On Identifying Fast Road Races: Decomposing Race Conditions and Individual Performance Level

Klaus Brinker[1] (✉)

Hamm-Lippstadt University of Applied Sciences, Department Hamm 1, Germany
klaus.brinker@hshl.de

Abstract. We address the question of how to identify fast road races in running by automatically decomposing race results into athlete performance and race condition components. Our approach does not require explicit modeling of influencing factors such as course terrain profiles. Favorable conditions have a substantial impact on race results in road running, and can be critical for meeting championship qualifying standards or for achieving personal bests. We frame this problem as an instance of weighted nonnegative matrix factorization and validate our approach using 6,000 real-world 10k race results from recent local to regional level races. Extensive experiments on both this real-world data and simulated data demonstrate the robustness of this method to high missing value rates and its ability to reduce bias in estimating race conditions compared to mean- or median-based approaches. Our approach also successfully recovered seasonal patterns in race conditions. The number of races and the rate of missing values were found to be the most important properties affecting accuracy, while the number of athletes had less impact.

Keywords: Endurance Sports · Nonnegative Matrix Factorization.

1 Introduction

In many endurance sports, athletic performance is measured in terms of the time required for completing a given course, or the distance covered within a given time. In both cases, a single value summarizes a complex interplay of underlying factors that affect each athlete's overall performance in a particular race. Unraveling running performance has been approached in the literature from a variety of perspectives, among others, using environmental properties, physiological measurements, training statistics, and course characteristics.

We study a related yet distinct problem: distinguishing general race conditions from individual performance levels. General race conditions encompass factors that vary between races but remain approximately constant within each race, such as weather or course topology. In contrast, individual performance levels pertain to characteristics specific to each athlete. This distinction allows us to separately analyze the impact of external race conditions and individual athletic abilities on performance outcomes.

We propose an automatic data-driven method based on solving an associated optimization problem for unpacking these two groups of components. In the most basic model instantiation, we unpack race results as the product of two particular values, one of them linked to the race and the other one to the athlete. Our approach is suitable for endurance sports where *individual performance* is the dominant factor, and components such as team tactics and race dynamics are less important. While this assumption is met for typical road races in running and time trial competitions in cycling, it is not met in cycling road races. Note that naive approaches for quantitatively capturing race conditions, such as mean or median race results, dependent heavily on the athlete performance distribution within races and are easily biased. We provide experimental evidence that our approach is less prone to this problem as it *simultaneously* conducts race and athlete unpacking.

Several interesting applications are facilitated by decomposing race condition and performance level, such as estimating and comparing race results of athletes who did not compete in the same race in a sound manner [11], or computing flat equivalent distances for comparing race courses [12]. If we assume that static race conditions, such as elevation profile or surface properties, are dominant, more favorable race conditions may be identified based on associated condition values for future race schedule planning, among others. More technically, both [11] and [12] build on a general matrix factorization solved by alternating least squares optimization, while our approach is based on nonnegative matrix factorization [7], a matrix decomposition technique which has received increasing attention and has been considered in a variety of fields, among those are astrophysics [10] and bioinformatics [5].

Our main contributions include a practical, entirely data-driven, and parameter-free method for effectively separating race conditions and athlete performance using nonnegative matrix factorization as the core computational technique. This approach is well-suited to real-world race data and handles missing values in a principled way with limited computational demands. We introduce an intuitive and interpretable normalization method for practitioners and compiled a real-world dataset of approximately 6,000 recent 10k race results from local to regional level running road races in Germany (2022-2023) using publicly available data sources. To evaluate the performance of our method and study the impact of specific dataset properties on estimation accuracy, we conducted extensive experiments on both this real-world dataset and an additional simulated one. Furthermore, nonnegative matrix decomposition provides a general data-driven framework for analyzing race properties and athlete profiles without explicit feature modeling.

This paper is structured as follows: We discuss related research from the fields of sports performance science and machine learning in the following section. In Section 3, we introduce our approach both from a technical perspective, including the underlying mathematical optimization problem and the solution strategy based on weighted nonnegative matrix factorization, and from a conceptual perspective with respect to interpreting the decomposition results and

practical applications. We discuss experimental results on real-world road running races and simulation data in Section 4 to provide a comprehensive empirical analysis of the properties of our approach. Finally, in Section 5, we summarize our findings and discuss promising directions for future research.

