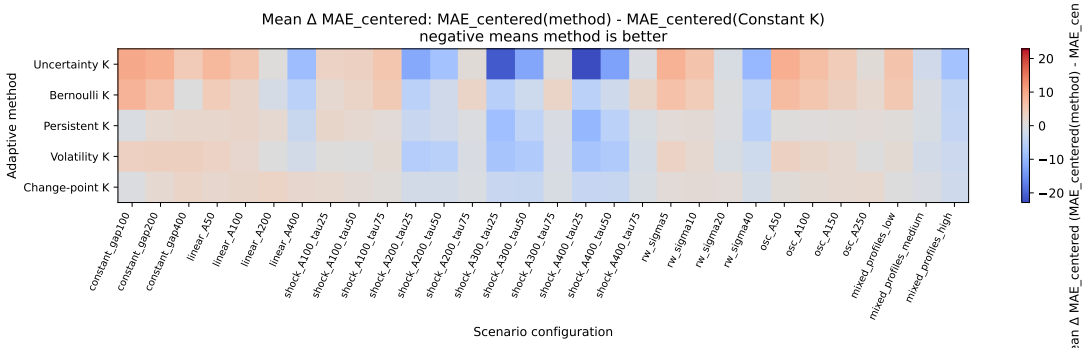

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

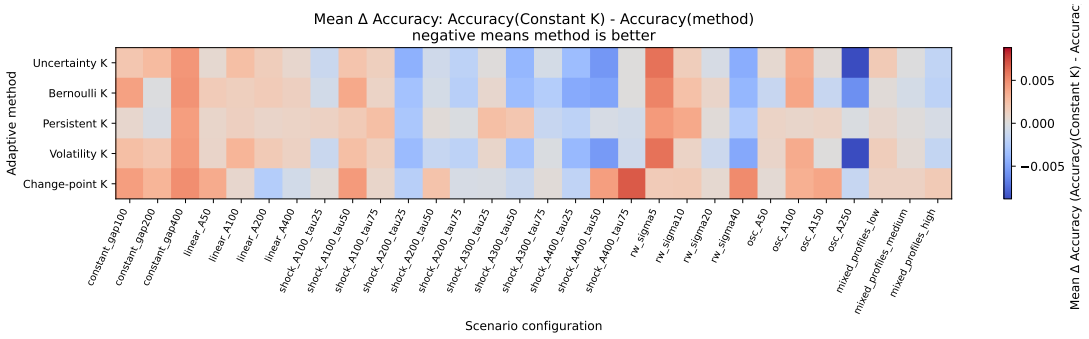



(c) $N = 20$ players, $T = 38$ matchdays.

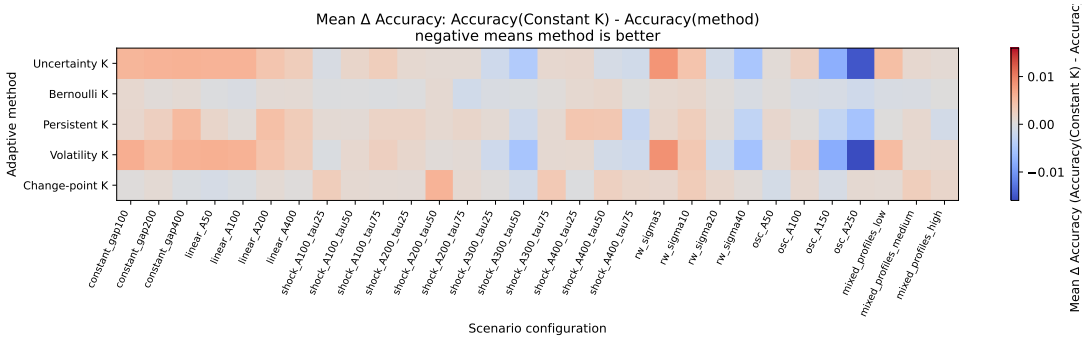



(d) $N = 20$ players, $T = 76$ matchdays.

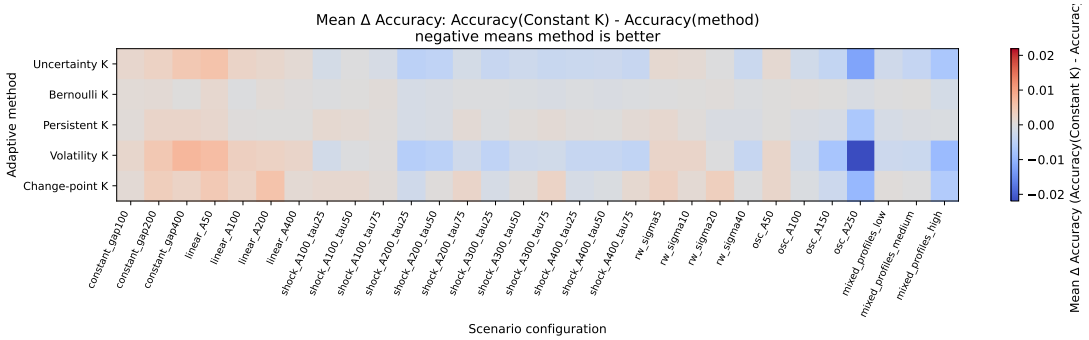
Figure 5.5: MAE performance of the analytical methods relative to the constant-$K$ baseline.



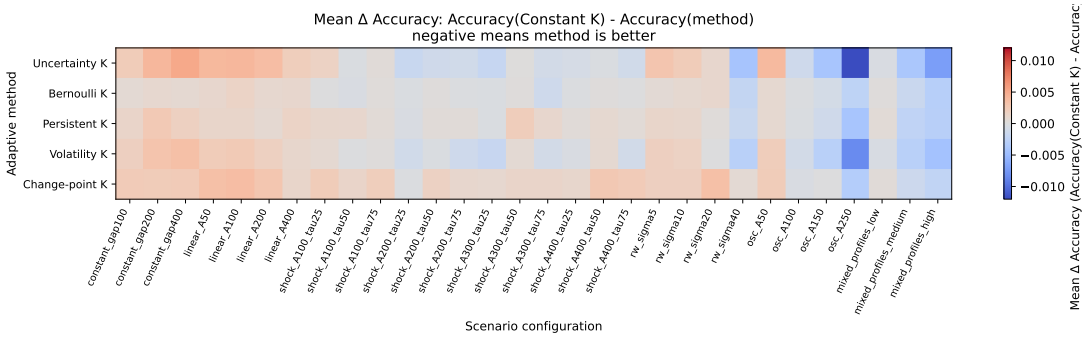

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.




(c) $N = 20$ players, $T = 38$ matchdays.




(d) $N = 20$ players, $T = 76$ matchdays.

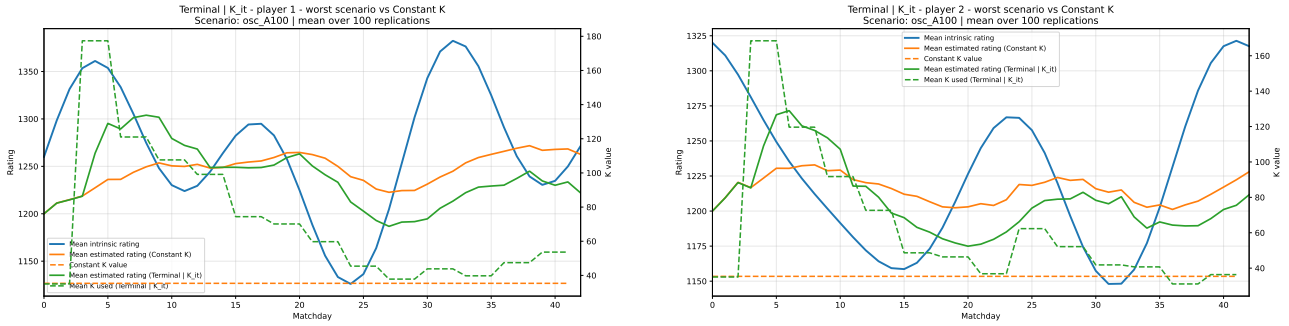
Figure 5.6: Accuracy performance of the analytical methods relative to the constant-$K$ baseline.

## 5.4 In-depth analysis

This section dives into the rating estimation process by looking at the curves for $\hat{E}_{it}$, selecting the best and worst scenarios encountered to see how the methods behave. Figures 5.7 to 5.12 display for a selected method the average over the 100 replications for the true rating of a selected player, its estimated rating by the constant $K$ rule and the considered method, and the $K$ values over time. The goal is to have an eye on how the rating estimation curve evolve in response to a significant $K$ adaptation. This will allow us to confirm or refute our observations regarding the overall results. I have selected a few representative cases to analyze in this section, but the graphs for all simulated players for each scenario and each method are available in [rob26].

### 5.4.1 Numerical methods

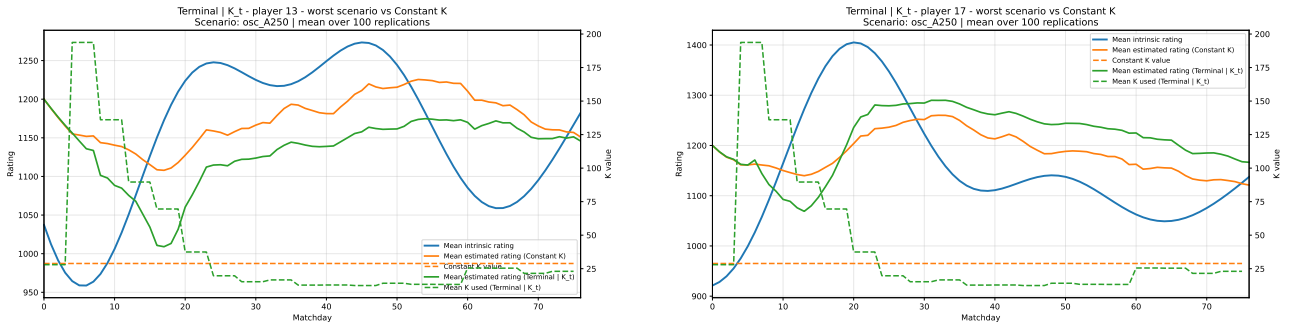
For this family of methods, the oscillatory profiles scenarios obtained the worst results. On Figures 5.7 and 5.8 we can see that the $K$ value has a tendency to take a big value at the beginning of the season and decrease afterwards. This leads to a faster reaction facing the error of the initial prediction $\hat{E}_{i1}$ but weaker ability to detect the long term trajectory of the rating. Those observations can constitute an explanation to the fact that numerical methods lead sometimes to better rating approximation, despite the weaker predictive ability. The constant $K$ method does not commit to deviate too much, which keeps it closer to the mean rating value encountered in oscillatory scenarios, explaining the possible weaker rating approximation but better Brier or accuracy for the predictions.




(a) Player 1 (b) Player 2

Figure 5.7: Worst-case scenario for numerical method `Term_`$K_{it}$ with $N = 4$ and $T = 42$.

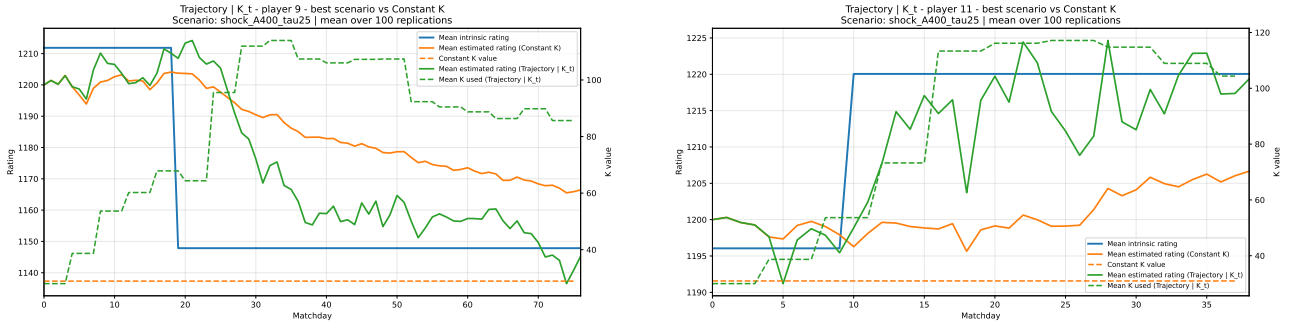
By looking at Figure 5.9, corresponding to the best-case scenario for method Trajectory $K_t$, we can see that big shocks in players' rating constitute easier profiles to estimate by this method. The value of $K$ rapidly increases right after the shock has occurred, leading to a faster convergence to the new stationary rating. However, Figure 5.9b gives the clearest picture about the method's inability to detect that the estimation has approached the true value and should therefore slower the learning. In fact, the $K$ value struggles to decrease after the shock has been detected. This observation is consistent with what is observed for the worst case scenarios.