2 Related Work

A related matrix factorization method has been considered in [11] for *race pace* results in multi-distance running races as an initial processing step. Beyond directly utilizing decomposition values for race pace prediction for known races, the authors employ race characterization vectors as targets in a linear regression model. Here, the inputs are distance and elevation features computed from global navigation satellite system (GNSS) route data. The final combined model allows for race pace prediction for races not included in the initial dataset. While adopting a similar factorization step for race time or covered distance, we focus on race condition estimation for fixed-distance or fixed-time races. Building upon [11], the authors in [13] incorporate a physiological model for adjusting different race distances beyond a straightforward average pace approach. The aim of this approach is to compute hypothetical flat equivalent distances for running routes.

Several studies have addressed the impact of ambient weather conditions on endurance performance in road running, including air temperature [1,8,16], relative humidity [1,16], air pressure [8,16], solar radiation [1], precipitation [8], and wind [8]. In addition to these typically highly variable features, more stable properties correlated with performance have been studied as well, among these are altitude [16] and course terrain profile [14], which includes elevation profile and the number and type of turns. In contrast to this, physiological and technological differences between athletes may be considered as an orthogonal dimension of overall performance. Physiological features include age and sex [17], among the technological ones is advanced footwear technology [9]. In this paper, we do not aim to study individual factors but rather to separate both groups of components in a data-driven manner. This setting allows for a substantially broader application range, as no measurements beyond basic race results are required.

3 Race Result Decomposition

In this section, we propose a model for decomposing endurance race results into two distinct components: those associated with the race itself and those associated with the individual athletes. First, we elaborate on the underlying motivation and assumptions, before we introduce technical details of our decomposition model. Then, we proceed by discussing its implication on analyzing race data.

We focus on endurance race results where individual performance is the predominant factor in races and components, such as team tactics, race dynamics, and drafting, are less important. While this may be considered a reasonable assumption for typical road races in running and time trial competitions in cycling,

among others, it is less so in typical cycling road races. From a data perspective, we consider races where outcomes are measured by a single value, typically by a time or distance measurement. Hence, we start of a straightforward race results collection comprising a set of races and athletes: For each athlete-race combination, this dataset contains either a valid race result or a special flag, if the athlete did not compete in this particular race. This type of dataset can be compiled for a variety of endurance sports easily, and we will present real-world results for regional running road races in the experimental results section.

We assume each athlete's fitness level to remain *approximately* constant for the analyzed period of time, hence, multi-season datasets are less suitable for our approach. Drawing on this assumption, race results from athletes competing in *multiple competitions* can be used for estimating race conditions, as athlete performances in races correlate with more or less favorable race conditions. Obviously, race data from athletes competing in a single race only does not provide meaningful information for this purpose. As races are not deterministic processes, we have to consider a random component which impacts race results and captures incomplete knowledge of the overall process. Starting from a potentially large dataset comprising results from multiple athletes and races, we aim at limiting the impact of this inherent randomness in each individual result.

As noted above, separating race- and athlete-related components in this type of data becomes much simpler once we have access to either one. However, since neither is known *a priori*, the major challenge is to unpack both simultaneously.

3.1 Decomposition Model

Assume we are given an overall number of m athletes competing in n races. We proceed from a $m \times n$ nonnegative real matrix of race results $X \in \mathbb{R}_+^{m \times n}$, with $X_{ij} \geq 0$ storing the outcome for athlete i in race j. In this setting, race outcomes X_{ij} typically represent a time measurement for a given race course, or the distance covered within a fixed period of time on a given race course.

We consider an unsupervised approach by applying nonnegative matrix factorization (NMF) [7] to the race result matrix X to solve the decomposition problem. Hence, we aim at solving the following optimization problem for a given rank d with $1 \leq d \leq \min(m, n)$:

$$\arg\min_{A,\,R} \frac{1}{2} \sum_{ij} (X_{ij} - \sum_{k} A_{ik} R_{kj})^2 = \arg\min_{A,\,R} \frac{1}{2} \sum_{ij} (X - AR)_{ij}^2 \tag{1}$$

where $A \in \mathbb{R}_+^{m \times d}$ and $R \in \mathbb{R}_+^{d \times n}$. Hence, NMF computes an approximation AR of rank d of the race result matrix X:

$$X \approx AR. \tag{2}$$

In practice, we typically need to account for a significant number of missing values in X, as only a subset of athletes competes in a particular given race. Weighted nonnegative matrix factorization (WNMF) [15] is an extension of

NMF, which allows to consider missing data by adding a binary $m \times n$ matrix P in the optimization problem, where $P_{ij} = 0$, if X_{ij} is missing, otherwise $P_{ij} = 1$:

$$\arg\min_{A,\,R} \frac{1}{2} \sum_{ij} P_{ij}(X - AR)^2_{ij}. \tag{3}$$

Note that solutions are not unique as simultaneously permuting columns of A and rows of R, or rescaling both matrices, yields the same objective value.

While it is obvious that the race data matrix X contains only nonnegative values, it is less obvious why we should restrict our solution to nonnegative decomposition matrices A and R. Other decomposition techniques, such as principal component analysis (PCA) and vector quantization (VQ), do not impose this constraint. However, as has been demonstrated in the seminal paper [7], PCA and VQ tend to learn linear combinations that typically involve complex cancellations and where many basis elements lack intuitive meaning. In contrast, in NMF no subtractions can occur and more intuitive, parts-based representations are learned, an appealing property which has led to increasing attention in recent years in several application fields, among those are astrophysics [10] and bioinformatics [5]. From a statistical perspective, solutions of (3) are maximum likelihood estimators for A and R, if we assume additive i.i.d. Gaussian noise with mean 0 and standard deviation σ for race results X_{ij} [2].

From a computational perspective, it was proved in [3] that the optimization problem (3) is NP-hard, even for rank $d = 1$. Computing the global optimum solution is therefore not realistic in general, and we have to resort to approximation algorithms. As a side note, for the rank $d = 1$ case *without* missing data, the global minimizer can be computed in a straightforward manner based on singular value decomposition [4]. For our experiments, we implemented a straightforward multiplicative update approach [6] to compute approximate solutions based on the matrix update rules for WNMF given in [4]. While more sophisticated optimization techniques are available, this straightforward approach proved to be both efficient and robust in our race data experiments.

3.2 Interpreting Decomposition Results

In this section, we will elaborate more on interpreting matrix decompositions for race results. For a given rank value d, the $X \in \mathbb{R}_+^{m \times n}$ race result matrix is factorized into $X \approx AR$, where m and n correspond to the number of athletes and races, respectively. Here, the $m \times d$ matrix A is associated with *athletes*, and more precisely, row k stores a d-dimensional nonnegative characterization for athlete k. Likewise, the $d \times n$ matrix R is linked to *races*, and column l stores a d-dimensional nonnegative characterization of race l. The choice of the rank d is the only hyperparameter in our model and controls the length of the athlete and race characterizations. We focus on the case $d = 1$. Here, the approximation of a race result X_{ij} for athlete i in race j simplifies to the product of two values, $X_{ij} \approx A_i R_j$, where $A_i \stackrel{\text{def}}{=} A_{i1}$ and $R_j \stackrel{\text{def}}{=} R_{1j}$. Therefore, the value A_i can

be interpreted as the basic fitness level for athlete i, while R_j integrates all performance-related properties of race j. These race-related properties typically vary between races, but remain roughly constant for all athletes in the same race. As stated above, typical factors relevant in many endurance races are:

- weather (temperature, humidity, precipitation, wind)
- course terrain profile (elevation profile, surface, number / type of of turns)
- altitude

As noted above, there exists a scaling degree of freedom for solutions of the optimization problem (3). We propose the following average race normalization[1]:

$$R^{\text{avg}} \stackrel{\text{def}}{=} \frac{1}{n} \sum_j R_j \tag{4}$$

$$A_i^{\text{norm}} \stackrel{\text{def}}{=} A_i \cdot R^{\text{avg}} \;\text{ and }\; R_j^{\text{norm}} \stackrel{\text{def}}{=} \frac{R_j}{R^{\text{avg}}} \tag{5}$$

Normalization yields better interpretability of the decomposition results:

$$X_{ij} \approx A_i^{\text{norm}} \cdot R_j^{\text{norm}} \tag{6}$$

While A_i^{norm} represents a reference race result for athlete i assuming average race conditions, R_j^{norm} is normalized such that a value of 1.0 corresponds to *average* race conditions in the dataset.

3.3 Applications

In this section, we elaborate more on applications for the decomposition matrices A^{norm} and R^{norm} and rank value $d = 1$. First of all, R^{norm} allows to compare general race conditions for two races k and l in a principled quantitative manner. For example, suppose that $R_k^{\text{norm}} = 1.03$ and $R_l^{\text{norm}} = 0.99$, then race results in race k will typically be approximately 4% higher compared to l. In contrast to simple race statistics, such as mean or median results, race decomposition approaches are potentially more robust with respect to the athlete performance distribution and allow comparing races with more or less elite level participation as has been noted in [11,12], as long as there is some participation overlap. In the experimental section, we will demonstrate this problem on real-world data.