(a) Player 13 (b) Player 17

Figure 5.8: Worst-case scenario for numerical method `Term_`$K_t$ with $N = 20$ and $T = 76$.

Better MAE result is a consequence of the numerical method ability to react fast, but it often leads to worse predictive quality regarding Brier score and accuracy.



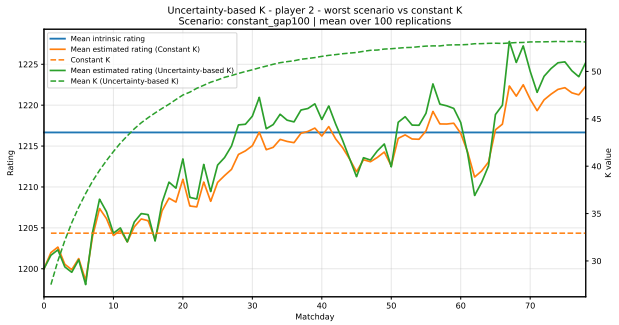

(a) Player 9, $N = 20$ and $T = 76$. (b) Player 11, $N = 20$ and $T = 38$.

Figure 5.9: Best-case scenario for numerical method `Traj_`$K_t$.

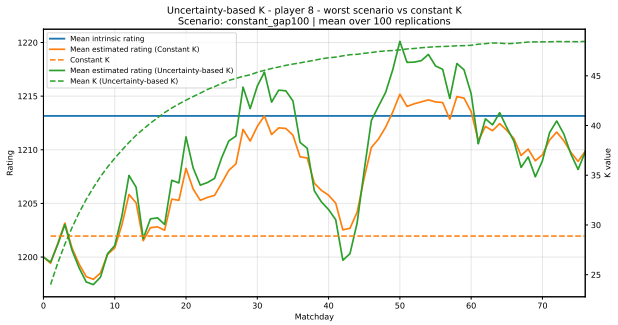
### 5.4.2 Analytical methods

Figure 5.10 shows the worst case scenario encountered by the Uncertainty-based $K$ method. One thing to notify here is that the average $K$ value over all replications tends to grow proportionally to a square root-like shape. This leads to a more and more reactive behavior as the season goes on. The main direct consequence is a faster ability to adjust the initial prediction error $\hat{E}_{i1}$ but an undesirable oscillatory behavior in a long run. This gives evidence why the constant rating profiles are more likely to be wrongly estimated by this adaptive method.

Looking at the best-case scenarios encountered by the Empirical volatility $K$ method shown in Figures 5.11 and 5.12, we already see an improvement compared to the Uncertainty $K$ method worst scenario : The $K$ curve has now a concave shape with maximum value right after a shock is detected. This allows the method to progressively decrease its learning rate after the adaptation has been applied. Figures 5.11b and 5.12b appear to be particularly favorable cases because the initial error correction is directly followed by the shock. The responsiveness of the method then easily outperforms the constant $K$ to reach the true rating as fast as possible. As a conclusion, at this stage, we have observed that both numerical and analytical adaptive methods share a tendency to provide a larger relative advantage on unstable profiles.

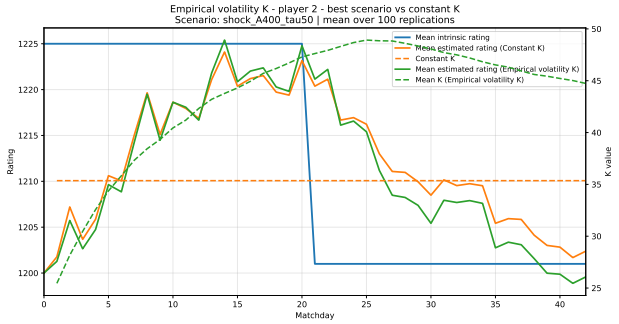



(a) Player 2, $N = 4$ and $T = 78$.

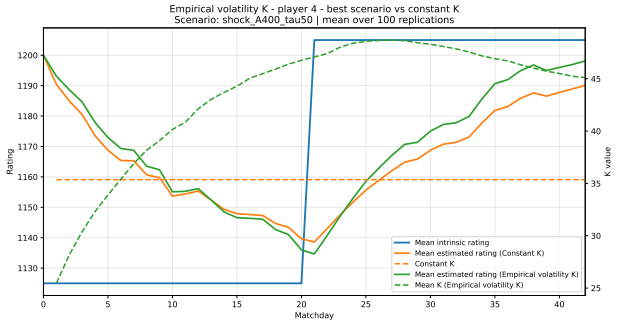



(b) Player 8, $N = 20$ and $T = 76$.

Figure 5.10: Worst-case scenario for analytical method `Uncertainty_`$K$.

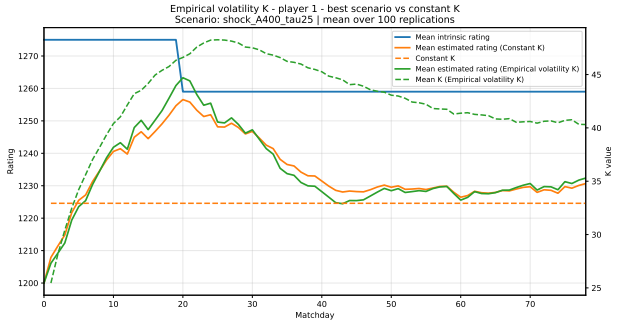



(a) Player 2

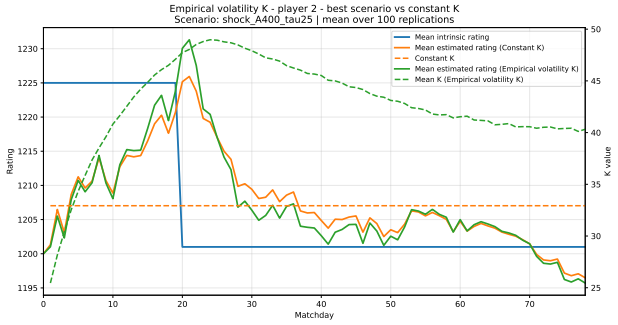



(b) Player 4

Figure 5.11: Best-case scenario for analytical method `Volatility_`$K$ with $N = 4$ and $T = 42$.




(a) Player 1




(b) Player 2

Figure 5.12: Best-case scenario for analytical method `Volatility_`$K$ with $N = 4$ and $T = 78$.

## 5.5 Euroleague basketball data

The simulated experiments identify situations in which adaptive learning rates can be useful, but the decisive test is whether these mechanisms transfer to real observations. The Euroleague experiment therefore evaluates the same families without access to intrinsic team ratings. As a consequence, only Brier score and cumulative accuracy can be considered.

The comparison is particularly demanding for the adaptive methods. Unlike the simulations, the real data do not provide explicit information about whether or when a team's underlying strength changes. Moreover, several sources of variation that are not taken into account in the simulation—such as roster changes, injuries, competition-specific effects, or temporary shape—may simultaneously influence match outcomes. The season-by-season heatmaps are therefore useful not only for comparing mean performance but also for assessing whether an apparent improvement is persistent across seasons.

### 5.5.1 Numerical methods

The Euroleague results reinforce the negative conclusion obtained for the numerical family in simulation. As shown in Table 5.5, every numerical adaptive method produces both a larger mean Brier score and a lower mean cumulative accuracy than the constant-$K$ baseline. Constant $K$ reaches an accuracy of 0.670715, whereas even the best numerical Brier score, obtained by the trajectory $K_t$ method, corresponds to a 2.769% deterioration (0.219455 compared with 0.213542 for constant $K$). For accuracy, the best numerical value is obtained by the terminal $K_{ijt}$ method at 0.656040, which also remains clearly below the baseline. We can also notice that the Shrinking $K$ sqrt method still performs below the constant $K$ rule, even in this practical case.

Figure 5.13 shows that this aggregate deterioration is not generated by every season individually. Some methods slightly outperform constant $K$ in isolated seasons, as indicated by negative cells (for example : `2015-2016` season for accuracy). However, these gains are compensated by larger deterioration in other seasons, and none of the proposed numerical rule displays a consistently favorable pattern across the thirteen retained seasons.

The real-data results therefore support the interpretation that retrospective optimization of $K$ is too unstable for this forecasting problem in its current form. This additional flexibility does not result in improved out-of-sample forecasts and appears particularly vulnerable to season-specific noise.

### 5.5.2 Analytical methods

For the Euroleague analytical experiment, the retained fixed configurations shown in Table 5.6 are drawn from the simulated 20 players case with short season[1] :

The analytical methods are again considerably more competitive. Among the five rules, the Bernoulli-variance method obtains the best mean Brier score, improving on constant $K$ by

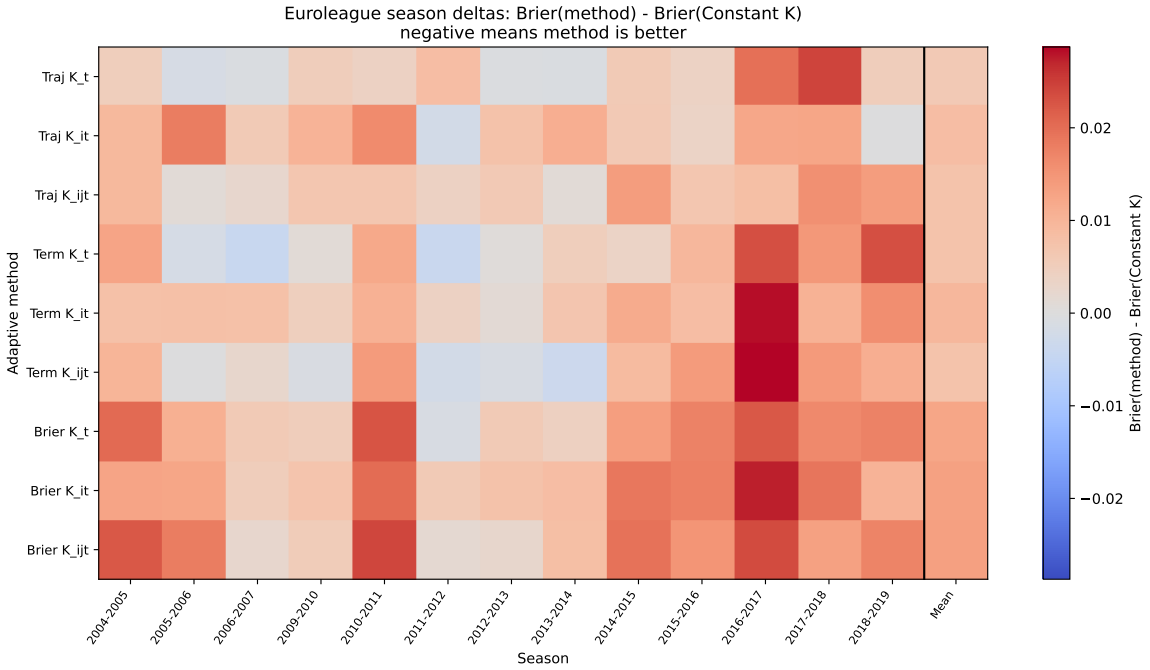
[1] I made this choice because this format is the closest to euroleague format among the four simulated ones.