Conceptually, our approach is of retrospective nature and aims at analyzing past race results, since environmental conditions may change in the future. However, R^{norm} may allow identifying more favorable races, assuming that static race conditions, such as the course terrain profile, are dominant. In [12], a two-step approach is considered for computing hypothetical equivalent flat distances for comparing races, where race result matrix factorization is combined with an elevation profile regression step.

As A^{norm} encodes general athlete fitness levels, we can compare two athletes k and l by their associated values A_k^{norm} and A_l^{norm}, even though they did not

[1] Median-based normalization is another straightforward option.

compete in the same race (see [11]). Beyond ranking athletes by their fitness level, our normalization method allows for a quantitative comparison as fitness levels represent hypothetical race results for an average race. Moreover, assuming that athlete i did not compete in race j, a hypothetical race result can be estimated by $\hat{X}_{ij} \stackrel{\text{def}}{=} A_i^{\text{norm}} \cdot R_j^{\text{norm}}$, which technically is a data imputation approach for missing results. For actual race results, residual values

$$D_{ij} \stackrel{\text{def}}{=} X_{ij} - A_i^{\text{norm}} \cdot R_j^{\text{norm}} = X_{ij} - \hat{X}_{ij}$$

provide some insight into whether the outcome for athlete i in race j is above or below the expected level.

In many endurance sports, scoring systems are used to aggregate results over a series of races. Here, race condition normalization X_{ij}/R_j^{norm} may be considered as a preprocessing step for result aggregation as an alternative to uncalibrated and rank-based approaches.

4 Experimental Results

In this section, we present empirical results of our approach on race data decomposition. Our first study is devoted to real-world running race data, while our second study considers simulated data to provide a more fine-grained analysis of the properties of our approach.

In our experimental studies, we focus on the case rank $d = 1$, where race results are approximated by

$$X_{ij} \approx A_i^{\text{norm}} \cdot R_j^{\text{norm}}.$$

As stated above, normalized matrix decomposition values have a straightforward interpretation for $d = 1$, since A^{norm} represents reference race results for athletes under average race conditions and R^{norm} encodes race-related performance components. As we will discuss below, missing data is a major issue in our real-world experimental study.

As stated in Section 3.1, solving the underlying WNMF problem is computationally very demanding as it is NP-hard. For our experimental study, we used an iterative approximation approach based on multiplicative updates [6] using the WNMF update rules derived in [4] for the weighted sum of squared differences loss function. The matrices A and R were initialized randomly by sampling elements from a uniform distribution over $[0, 1]$. The factorization algorithm stopped once the relative loss change between two subsequent iterations fell below a tolerance value of 10^{-8}. Normalization, i.e., computing A^{norm} and R^{norm}, was conducted as a postprocessing step (see Equations (4) and (5)).

4.1 Real-World Data for Road Running

Data Preparation and Characterization We conducted a study on results from regional road races in running. The dataset consists of publicly available

results [2] from local to regional level 10k road races in Westphalia, a region of northwestern Germany and a regional district of the German Athletics Association (DLV), for the years 2022 and 2023. The raw dataset was compiled on 2024/11/06 using regional district (= *Westfalen*) and race category (= *road race*) as filter criteria. Note that these races are characterized by relatively flat to moderately hilly courses with limited elevation change. Hence, altitude level can be neglected as a relevant performance impacting factor for these races.

The raw dataset contains 6784 individual results for 50 races in 2022 and 9565 results for 52 races in 2023 with valid entries for the considered fields, i.e., result time, name, sex, year of birth. Since this dataset does not include any unique identifier for athletes, such as a national athlete ID, we consider results to originate from the same athlete, if all of the following three records are identical: name, sex, and year of birth. After matching records, we applied the following two-step filtering process to compose a common dataset for all experiments. Step one consists of removing athletes who competed in a single race only, as we assume these records to provide limited information for the decomposition process. In step two, we removed races with less than 10 valid result entries remaining after the athlete removal step.

After preprocessing, the final dataset for 2022 consists of 2302 result records for 904 athletes in 40 races, and 3832 result records for 1309 athletes in 47 races (see Table 1 for more details on the dataset properties). For each year, we composed a m-by-n nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$ for all combinations of athletes and races, where X_{ij} stores the result for athlete i in race j. Note that 93.6% and 93.8% of the matrix entries are missing for 2022 and 2023, respectively, due to the fact that athletes competed in a subset of races only. Hence, dealing with missing data is of particular importance here.

Results on Algorithmic Stability We conducted an initial set of experiments on this dataset to analyze algorithmic initialization sensitivity by solving the decomposition problem for 100 random initializations of A and R. Despite the huge number of missing values, the algorithm was rather robust, as indicated by the *maximum* standard deviation for matrix entries in A^{norm} and R^{norm} of $4.08 \cdot 10^{-1}$ and $8.58 \cdot 10^{-5}$ for 2022, and $9.09 \cdot 10^{-2}$ and $2.96 \cdot 10^{-5}$ for 2023. Moreover, robustness with respect to random initialization can be further increased by reducing the termination tolerance value.

Results on Race-Factors We computed race result decompositions for 2022 and 2023 (detailed results are given in the supplementary section in Tables 8 and 9). The race factors R_j^{norm} are compared with normalized mean and median results, i.e., a normalized mean or median value of 1.0 corresponds to a global dataset mean or median race result.

[2] See `https://ladv.de/westfalen/ergebnisse` or as an alternative source `https://ergebnisse.leichtathletik.de`, as part of the official website of the German Athletics Association.

Table 1: Race dataset properties (after preprocessing).

		Year	
		2022	2023
Results		2302	3832
Races		40	47
Athletes		904	1309
Race results (in minutes)	min	31:03	31:18
	max	1:21:33	1:26:26
	mean	48:33	48:52
	median	47:56	48:16
	SD	8:20	8:36
Result entries (per race)	min	15	10
	max	168	328
	mean	57.5	81.5
	median	49.0	53.0
	SD	38.9	75.7
Result entries (per athlete)	min	2	2
	max	13	14
	mean	2.5	2.9
	median	2.0	2.0
	SD	1.1	1.6

The Pearson correlation coefficients calculated between mean / median race results and R_j^{norm} are 0.509 and 0.432 for the year 2022, and 0.226 and 0.148 for the year 2023. These results indicate a rather limited positive linear correlation between the r-factors and both the mean and median race outcomes, hence, providing initial evidence that these approaches for capturing race-related components are conceptually different.

As most of the race courses in this dataset are characterized by limited elevation change and took place at a similar low altitude level, we hypothesized that weather conditions are of particular importance for differences in race conditions and hence in race results. Therefore, we grouped the races into monthly bins and computed normalized mean and median results, and mean r-factors. There are some important observations to be pointed out in Tables 3 and 4. The overall range of normalized mean and median values is substantially larger. Moreover, we can observe a seasonal pattern for r-factor values both in 2022 and 2023, while this pattern is much less pronounced, if at all visible, for mean and median values. Figure 1 provides bar chart visualizations for this seasonal pattern. With respect to r-factors, the best race conditions could be observed in April and October for 2022, and in February and October for 2023.

Apart from seasonal patterns, race-specific comparisons provide some evidence that r-factor values are less biased with respect to more or less elite level athlete distribution: In 2023, races 9 and 10 overlap as the state road championships (race 10) have been conducted as part of race 9 (see Table 9 in the supplementary section). Here, the mean and median for both athlete subsets

Table 2: Evaluation of imputation error for WNMF-imputation and athlete-based mean and median imputation. Year-wise column-normalized background coloring is added to emphasize differences.

Year	Imputation Method	MAE	MRE
2022	WNMF	106.3	0.0354
	mean	117.6	0.0391
	median	117.9	0.0391
2023	WNMF	107.4	0.0352
	mean	112.9	0.0370
	median	112.7	0.0368

are substantially different (mean: 0.963 vs. 0.846 and median: 0.935 vs. 0.824), presumably due to the fact that the state championship (sub-)race comprises a more competitive athlete subset. When interpreted as a race condition characterization, mean and median values are obviously misleading here, as both races have been conducted simultaneously on the same course. In contrast, the r-factor values for these races are considerably less impacted by the athlete performance distribution (1.006 vs. 1.000).

Imputation Results We conducted an additional experiment focusing on data imputation only. In each subexperiment, we removed a single valid result entry X_{ij} from the data matrix, computed both decomposition matrices A^{norm} and R^{norm} using the remaining entries, and estimated this additional missing entry (see (6)). We compared each estimated entry with the ground truth one using the well-known mean absolute error (MAE) and the mean relative error (MRE) measures as evaluation measures, and averaged the results over all subexperiments for non-missing result entries in X. As alternative imputation methods, we considered athlete-based mean and median value imputation.