| **Method** | **Final Brier score** | **Final cumulative accuracy** |
|---|---|---|
| *Baseline methods* | | |
| Constant $K$ | 0.213542 | 0.670715 |
| Shrinking $K$ sqrt | 0.218542 | 0.661998 |
| *Numerical adaptive methods* | | |
| Trajectory – $K_t$ | 0.219455 | 0.651897 |
| Trajectory – $K_{it}$ | 0.221998 | 0.644646 |
| Trajectory – $K_{ijt}$ | 0.220845 | 0.651090 |
| Terminal – $K_t$ | 0.220873 | 0.651642 |
| Terminal – $K_{it}$ | 0.223149 | 0.643190 |
| Terminal – $K_{ijt}$ | 0.220908 | 0.656040 |
| Brier terminal – $K_t$ | 0.225900 | 0.645931 |
| Brier terminal – $K_{it}$ | 0.226755 | 0.639243 |
| Brier terminal – $K_{ijt}$ | 0.226847 | 0.642059 |

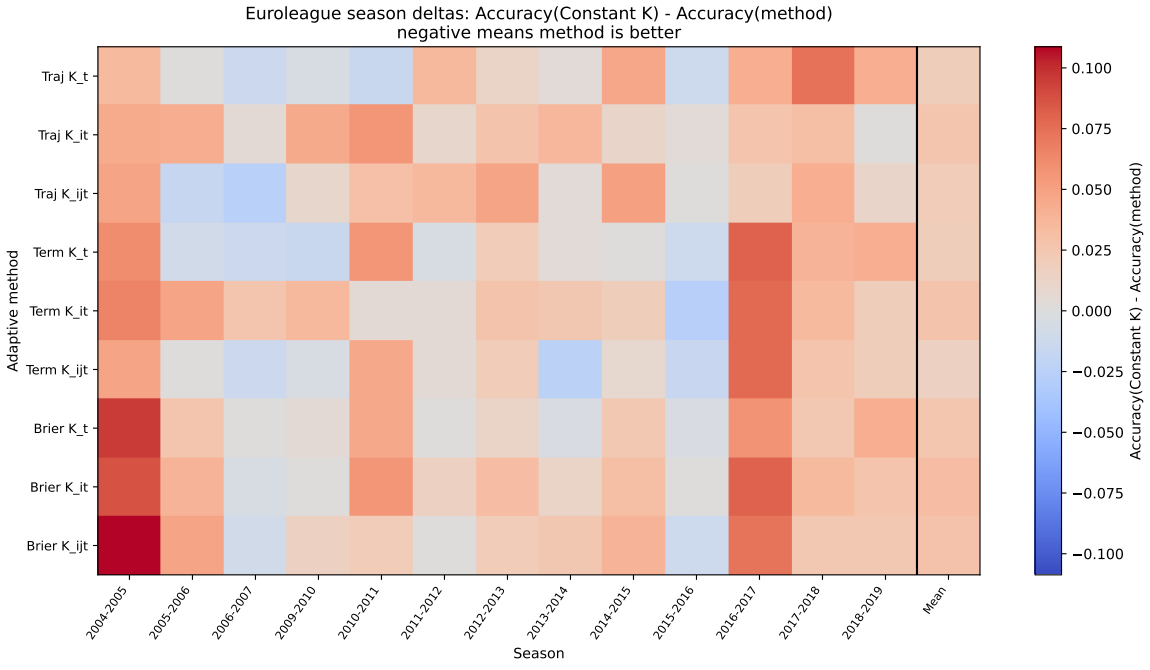
Table 5.5: Mean predictive performance of the baseline and numerical adaptive $K$-factor methods over the selected Euroleague seasons.

| **Method** | $\mathbf{K}_{\min}$ | $\mathbf{K}_{\max}$ | $\boldsymbol{\eta}$ | **c** | $\boldsymbol{\delta}$ | **h** | $\boldsymbol{\tau}$ |
|---|---|---|---|---|---|---|---|
| Uncertainty-based $K$ | 0 | 100 | 0.05 | 0.40 | – | – | – |
| Bernoulli variance $K$ | 5 | 40 | – | – | – | – | – |
| Persistent residual $K$ | 20 | 60 | 0.05 | 0.10 | – | – | – |
| Empirical volatility $K$ | 0 | 100 | 0.05 | 0.20 | – | – | – |
| Change-point $K$ | 20 | 60 | – | – | 0.05 | 2.00 | 8 |

Table 5.6: Selected parameters for analytical methods for Euroleague experiment, retained from the $N = 20$ players and $T = 38$ simulation.




(a) Brier score relative to constant $K$.




(b) Accuracy relative to constant $K$.

Figure 5.13: Performance of the numerical adaptive $K$-factor methods relative to the constant-$K$ baseline applied to the Euroleague dataset.

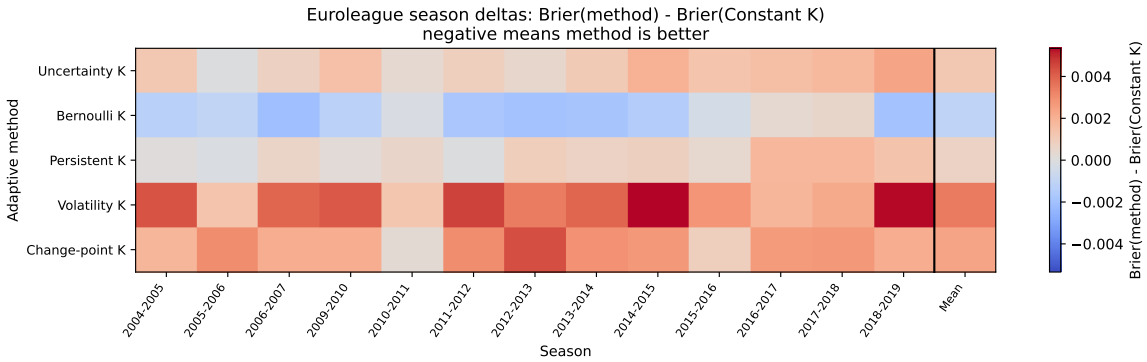
0.513% (0.212446 compared with 0.213542). This corresponds to a small but genuine average improvement in probabilistic prediction. Its cumulative accuracy, 0.669958, remains extremely close to, but slightly below, the constant-$K$ value of 0.670715.

The other analytical methods do not improve the aggregate Brier score. Persistent residual $K$ remains comparatively close to the baseline, whereas the empirical-volatility method, which performed well in several simulated shock scenarios, shows a 1.587% deterioration in Brier score (0.216930 compared with 0.213542 for constant $K$) and an accuracy of 0.656269. This inconsistency is revealing: the volatility rule appears effective when the simulated change mechanism matches its intended behavior, but real season-to-season dynamics are more complex and may cause the rule to react to fluctuations that do not correspond to persistent changes in team strength.

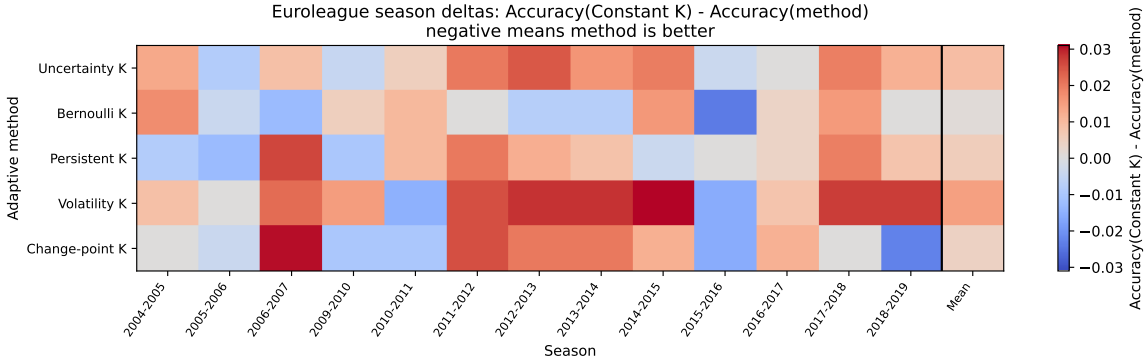
Figure 5.14 also reveals substantial heterogeneity between seasons. The Bernoulli method produces favorable Brier differences in several seasons, while other methods alternate between improvements and deterioration. The aggregate advantage of the Bernoulli rule should therefore be interpreted as modest rather than as evidence of uniform dominance. Despite the fact that this method did not stand out in the simulations, it is the only candidate in the real-data experiment that improves the principal metric of this thesis relative to the calibrated constant-$K$ benchmark .

| **Method** | **Final Brier score** | **Final cumulative accuracy** |
|---|---|---|
| *Baseline methods* | | |
| Constant $K$ | 0.213542 | 0.670715 |
| Shrinking $K$ sqrt | 0.218542 | 0.661998 |
| *Analytical adaptive methods* | | |
| Uncertainty-based $K$ | 0.214678 | 0.661539 |
| Bernoulli variance $K$ | **0.212446** | 0.669958 |
| Persistent residual $K$ | 0.214230 | 0.665224 |
| Empirical volatility $K$ | 0.216930 | 0.656269 |
| Change-point $K$ | 0.215886 | 0.666480 |

Table 5.7: Mean predictive performance of the baseline and analytical adaptive $K$-factor methods over the selected Euroleague seasons.




(a) Brier score relative to constant $K$.




(b) Accuracy relative to constant $K$.

Figure 5.14: Performance of the analytical adaptive $K$-factor methods relative to the constant-$K$ baseline applied to the Euroleague dataset.

## 5.6 Main findings

Three principal conclusions emerge from the experiments. First, the calibrated constant-$K$ Elo system is a strong benchmark. The deterministic square-root decrease performs systematically worse, showing that reducing responsiveness as information accumulates is not appropriate when strengths may evolve over time.

Second, adaptive learning rates are able to improve the reconstruction of dynamic intrinsic ratings, particularly after abrupt changes. This effect is much clearer for the analytical methods than for the numerical ones. However, improvements in MAE do not automatically lead to improvements in predictive metrics. The numerical methods illustrate this distinction particularly strongly: despite several favorable rating tracking results, all 9 methods perform worse than constant $K$ in aggregate Brier score and accuracy.

Third, the analytical rules provide the most encouraging predictive results. In simulations, uncertainty- and volatility-based adaptations outperform constant $K$ in several non-stationary scenarios and occasionally in aggregate performance. On Euroleague data, the Bernoulli-variance rule achieves the lowest mean Brier score of all deployable methods, although the improvement over constant $K$ is small and is not accompanied by an accuracy gain. Overall, the results support adaptive $K$ primarily as a mechanism for handling identifiable non-stationarity rather than as a universally superior replacement for a well-calibrated constant learning rate.

# Chapter 6

# Discussions & Conclusion

## 6.1 Discussions

The experimental results nuance the initial intuition that allowing the Elo update factor to depend on past observations should systematically improve prediction. Adaptivity is useful only when the information used to modify $K$ is sufficiently informative about a genuine change in the underlying strength. Otherwise, the additional responsiveness amplifies the stochastic component of match outcomes. The constant-$K$ baseline performs well precisely because it represents a fixed compromise between these two effects.

This explains why the analytical methods appear more successful than the numerical procedures. The numerical rules optimize a retrospective objective over a large search space and therefore have enough flexibility to respond to singular past outcomes. Their poor predictive results suggest that this flexibility creates an important generalization problem. In contrast, the analytical rules restrict adaptation to an interpretable statistic of the residual process. This structural constraint appears to act as a form of regularization and produces more stable behaviour across scenarios.

The results also emphasize the importance of distinguishing the quality of the latent rating estimate from the quality of the resulting prediction. MAE is valuable in simulation because it directly measures whether the algorithm follows the intrinsic strengths, but it is not itself the final prediction objective. In practical applications, intrinsic ratings are unavailable, and a method must ultimately be judged through observable predictive quantities such as the Brier score and accuracy. The strong MAE improvements obtained in some simulations should therefore be viewed as evidence that adaptive $K$ can track non-stationarity, rather than as sufficient evidence of superior forecasting.

Several limitations should also be kept in mind. The simulated scenarios, while deliberately diverse, remain stylized representations of ability dynamics. The Euroleague validation contains only the retained seasons available in the dataset and does not provide a sufficient validation process. Furthermore, the analytical hyperparameters are calibrated on simulated environments and then transferred to real data, while the numerical procedures rely on approximate stochastic optimization for computational feasibility. This means that every candidate is not considered,

the method is then probably missing the very best one sometimes. These choices make the comparison transparent and reproducible, but they also mean that small performance differences, such as the Brier-score advantage of the Bernoulli-variance rule, should be interpreted cautiously.