The imputation results are shown in Table 2, where for both 2022 and 2023, mean and median achieve comparable estimation accuracy, while WNMF-based imputation is the most accurate method in this evaluation. As stated in the dataset properties Table 1, the mean number of result entries per athlete is 2.5 and 2.9, respectively. Moreover, the mean standard deviation per athlete is 71.0 and 73.3, respectively. Hence, considering this typically very small number of highly variable data points per athlete (out of which one is masked out in the above stated evaluation process), we should not expect a high level of accuracy when estimating individual results. In a secondary evaluation, we confirmed that imputation accuracy generally increases for all methods with the number of results available for the considered athlete.

4.2 Simulated Data

In addition to experiments on real-world data, we conducted a series of experiments on simulated race data in order to analyze the algorithmic sensitivity

Table 3: Normalized mean, median, and r-factors for races grouped by month for 2022. Note that the dataset does not include outdoor race results for 1/2022. Column-normalized background coloring is added to emphasize differences.

Month	Athletes	mean	median	**r-factor**
2	204	1.017	1.020	1.008
3	256	0.957	0.952	0.996
4	169	1.030	1.039	0.990
5	180	0.988	0.993	0.992
6	102	0.980	0.974	1.003
7	235	1.032	1.043	1.016
8	156	1.018	1.009	1.016
9	273	1.049	1.046	1.014
10	252	0.963	0.968	0.985
11	182	1.004	0.992	0.992
12	293	0.972	0.974	1.004

Table 4: Normalized mean, median, and r-factors for races grouped by month for 2023. Column-normalized background coloring is added to emphasize differences.

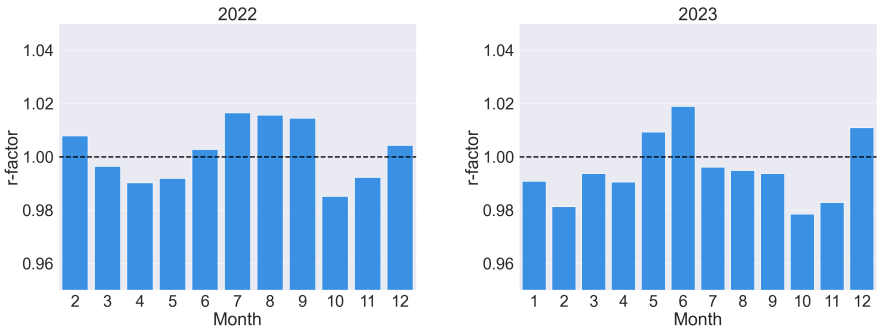
Month	athletes	mean	median	**r-factor**
1	507	1.010	1.009	0.991
2	212	1.013	1.013	0.981
3	239	0.983	0.979	0.994
4	201	1.040	1.048	0.990
5	308	1.022	1.014	1.009
6	260	1.001	0.977	1.019
7	409	1.021	1.031	0.996
8	820	0.969	0.966	0.995
9	353	1.031	1.047	0.994
10	176	0.930	0.925	0.978
11	63	0.953	0.969	0.983
12	284	1.010	1.005	1.011

with respect to the missing value rate, the number of races, and the number of athletes. The conceptual difference between real-world data and simulated data is that the data-generating process is known for the latter one, and more specifically the ground truth matrix entries $\hat{A}_i^{\text{norm}}$ and $\hat{R}_j^{\text{norm}}$. Hence, we are able to quantitatively evaluate the accuracy of our approach in estimating these matrices given noisy race data $\hat{X}$.

The ground truth athlete-related entries $\hat{A}_i^{\text{norm}}$ were randomly sampled from a uniform distribution over $[1800, 2700]$ corresponding to reference race times between 30 min and 90 min. The race-related entries $\hat{R}_j^{\text{norm}}$ were randomly sampled from a uniform distribution over $[0.9, 1.1]$. Both choices are based roughly on the observed real-world value ranges. Race results $\hat{X}_{ij}$ were computed as

$$\hat{X}_{ij} = \hat{A}_i^{\text{norm}} \cdot \hat{R}_j^{\text{norm}} + \epsilon_{ij},$$

Fig. 1: Average r-factors for races grouped by month for 2022 and 2023. Note that the dataset does not include outdoor race results for 1/2022. The reference line corresponds to the mean r-factor among all races for the considered year.



where the additive noise components ϵ_{ij} were randomly sampled from a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 70$.

We used the following approach for simulating missing values in $\hat{X}$: In each row, one entry is randomly selected as a valid data element (which corresponds to the requirement that each athlete has to compete in at least one race). Likewise, we randomly select one entry in each column to be a valid data element (which corresponds to the requirement that each race is associated with at least one result). Note that these two steps are conducted independently of each other. All remaining entries are randomly assigned to the valid and missing data categories in order to match the required overall missing data rate. For each choice of parameters, we computed average MAE and MRE values over 100 repeated simulation runs.

Missing Values In a first experimental evaluation, we fixed the number of athletes (= 1000) and the number of races (= 50), while varying the fraction of missing values from 0.0 to 0.8 in equal steps of 0.1, and between 0.90 and 0.97 with a step size of 0.01. Note that there is a limit for the fraction of missing values of 0.97 here, as for 0.98, which corresponds to 1000 missing values, all missing entries would already be required for ensuring the minimum valid data constraint for athletes only.

The results shown in Table 5 indicate that up to a missing value fraction of roughly 0.9 there is a continuous, but limited increase in the estimation error measures for A, while for R the estimation error is at a rather stable level up to a missing value fraction of 0.96. Then, all error measures increase substantially at the final missing value fraction of 0.97.

Number of Athletes A second experimental evaluation is devoted to the number of athletes. Here, we fixed the number of races (= 50) and the missing value

Table 5: Error evaluation (MAE, MRE) for A^{norm} and R^{norm} using simulated race data with varying missing value rates. Column-normalized background coloring is added to emphasize differences.

missing	A^{norm}		R^{norm}	
values	MAE	MRE	MAE	MRE
0.00	18.4	0.0082	0.0075	0.0075
0.10	18.0	0.0081	0.0072	0.0072
0.20	17.3	0.0077	0.0067	0.0067
0.30	17.5	0.0078	0.0067	0.0067
0.40	16.5	0.0074	0.0059	0.0059
0.50	17.4	0.0078	0.0059	0.0059
0.60	20.0	0.0090	0.0071	0.0071
0.70	19.7	0.0088	0.0061	0.0061
0.80	23.1	0.0103	0.0066	0.0066
0.90	30.4	0.0137	0.0069	0.0070
0.91	31.8	0.0143	0.0067	0.0067
0.92	33.7	0.0152	0.0069	0.0069
0.93	36.6	0.0165	0.0077	0.0077
0.94	38.0	0.0171	0.0068	0.0068
0.95	42.0	0.0189	0.0077	0.0076
0.96	46.8	0.0211	0.0085	0.0085
0.97	53.4	0.0240	0.0104	0.0104

rate ($= 0.9$), while the number of athletes varied between 100 and $10,000,000$ with a multiplicative step factor of 10. Note that our missing value model requires the number of athletes to be greater than 50.

The results shown in Table 6 indicate that the error measures are at a rather stable level over a very large range of the number of athletes, except for the initial number of 100 athletes.

Number of Races A third experimental evaluation is devoted to the number of races. Here, we fixed the number of athletes ($= 1000$) and the missing value rate ($= 0.9$) while varying the number of races between 20 and 100 with a step size of 10. Note that our missing value model requires the number of races to be greater than 10.

The results given in Table 7 show a continuous error decrease with an increasing number of races. While for 20 and 30 races, the associated error values are comparatively high, error values decrease at a substantially lower rate starting at 40 .

Discussion In our experimental evaluations on simulated data, the missing value rate and the number of races had the largest impact on the considered error measures. For the missing value rate, the impact on A^{norm} seems to be more pronounced, while the estimation accuracy was more stable for R^{norm}.

Table 6: Error evaluation (MAE, MRE) for A^{norm} and R^{norm} using simulated race data with varying numbers of athletes. Column-normalized background coloring is added to emphasize differences.

	A^{norm}		R^{norm}	
athletes	MAE	MRE	MAE	MRE
100	32.0	0.0144	0.0114	0.0114
1000	30.5	0.0137	0.0070	0.0070
10000	30.9	0.0139	0.0067	0.0067
100000	31.0	0.0139	0.0069	0.0069
1000000	30.2	0.0136	0.0064	0.0064
10000000	30.6	0.0137	0.0065	0.0065

Table 7: Error evaluation for simulation race data with varying numbers of races. Column-normalized background coloring is added to emphasize differences.