From a practical standpoint, the additional complexity of an adaptive method is therefore justified only if it provides a sufficiently stable advantage. The present results do not support replacing constant $K$ by a highly flexible optimization-based rule. They do, however, motivate further investigation of low-dimensional analytical or hybrid adaptations capable of remaining close to the calibrated baseline under stable conditions while increasing responsiveness when there is strong evidence of a structural change.

## 6.2 Answers to research questions

**How can the predictive quality of a rating system be formally evaluated ?**

The experiments confirm the usefulness of combining several complementary metrics. The Brier score is the principal criterion because it evaluates the complete predicted probability rather than only the most likely outcome and, unlike a rating-error metric, remains observable on real data. Accuracy provides an intuitive secondary measure of the proportion of correctly predicted winners, although it discards information about prediction confidence. In simulated environments, centered MAE additionally measures how closely the estimated ratings reproduce the intrinsic rating structure while respecting the translation invariance of the Elo model. Together, these metrics distinguish probabilistic forecasting quality, classification performance, and latent-rating reconstruction.

**Can we design an adaptation rule for parameter $K$ that performs better than the usual constant-$K$ rule ?**

The answer is positive in specific settings but negative if "better" is understood as a uniform improvement across environments and metrics. Several analytical adaptations outperform constant $K$ in particular non-stationary simulated scenarios, especially after abrupt changes in intrinsic strength, and they frequently improve centered MAE. The uncertainty-based and empirical-volatility methods also obtain small aggregate predictive gains in some simulated configurations. In the Euroleague experiment, the Bernoulli-variance method improves the mean Brier score by 0.513% (from 0.213542 to 0.212446).

However, none of the proposed adaptive method dominates constant $K$ consistently. The 9 numerical methods all deteriorate aggregate Brier score and accuracy, and most analytical gains are scenario-dependent. Even the Euroleague Bernoulli-variance rule does not improve cumulative accuracy. The experiments therefore do not identify a universally superior adaptive $K$ rule. Instead, they show that adaptation can be beneficial when it is sufficiently constrained and when the underlying environment contains changes that the adaptation statistic is able to detect.

## 6.3 Conclusion

This thesis investigated whether the learning rate of the Elo Rating System can be adapted to observed information in order to improve sports-outcome prediction. 9 optimization-based methods and 5 analytical adaptation rules were developed and evaluated against a calibrated constant-$K$ Elo model, a deterministic decreasing-$K$ rule, and an oracle benchmark. The comparison was conducted over a large simulation experiment containing stationary and time-varying rating processes and was subsequently extended to Euroleague basketball data.

The results show that the problem is fundamentally governed by the trade-off between responsiveness and stability. A variable $K$ can improve the tracking of evolving player strengths, particularly after abrupt changes, but reacting to observations also creates the risk of interpreting random outcomes as changes in ability. This explains why improved rating reconstruction does not systematically translate into improved prediction. To illustrate this conclusion, we could say that the proposed adaptive methods fall on the responsiveness side of the trade-off, straying too far from the desired middle ground.

Among the investigated approaches, the numerical optimization methods do not provide a convincing alternative to constant $K$: their additional flexibility results in worse aggregate predictive performance both in simulation and on Euroleague data. The analytical methods are more promising. Their constrained and interpretable adaptation mechanisms produce substantial rating-tracking improvements in several dynamic environments and occasional predictive gains. In particular, the Bernoulli-variance rule achieves a slightly lower mean Brier score than constant $K$ on the Euroleague data.

The main conclusion is therefore not that, among the proposed methods, an adaptive learning rate universally outperforms the classical Elo update. Rather, a well-calibrated constant $K$ remains a remarkably robust solution, while adaptive rules become useful when they can identify meaningful non-stationarity without reacting excessively to observation noise. Future work could consequently focus on hybrid procedures that retain a stable baseline learning rate and activate stronger adaptation only when sufficient evidence of a persistent change is available, as well as on validation across additional sports, competitions, and bigger real-world datasets.

# Bibliography


[ACDA22] Giovanni Angelini, Vincenzo Candila, and Luca De Angelis. Weighted elo rating for tennis match predictions. *European Journal of Operational Research*, 297(1):120–132, 2022.

[BLSR20] Heejong Bong, Wanshan Li, Shamindra Shrotriya, and Alessandro Rinaldo. Nonparametric estimation in the dynamic bradley-terry model. In *International Conference on Artificial Intelligence and Statistics*, pages 3317–3326. PMLR, 2020.

[Bri50] GW Brier. Verification of forecasts expressed in terms of probability,„monthly weather review, 78 (1), 1950.

[BV22] M. Branders and A. Vekemans. Recovering weights from comparison results in extensions of btl model. Master's thesis, UCLouvain, 2022.

[Chea] Chess.com. Chess terms : Elo rating system. `https://www.chess.com/terms/elo-rating-chess`.

[Cheb] Chess.com. Chess terms : Round-robin tournament. `https://www.chess.com/terms/round-robin-chess-tournament`.

[Elo78] Arpad E. Elo. *The Rating of Chessplayers, Past and Present.* Arco Publishing, New York, 1st edition, 1978.

[Elo86] Arpad E. Elo. *The Rating of Chessplayers, Past and Present.* Arco Publishing, New York, 2nd edition, 1986.

[FIF18] FIFA. Revision of the fifa/coca-cola world ranking. `https://digitalhub.fifa.com/m/f99da4f73212220/original/edbm045h0udbwkqew35a-pdf.pdf`, 2018.

[GJ99] Mark E Glickman and Albyn C Jones. Rating the chess rating system. *Chance-Berlin Then New York-*, 12:21–28, 1999.

[Gla] Vadim Gladky. Basketball - euroleague results 2003-2020. `https://www.kaggle.com/datasets/vadimgladky/euroleague-basketball-results-20032019`.

[Gli95] Mark E Glickman. The glicko system. *Boston University*, 16(8):9, 1995.

[Gli99] Mark E Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 48(3):377–394, 1999.

[Gli12] Mark E Glickman. Example of the glicko-2 system. *Boston University*, 28:2012, 2012.

[GR07] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[HA10] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010.

[HMG06] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006.

[Ing21] Martin Ingram. How to extend elo: a bayesian perspective. *Journal of Quantitative Analysis in Sports*, 17(3):203–219, 2021.

[LG21] Jan Lasek and Marek Gagolewski. Interpretable sports team rating models based on the gradient descent algorithm. *International Journal of Forecasting*, 37(3):1061–1071, 2021.

[lic] lichess.org. Chess rating systems. `https://lichess.org/page/rating-systems`.

[LKH14] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4):2065–2073, 2014.

[Mas22] Bastien Massion. Tennis matches outcome prediction via low-rank approaches. Master's thesis, Université catholique de Louvain (UCLouvain), 2022.

[Pag54] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[rob26] robinrochet. Code repository for "predicting sports outcomes: Rating approach". GitHub repository, 2026. Accessed: August 2026. URL: `https://github.com/robinrochet/sportsoutcomesrating`.

[Sol22] Nate Solon. How Elo ratings actually work. `https://www.zwischenzug.gg/p/how-elo-ratings-actually-work`, December 2022. Zwischenzug Chess (Substack newsletter). Accessed 2026-08-16.

[ST23] Leszek Szczecinski and Raphaëlle Tihon. Simplified kalman filter for on-line rating: one-fits-all approach. *Journal of Quantitative Analysis in Sports*, 19(4):295–315, 2023.

[VHB+26] Hanke Vermeiren, Abe D Hofman, Maria Bolsinova, Han LJ Van der Maas, and Wim Van Den Noortgate. Balancing stability and flexibility: investigating a dynamic k value approach for the elo rating system in adaptive learning environments: H. vermeiren et al. *User Modeling and User-Adapted Interaction*, 36(1):4, 2026.

[Wik26a] Wikipedia contributors. Bradley–terry model, 2026. Wikipedia, The Free Encyclopedia. URL: `https://en.wikipedia.org/wiki/Bradley%E2%80%93Terry_model`.

[Wik26b] Wikipedia contributors. Elo rating system, 2026. Wikipedia, The Free Encyclopedia. URL: `https://en.wikipedia.org/wiki/Elo_rating_system`.

# Technical Resources

The computational work presented in this thesis was carried out using Python, which was used to implement the rating methods, perform the simulations, optimize model parameters, process the experimental data, and generate the numerical results. The main libraries used include NumPy and Pandas for numerical computation and data manipulation, Matplotlib for data visualization, and Numba to accelerate the most computationally intensive parts of the simulations. The code was developed and executed using Visual Studio Code. ChatGPT (version GPT-5.6) was used as a coding assistant for debugging, and as a writing assistant for layout purposes.

# Appendix A

# Complete Description of the Simulated Leagues

## A.1 Common rating constructions

Define the initial equally spaced strength vector with total gap $G$ by

$$b_i(G) = 1200 + \frac{G}{2} - \frac{(i-1)G}{N-1}, \qquad i = 1, \ldots, N.$$

Thus, before the final re-centring and clipping operation, player 1 is the strongest and player $N$ is the weakest.

### A.1.1 Stationary profiles

For the constant family,

$$E_{it} = b_i(G) \qquad \text{for all } t,$$

with $G \in \{100, 200, 400\}$. These scenarios measure performance when no adaptation to changing intrinsic strength is required. Increasing $G$ changes the predictability of the outcomes: larger gaps produce more unequal win probabilities.

### A.1.2 Linear drift profiles

The linear family starts from $b_i(200)$. Let $d_i(A)$ be an equally spaced vector from $+A$ to $-A$, centred to have mean zero. The profile is

$$E_{it} = b_i(200) + \frac{t}{T} d_i(A),$$

with $A \in \{50, 100, 200, 400\}$. Strong players therefore tend to drift upward while weak players drift downward before the final common re-centring. The parameter $A$ controls the maximum seasonal displacement and tests gradual non-stationarity.

### A.1.3 Abrupt shock profiles

The shock family starts from $b_i(150)$. At the shock time

$$t_{\text{shock}} = \text{round}(\tau T),$$

half of the players receive a displacement $+A$ and the other half receive $-A$. The assignment is randomly shuffled in every replication and is centred to have mean zero. Hence

$$E_{it} = \begin{cases} b_i(150), & t < t_{\text{shock}}, \\ b_i(150) + s_i, & t \geq t_{\text{shock}}, \end{cases}$$

where $s_i \in \{-A, +A\}$. The amplitudes are $A \in \{100, 200, 300, 400\}$ and the relative times are $\tau \in \{0.25, 0.50, 0.75\}$. These twelve scenarios separate the effect of shock size from the amount of post-shock data available for recovery.

### A.1.4 Random-walk profiles

The random-walk family starts from $b_i(150)$. At every matchday,

$$E_{it} = E^{\star}_{i,t-1} + \varepsilon_{i,t}, \qquad \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma^2),$$

where the innovation vector is centred across players before it is added. The standard deviations are $\sigma \in \{5, 10, 20, 40\}$. These cases represent persistent but unpredictable local changes in strength.

### A.1.5 Oscillating profiles

The oscillating family starts from $b_i(120)$ and uses

$$E_{it} = b_i(120) + A \sin\left(\frac{2\pi t}{P_i} + \phi_i\right).$$

The phases are equally spaced over $[0, 2\pi)$, and the periods are equally spaced from

$$\max\left(12, \frac{T}{3}\right) \quad \text{to} \quad \max(18, 0.9T).$$

The amplitudes are $A \in \{50, 100, 150, 250\}$. Because players have different phases and periods, the ranking can change repeatedly over a season.

## A.2 Complete scenario list

The same ordered list is used by the analytical and numerical scripts and by both league sizes.

Table A.1: Complete list of the thirty simulated scenarios.

| No. | Family | Parameters | Script label |
|---|---|---|---|
| 1 | Constant | $G = 100$ | `constant_gap100` |
| 2 | Constant | $G = 200$ | `constant_gap200` |
| 3 | Constant | $G = 400$ | `constant_gap400` |
| 4 | Linear | $A = 50$ | `linear_A50` |
| 5 | Linear | $A = 100$ | `linear_A100` |
| 6 | Linear | $A = 200$ | `linear_A200` |
| 7 | Linear | $A = 400$ | `linear_A400` |
| 8 | Shock | $A = 100,\ \tau = 0.25$ | `shock_A100_tau25` |
| 9 | Shock | $A = 100,\ \tau = 0.50$ | `shock_A100_tau50` |
| 10 | Shock | $A = 100,\ \tau = 0.75$ | `shock_A100_tau75` |
| 11 | Shock | $A = 200,\ \tau = 0.25$ | `shock_A200_tau25` |
| 12 | Shock | $A = 200,\ \tau = 0.50$ | `shock_A200_tau50` |
| 13 | Shock | $A = 200,\ \tau = 0.75$ | `shock_A200_tau75` |
| 14 | Shock | $A = 300,\ \tau = 0.25$ | `shock_A300_tau25` |
| 15 | Shock | $A = 300,\ \tau = 0.50$ | `shock_A300_tau50` |
| 16 | Shock | $A = 300,\ \tau = 0.75$ | `shock_A300_tau75` |
| 17 | Shock | $A = 400,\ \tau = 0.25$ | `shock_A400_tau25` |
| 18 | Shock | $A = 400,\ \tau = 0.50$ | `shock_A400_tau50` |
| 19 | Shock | $A = 400,\ \tau = 0.75$ | `shock_A400_tau75` |
| 20 | Random walk | $\sigma = 5$ | `rw_sigma5` |
| 21 | Random walk | $\sigma = 10$ | `rw_sigma10` |
| 22 | Random walk | $\sigma = 20$ | `rw_sigma20` |
| 23 | Random walk | $\sigma = 40$ | `rw_sigma40` |
| 24 | Oscillating | $A = 50$ | `osc_A50` |
| 25 | Oscillating | $A = 100$ | `osc_A100` |
| 26 | Oscillating | $A = 150$ | `osc_A150` |
| 27 | Oscillating | $A = 250$ | `osc_A250` |
| 28 | Mixed | Low, $s = 0.5$ | `mixed_profiles_low` |
| 29 | Mixed | Medium, $s = 1.0$ | `mixed_profiles_medium` |
| 30 | Mixed | High, $s = 1.5$ | `mixed_profiles_high` |

## A.3 Mixed-profile scenarios 28–30

The mixed scenarios deliberately place players with different forms of non-stationarity in the same league. Let $s \in \{0.5, 1, 1.5\}$ denote the low, medium, and high amplitude scale. Their

common parameters are

$$G_{\text{constant}} = 200s, \qquad A_{\text{linear}} = 200s,$$
$$A_{\text{shock}} = 300s, \qquad \sigma_{\text{rw}} = 20s,$$
$$A_{\text{osc}} = 150s.$$

Consequently, the three mixed levels use the parameter sets shown in Table A.2.

Table A.2: Parameters of the three mixed-profile scenarios.

| Level | Constant gap | Linear $A$ | Shock $A$ | RW $\sigma$ | Oscillation $A$ |
|---|---|---|---|---|---|
| Low | 100 | 100 | 150 | 10 | 75 |
| Medium | 200 | 200 | 300 | 20 | 150 |
| High | 300 | 300 | 450 | 30 | 225 |

### A.3.1 Four-player composition

The four-player mixed league contains one constant player, one linearly drifting player, one shocked player, and one random-walk player. An oscillating profile is not included because only four players are available. Before the common re-centring operation, the profiles are assigned as follows:

Table A.3: Player profiles in the four-player mixed scenarios.

| Player | Family | Definition |
|---|---|---|
| 1 | Constant | Remains at its initial base strength. |
| 2 | Linear | Moves upward by $A_{\text{linear}}$ over the season. |
| 3 | Shock | Moves downward by $A_{\text{shock}}$ from $0.50T$ onward. |
| 4 | Random walk | Receives independent Gaussian increments with standard deviation $\sigma_{\text{rw}}$. |

The initial base vector uses the mixed constant gap $G_{\text{constant}}$. Since the vector is re-centred at every time step, a change in one player also produces a common compensating shift in the displayed ratings of the other players while preserving relative differences.

### A.3.2 Twenty-player composition

The twenty-player mixed league contains four constant players, four linear players, six shocked players, three random-walk players, and three oscillating players. The profiles are interleaved across the initial ranking so that no family is assigned exclusively to the strongest or weakest players. Linear and shock directions refer to their pre-centring movement.

Table A.4: Player-level profiles in the twenty-player mixed scenarios.

| Player | Family | Direction | Change time |
|---|---|---|---|
| 1 | Constant | – | – |
| 2 | Linear | downward | – |
| 3 | Shock | downward | $0.25T$ |
| 4 | Random walk | – | – |
| 5 | Oscillating | – | – |
| 6 | Constant | – | – |
| 7 | Linear | upward | – |
| 8 | Shock | downward | $0.50T$ |
| 9 | Random walk | – | – |
| 10 | Oscillating | – | – |
| 11 | Constant | – | – |
| 12 | Linear | downward | – |
| 13 | Shock | downward | $0.75T$ |
| 14 | Random walk | – | – |
| 15 | Oscillating | – | – |
| 16 | Constant | – | – |
| 17 | Linear | upward | – |
| 18 | Shock | upward | $0.25T$ |
| 19 | Shock | upward | $0.50T$ |
| 20 | Shock | upward | $0.75T$ |

The four linear players alternate between downward and upward drifts. At each of the three shock times, one player jumps downward and one player jumps upward. The three random walks use independent innovations. The three oscillating players use distinct phases and periods generated over their subgroup. This construction creates a setting in which a single global $K_t$ must compromise between players with fundamentally different adaptation needs, whereas $K_{it}$ and $K_{ijt}$ can react heterogeneously.

## A.4 Randomness and reproducibility

The main simulations use the base seed 20260614. For scenario index $s$ and replication $r$, a deterministic offset is applied before separately generating intrinsic ratings and match outcomes. The preliminary sequence begins from a seed range shifted by 500,000,000, which ensures that calibration and evaluation seasons do not overlap. The same seed construction is used in the analytical and numerical scripts, so the underlying simulated seasons are reproducible and aligned across method families.

# Appendix B

# Euroleague Dataset and Preprocessing

## B.1 Supplied files and temporal coverage

Two CSV files support the experiment. The earlier file, `euroleague_dset_csv_03_19.csv`, contains 3715 matches and 21 variables. The updated file, `euroleague_dset_csv_03_20_upd.csv`, contains the same 3715 match identifiers and final scores, adds 117 observations from the beginning of the 2019–2020 season, and expands the schema to 33 variables. Both real-data scripts read the updated file.

The updated file contains 3832 rows dated from 3 November 2003 to 13 December 2019. Across the full unfiltered file, 95 distinct team-name strings appear. The code does not apply a separate alias-normalization table: after removing surrounding whitespace, every distinct string is treated as a distinct team identifier. This point is relevant because changes in naming conventions would be interpreted as changes of team identity.

## B.2 Raw variables

Table B.1 groups the 33 raw fields by their role. Only the date, team names, and final scores are retained for the rating experiment.

Table B.1: Variables available in the updated Euroleague file.

| Field(s) | Interpretation | Experimental use |
|---|---|---|
| `DATE` | Calendar date in `dd.mm.yyyy` format. | Retained and parsed. |
| `HT`, `AT` | Home-team and away-team name strings. | Retained as team identifiers. |
| `WINNER` | Source-provided winning-team name. It is redundant with the final scores. | Discarded. |

| Field(s) | Interpretation | Experimental use |
|---|---|---|
| HS, AS | Final home and away scores. | Retained to derive the binary outcome. |
| Q1H, Q1A, ..., Q4H, Q4A | Home and away points recorded for each of the four quarters. | Discarded. |
| Q1T, ..., Q4T | Total points scored by both teams in each quarter. | Discarded. |
| OTH, OTA | Home and away overtime points when supplied. Only nine rows contain non-missing overtime values. | Discarded. |
| P1H, P1A | Home and away points over the first half. | Discarded. |
| P2H, P2A | Home and away points over the second half. | Discarded. |
| P1T, P2T | Combined points scored by both teams in the first and second halves. | Discarded. |
| HTT, FTT | Additional source-provided aggregate scoring fields. They are redundant for winner prediction and are not required by the code. | Discarded. |
| GAPA1Q, ..., GAPA4Q, GAPT | Source-provided quarter-level and final score-gap features. | Discarded. |

The discarded in-game and post-game variables are not needed by an Elo model whose information set consists only of previous match outcomes. In particular, the winner, final margin, quarter scores, and half-time scores would either duplicate the target or introduce information unavailable before the match. The experiment therefore evaluates a pure rating-based forecasting problem rather than a model enriched with additional match statistics.

## B.3 Cleaning and derived features

The preprocessing sequence is identical in the analytical and numerical scripts:

- Empty strings in the required fields are converted to missing values, and rows missing DATE, HT, AT, HS, or AS are removed.
- DATE is parsed with the fixed format %d.%m.%Y. Team strings are stripped, and final scores are converted to numeric integers.
- Two tied rows are removed by the condition HomeScore $\neq$ AwayScore. One has a recorded score of 70–70 and the other 0–0. After this step, 3830 matches remain in the full cleaned file.

- The binary response at time $t$ is defined by

$$Y_{\text{home team vs away team},t} = \mathbf{1}\{\texttt{HomeScore} > \texttt{AwayScore}\}.$$

- A match from August to December is assigned to the season beginning in that calendar year. A match from January to July is assigned to the season ending in the previous calendar year.
- Rows are sorted by season, date, home team, and away team.

The resulting derived variables are summarized in Table B.2.

Table B.2: Variables created by the preprocessing code.

| Variable | Definition |
|---|---|
| `MatchDate` | Parsed calendar date. |
| `HomeTeam`, `AwayTeam` | Cleaned team-name strings. |
| `HomeScore`, `AwayScore` | Integer final scores. |
| `YHome` | Binary indicator of a home-team victory. |
| `SeasonStartYear` | Calendar year in which the assigned season begins. |
| `Season` | Label of the form `YYYY-YYYY`. |
| `Matchday` | Chronological group reconstructed so that a team appears at most once in a group. |

## B.4 Reconstruction of matchdays

The file does not contain a single common round or matchday field used by the scripts. Within each season, matches are sorted by date and team names. The algorithm begins with matchday 1 and maintains the set of teams already used in the current group. When the next match contains a team that has already appeared in that group, a new matchday is opened and the set is reset. This greedy rule guarantees that no team plays twice within a reconstructed matchday.

The resulting matchday number is a computational grouping rather than an official competition-round label. It is nevertheless sufficient for the reconstruction procedure because all predictions within a group are made from the ratings and adaptive states available at the beginning of that matchday, after which the accumulated rating changes are applied simultaneously.

## B.5 Season selection

For each constructed season, the code records its first and last date, number of distinct team strings, number of matches, and maximum reconstructed matchday. A season is considered sufficiently complete when

$$N_{\text{matchdays}} \geq 25 \qquad \text{and} \qquad N_{\text{matches}} \geq 150.$$

Table B.3: Euroleague season metadata after cleaning and matchday reconstruction.

| Season | Start | End | Teams | Matches | Matchdays | Status |
|---|---|---|---|---|---|---|
| 2003–2004 | 03 Nov 2003 | 01 May 2004 | 24 | 220 | 22 | Excluded |
| 2004–2005 | 01 Nov 2004 | 08 May 2005 | 26 | 230 | 25 | Retained |
| 2005–2006 | 31 Oct 2005 | 30 Apr 2006 | 24 | 231 | 25 | Retained |
| 2006–2007 | 24 Oct 2006 | 06 May 2007 | 24 | 230 | 25 | Retained |
| 2007–2008 | 22 Oct 2007 | 10 Apr 2008 | 24 | 227 | 23 | Excluded |
| 2008–2009 | 20 Oct 2008 | 03 May 2009 | 24 | 188 | 23 | Excluded |
| 2009–2010 | 29 Sep 2009 | 09 May 2010 | 30 | 200 | 26 | Retained |
| 2010–2011 | 24 Sep 2010 | 08 May 2011 | 33 | 198 | 28 | Retained |
| 2011–2012 | 29 Sep 2011 | 13 May 2012 | 36 | 200 | 26 | Retained |
| 2012–2013 | 11 Oct 2012 | 12 May 2013 | 24 | 250 | 31 | Retained |
| 2013–2014 | 17 Oct 2013 | 18 May 2014 | 24 | 250 | 31 | Retained |
| 2014–2015 | 23 Sep 2014 | 17 May 2015 | 31 | 258 | 33 | Retained |
| 2015–2016 | 15 Oct 2015 | 15 May 2016 | 24 | 250 | 31 | Retained |
| 2016–2017 | 28 Sep 2016 | 21 May 2017 | 18 | 261 | 39 | Retained |
| 2017–2018 | 12 Oct 2017 | 20 May 2018 | 16 | 260 | 36 | Retained |
| 2018–2019 | 11 Oct 2018 | 19 May 2019 | 16 | 260 | 37 | Retained |
| 2019–2020 | 03 Oct 2019 | 13 Dec 2019 | 18 | 117 | 13 | Excluded |

Thirteen seasons are retained, containing 3078 matches and 84 distinct team-name strings. The excluded 2003–2004, 2007–2008, and 2008–2009 seasons fail only the matchday threshold, while 2019–2020 fails both thresholds because the supplied file ends on 13 December 2019. The selected seasons contain between 16 and 36 distinct team-name strings and between 198 and 261 matches, which confirms that the dataset does not represent one fixed balanced league format over the full historical period.

## B.6 Use in the rating experiment

Each retained season is treated independently. Every team begins the season at rating 1500, and final ratings from the previous season are not carried forward. For a home team $i$ and away team $j$, the online methods use

$$\hat{P}_{ijt} = \frac{1}{1 + 10^{-\left(\hat{E}_{it} + 65 - \hat{E}_{jt}\right)/400}},$$

where 65 is the fixed home-advantage adjustment. The observed score margin does not affect the update; only $Y_{\text{home}}$ is used.

The constant baseline uses $K = 30.953333$[1], while the square-root baseline starts from

[1]Corresponding to the value retained by the simulated data experiment for 20 players and 38 matchdays.

40. The analytical and numerical method families process the same chronological matches and use the same home advantage. For the numerical methods, $K$ is re-optimized every four reconstructed matchdays from at most the previous forty matchdays.

Only the Brier score and classification accuracy are reported on this dataset. Intrinsic ratings are not observed, so centred MAE cannot be evaluated. The oracle method is no longer performed, due to the impossibility of getting $E_{it}$ values, or because of the inconsistency inherent in making predictions after an observation, leading to a 100% accuracy value that would not mean anything.

## B.7 Differences from the simulated leagues

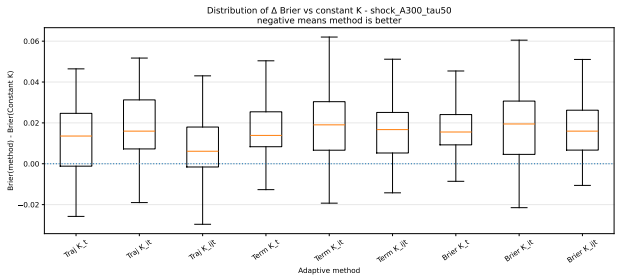
The real-data experiment differs from the controlled simulation in several important ways. The number and identity of teams change across seasons, schedules are not balanced repeated round robins, the number of matches per season is much smaller than the 760 matches of the simulated twenty-player league, and home advantage is explicitly modelled. Furthermore, the true rating trajectories are unknown, team labels are taken directly from the file, and matchdays are reconstructed rather than supplied. These differences make the Euroleague analysis a test of external robustness rather than a direct replication of the simulated setting.

# Appendix C

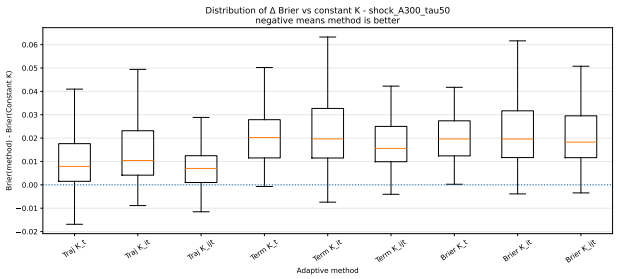
# Boxplots for representative scenarios

I selected some of the relevant Boxplots that illustrate the purpose of the analysis in Chapter 5.

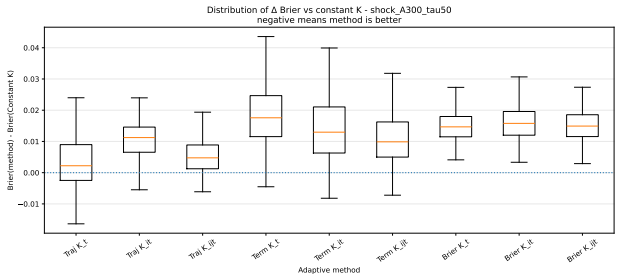
## C.1 Numerical methods



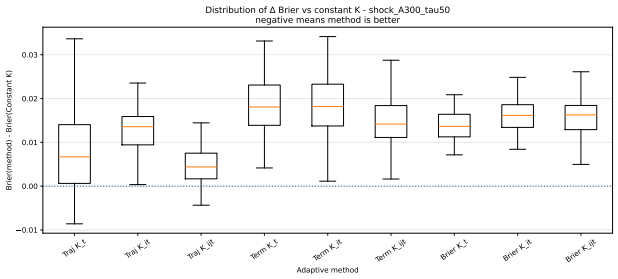

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

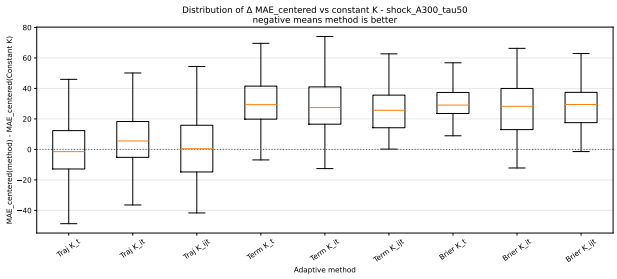



(c) $N = 20$ players, $T = 38$ matchdays.

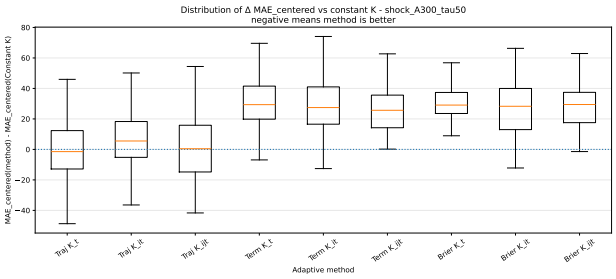



(d) $N = 20$ players, $T = 76$ matchdays.

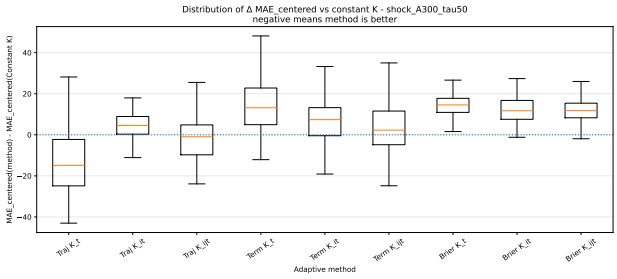
Figure C.1: Distribution of the paired Brier-score differences between the numerical adaptive methods and the constant-$K$ baseline for the `shock_A300_tau50` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.



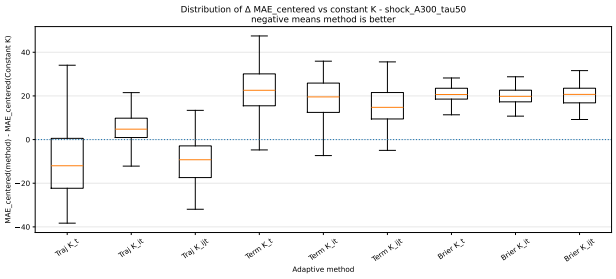

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

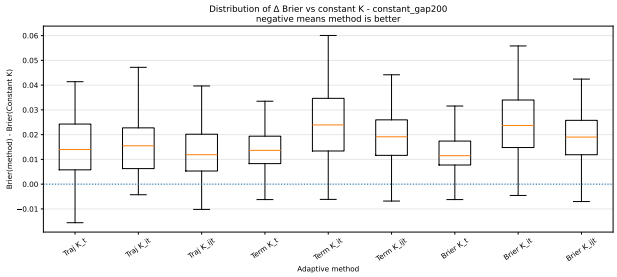



(c) $N = 20$ players, $T = 38$ matchdays.

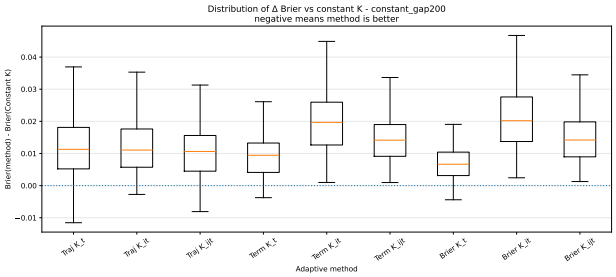



(d) $N = 20$ players, $T = 76$ matchdays.

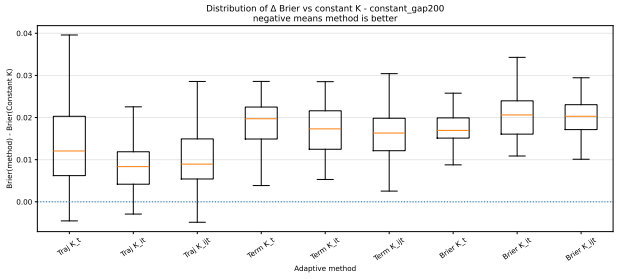
Figure C.2: Distribution of the paired MAE differences between the numerical adaptive methods and the constant-$K$ baseline for the `shock_A300_tau50` scenario. Negative values indicate that the adaptive method achieves a lower MAE than the constant-$K$ baseline.



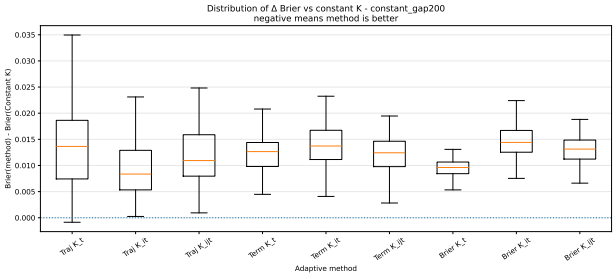

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

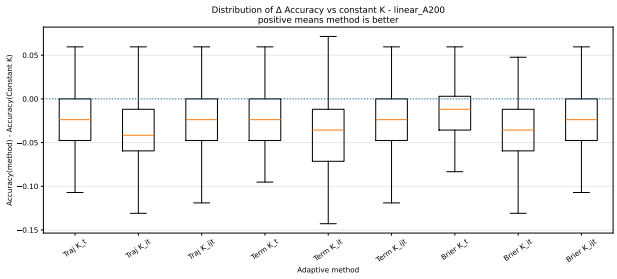



(c) $N = 20$ players, $T = 38$ matchdays.

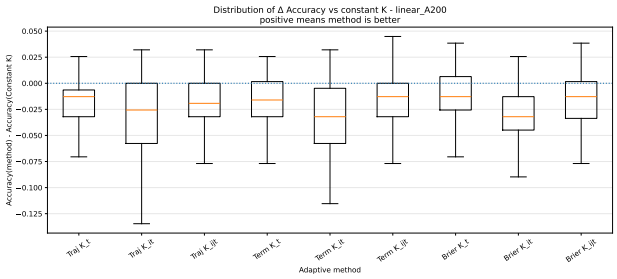



(d) $N = 20$ players, $T = 76$ matchdays.

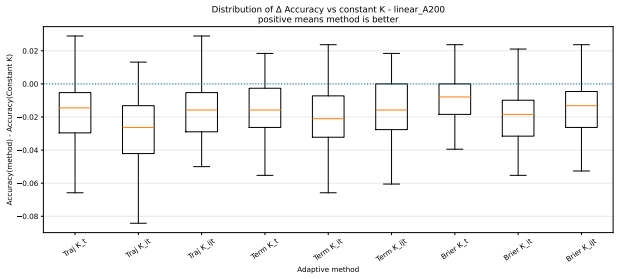
Figure C.3: Distribution of the paired Brier-score differences between the numerical adaptive methods and the constant-$K$ baseline for the `constant_gap200` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.



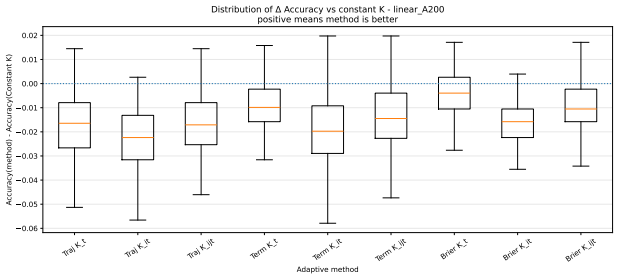

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

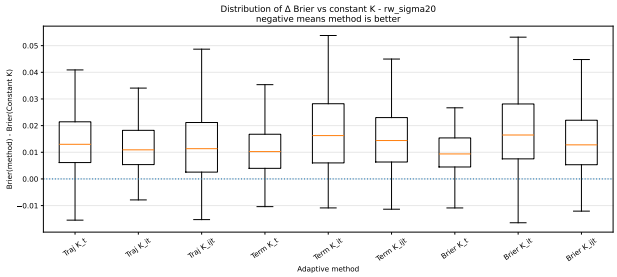



(c) $N = 20$ players, $T = 38$ matchdays.

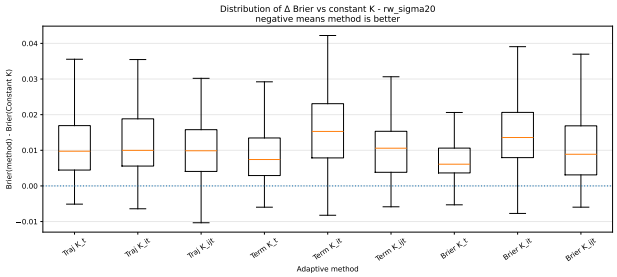



(d) $N = 20$ players, $T = 76$ matchdays.

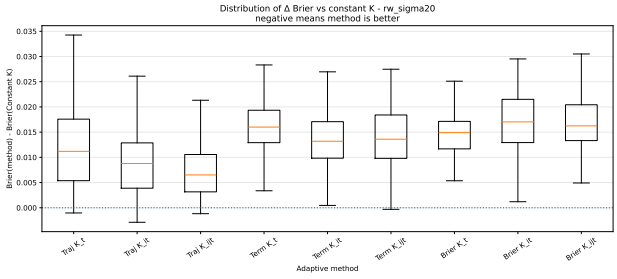
Figure C.4: Distribution of the accuracy differences between the numerical adaptive methods and the constant-$K$ baseline for the `linear_A200` scenario. Positive values indicate that the adaptive method achieves a higher accuracy score than the constant-$K$ baseline.



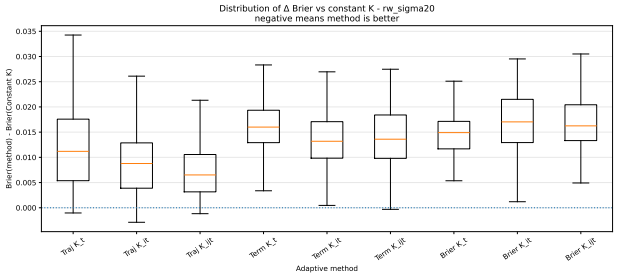

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.



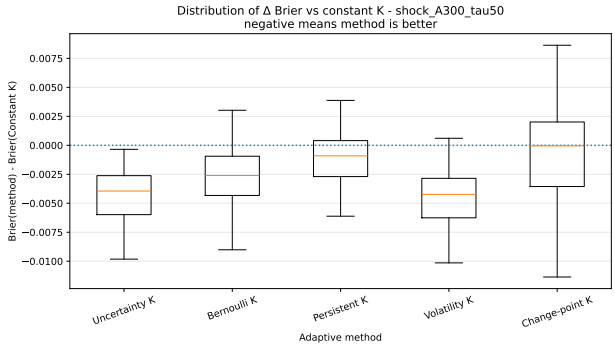

(c) $N = 20$ players, $T = 38$ matchdays.



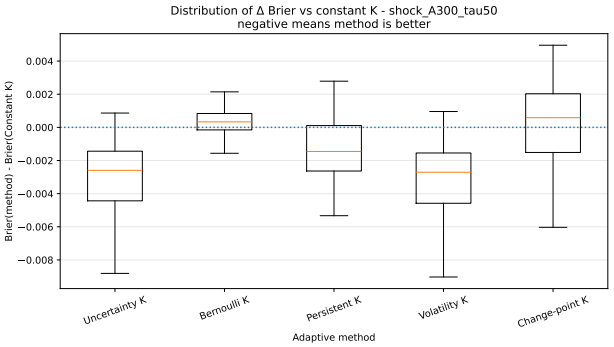

(d) $N = 20$ players, $T = 76$ matchdays.

Figure C.5: Distribution of the paired Brier-score differences between the numerical adaptive methods and the constant-$K$ baseline for the `rw_sigma20` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.

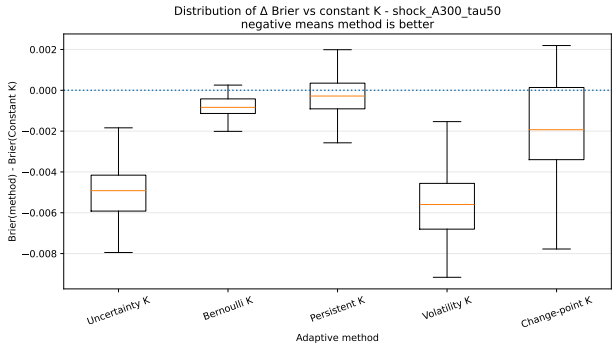
## C.2 Analytical methods



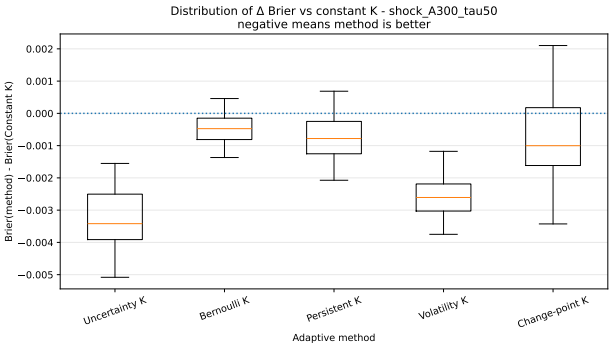

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

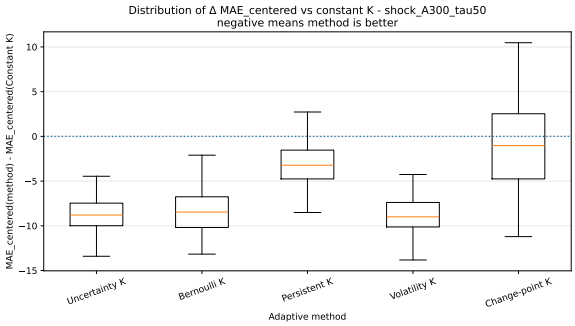



(c) $N = 20$ players, $T = 38$ matchdays.

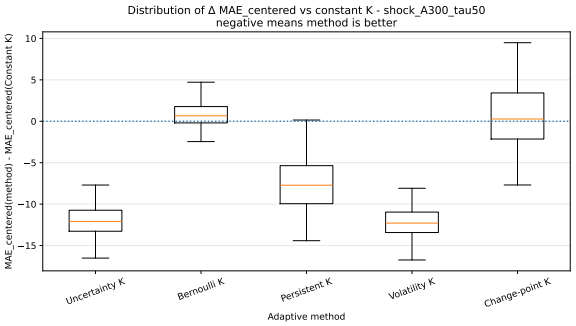



(d) $N = 20$ players, $T = 76$ matchdays.

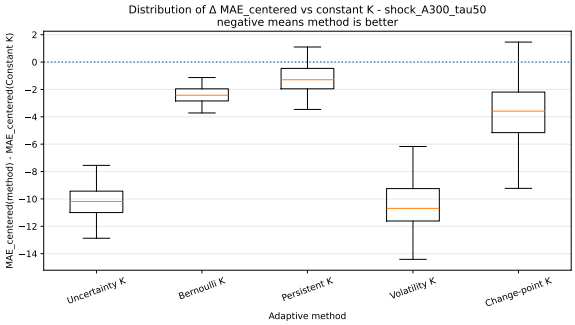
Figure C.6: Distribution of the paired Brier-score differences between the analytical adaptive methods and the constant-$K$ baseline for the `shock_A300_tau50` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.



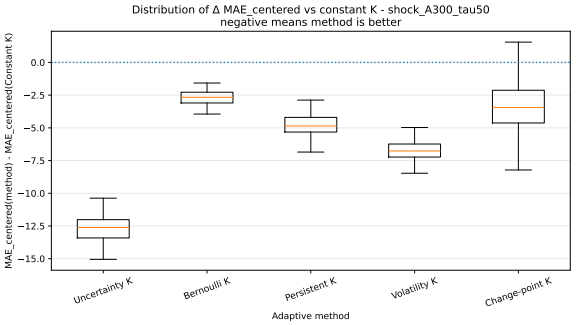

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

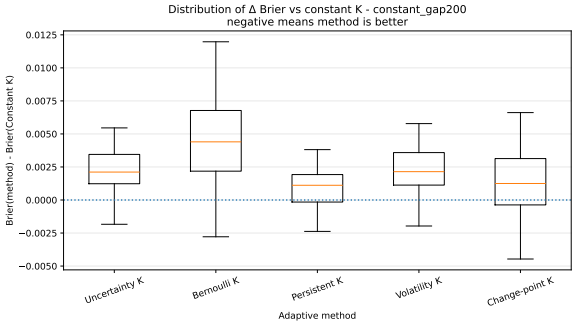



(c) $N = 20$ players, $T = 38$ matchdays.

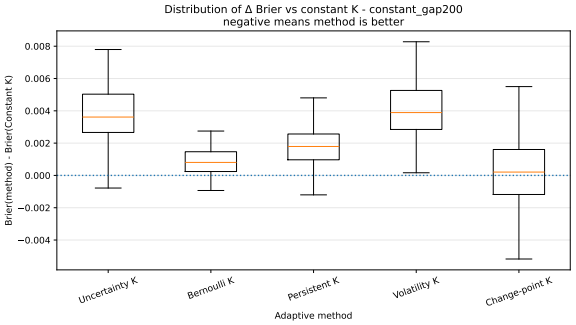



(d) $N = 20$ players, $T = 76$ matchdays.

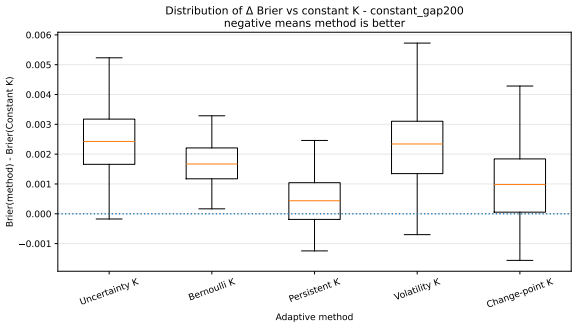
Figure C.7: Distribution of the paired MAE differences between the analytical adaptive methods and the constant-$K$ baseline for the `shock_A300_tau50` scenario. Negative values indicate that the adaptive method achieves a lower MAE than the constant-$K$ baseline.



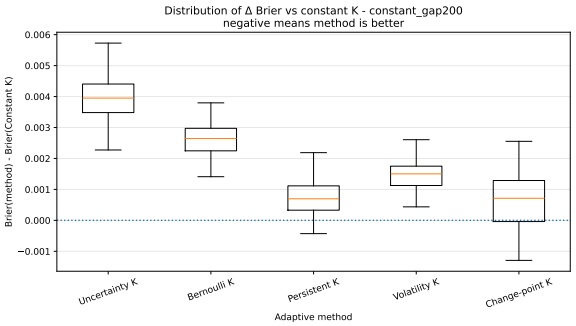

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

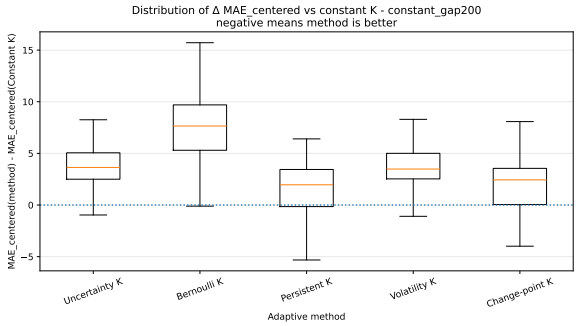



(c) $N = 20$ players, $T = 38$ matchdays.

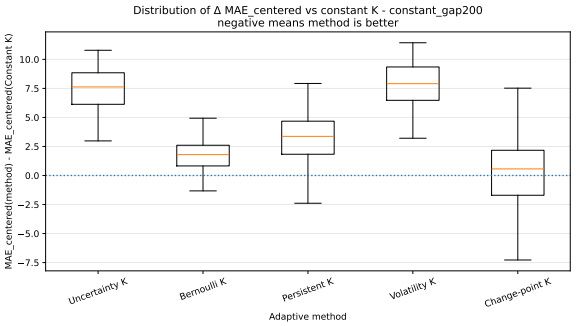



(d) $N = 20$ players, $T = 76$ matchdays.

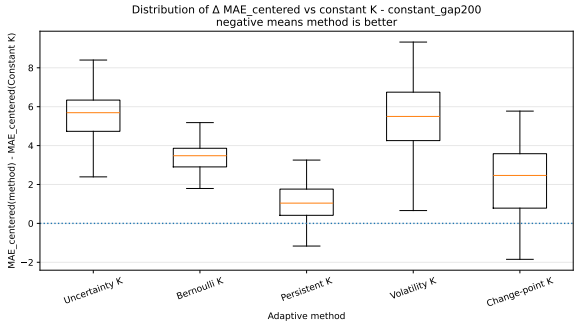
Figure C.8: Distribution of the paired Brier-score differences between the analytical adaptive methods and the constant-$K$ baseline for the `constant_gap200` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.



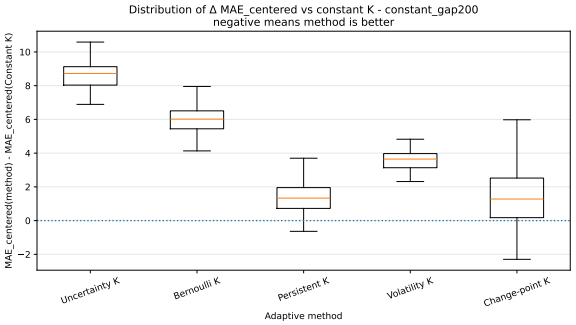

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

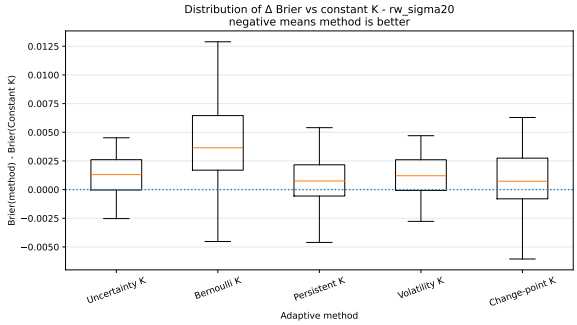



(c) $N = 20$ players, $T = 38$ matchdays.

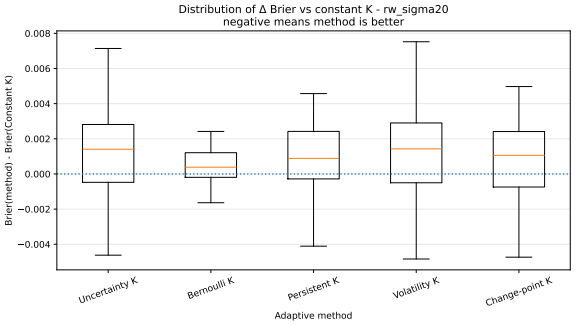



(d) $N = 20$ players, $T = 76$ matchdays.

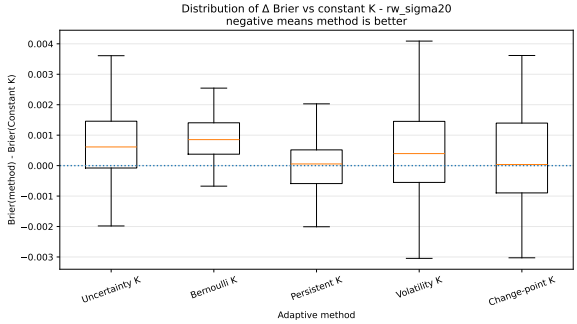
Figure C.9: Distribution of the paired MAE differences between the analytical adaptive methods and the constant-$K$ baseline for the `constant_gap200` scenario. Negative values indicate that the adaptive method achieves a lower MAE than the constant-$K$ baseline.



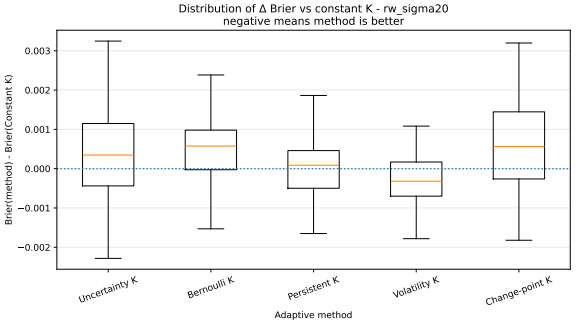

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.

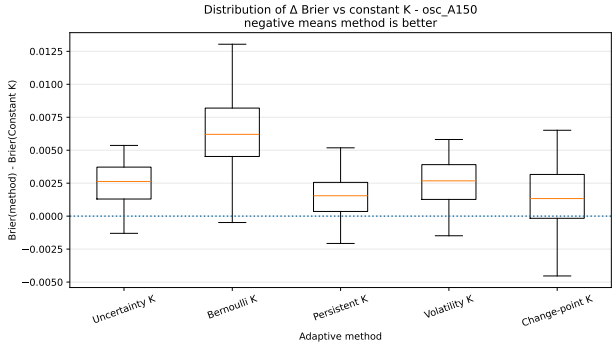



(c) $N = 20$ players, $T = 38$ matchdays.

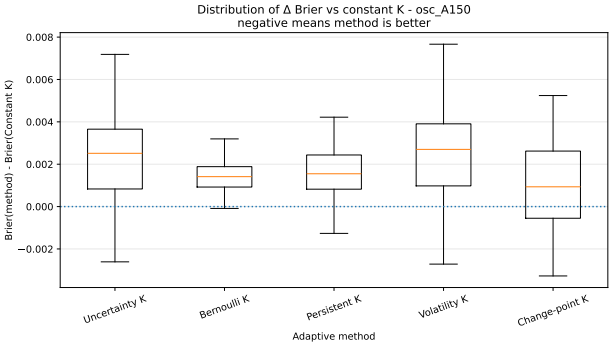



(d) $N = 20$ players, $T = 76$ matchdays.

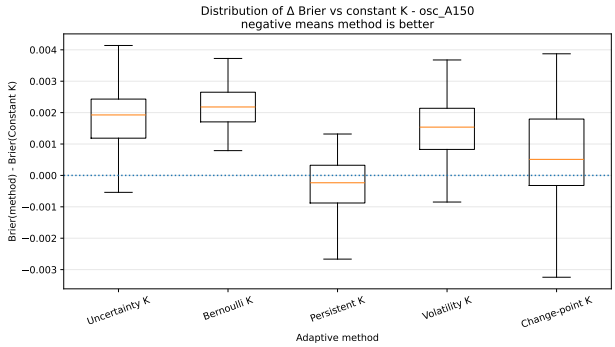
Figure C.10: Distribution of the paired Brier-score differences between the analytical adaptive methods and the constant-$K$ baseline for the `rw_sigma20` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.



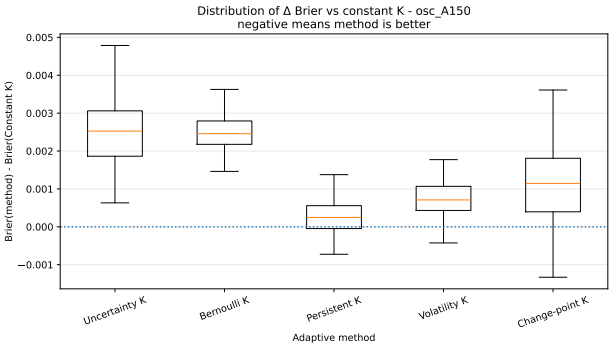

(a) $N = 4$ players, $T = 42$ matchdays.




(b) $N = 4$ players, $T = 78$ matchdays.



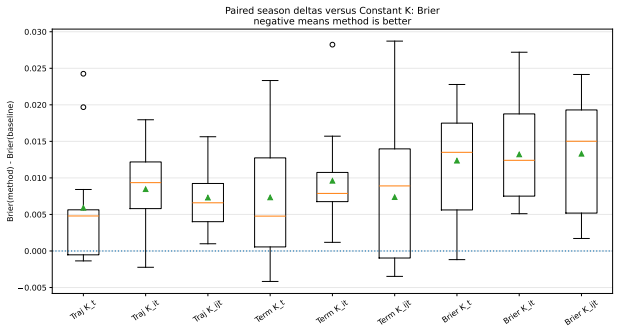

(c) $N = 20$ players, $T = 38$ matchdays.



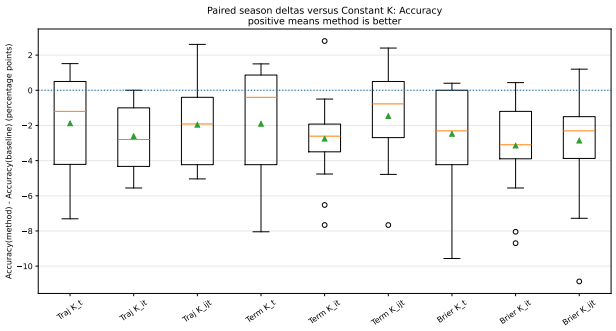

(d) $N = 20$ players, $T = 76$ matchdays.

Figure C.11: Distribution of the paired Brier-score differences between the analytical adaptive methods and the constant-$K$ baseline for the `osc_A150` scenario. Negative values indicate that the adaptive method achieves a lower Brier score than the constant-$K$ baseline.

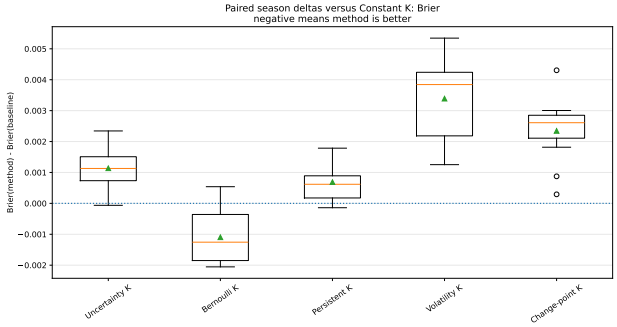
## C.3 Euroleague boxplots



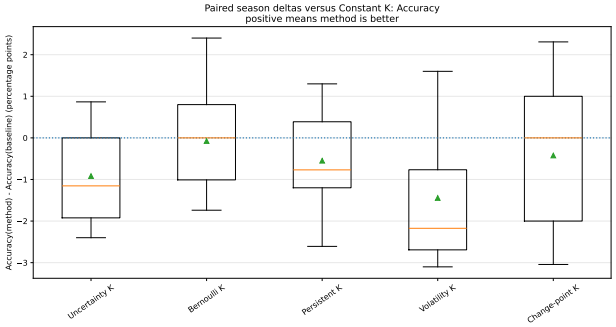

(a) Δ Brier score for numerical methods.




(b) Δ accuracy for numerical methods.




(c) Δ Brier score for analytical methods.




(d) Δ accuracy for analytical methods.

Figure C.12: Distribution of the paired Brier-score and accuracy differences for numerical and analytical adaptive methods applied to the Euroleague dataset.