	A^{norm}		R^{norm}	
races	MAE	MRE	MAE	MRE
20	49.0	0.0220	0.0103	0.0104
30	40.6	0.0182	0.0090	0.0090
40	35.0	0.0157	0.0081	0.0081
50	31.2	0.0140	0.0075	0.0076
60	27.4	0.0123	0.0062	0.0062
70	25.8	0.0116	0.0064	0.0064
80	23.5	0.0105	0.0055	0.0055
90	22.6	0.0101	0.0056	0.0056
100	21.0	0.0094	0.0052	0.0052

The impact of the number of races on A^{norm} and R^{norm} seems to be on a similar scale.

In contrast to these findings, there seems to be only a minor impact of the number of athletes on the estimation performance, as suggested by our experiments which cover a large range of values.

5 Conclusions

We propose a novel data-driven approach for separating two essential groups of components from a set of race results: Race conditions and individual performance level. Based on nonnegative matrix factorization for dimensionality reduction, we unpack results into race- and athlete-related quantitative characterizations. Their dimensionality is the only hyperparameter of our method.

We focus on the one-dimensional case in our experimental evaluation, where the normalized values associated with athletes and races have a straightforward interpretation: The athlete component represents a reference race result assuming average conditions and hence a quantification of the individual performance

level, while the race value integrates general factors such as weather and course terrain profile.

In the experimental section, we conduct a series of evaluations on real-world race data for local to regional level 10k road running races and simulated race data. We demonstrate that our approach is a practical method for real-world data, and in particular is a mathematically sound and well-suited method for dealing with a high missing value rate. We show that our method provides more robust estimates for race conditions with respect to the particular athlete distribution competing in a race, while simple mean- or median-based techniques suffer from obvious drawbacks and compute biased race estimates. Moreover, we were able to recover seasonal patterns from race-data only, which is consistent with the fact that weather conditions are of particular importance for the considered races which are characterized by limited elevation change and a similar altitude level. A series of experiments on simulated race data suggests that our approach provides stable estimates over a wide range of dataset properties. More precisely, in our experiments, the missing value rate and the number of races had a larger impact on the estimation accuracy, while the number of athletes had less impact. The computational complexity of our approach is rather low and should not be a limiting factor for typical applications.

Unpacking race conditions and individual performance levels provides a variety of interesting applications, such as comparing results of athletes who did not compete in the same race in a sound quantitative manner, comparing past race conditions and potentially identifying favorable future ones, and scoring systems for race series.

We focused on the case $d = 1$, where all components related to particular races and athletes were integrated into associated scalar values. However, for $d > 1$ this approach provides a framework for computing multidimensional characterizations as well, which, due to the nonnegativity constraint (see Section 3.1), facilitates feature interpretability. From a conceptual and terminological point of view, it is more appropriate to refer to multidimensional race and athlete characterizations as race condition and athlete profiles rather than race condition values and athlete performance levels. Then $\sum_k A_{ik} R_{kj}$ can be interpreted as matching race and athlete *profiles* against each other. It will be interesting to analyze and potentially link individual components of these profiles to known performance factors or explore novel ones. These more fine-grained characterizations may provide an interesting starting point for better understanding the strengths and weaknesses of athletes with respect to race conditions, and allow for better matching those to specific race demands in the future.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ely, M.R., Cheuvront, S.N., Roberts, W.O., Montain, S.J.: Impact of weather on marathon-running performance. Medicine and science in sports and exercise **39**(3),

487–493 (2007)
2. Gillis, N.: Nonnegative Matrix Factorization. Society for Industrial and Applied Mathematics, Philadelphia, PA (2020)
3. Gillis, N., Glineur, F.: Low-rank matrix approximation with weights or missing data is np-hard. SIAM Journal on Matrix Analysis and Applications **32**(4), 1149–1165 (2011)
4. Ho, N.D.: Nonnegative Matrix Factorization - Algorithms and Applications. Ph.D. thesis, Université catholique de Louvain (2008)
5. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Mohan, R., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., Regev, A.: Identifying disease-critical cell types and cellular processes by integrating single-cell rna-sequencing and human genetics. Nature Genetics **54**(10), 1479–1492 (2022)
6. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Leen, T., Dietterich, T., Tresp, V. (eds.) Advances in Neural Information Processing Systems. vol. 13. MIT Press (2000)
7. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
8. Nikolaidis, P.T., Di Gangi, S., Chtourou, H., Rüst, C.A., Rosemann, T., Knechtle, B.: The role of environmental conditions on marathon running performance in men competing in boston marathon from 1897 to 2018. International journal of environmental research and public health **16**(4), 614 (2019)
9. Ortega, J.A., Healey, L.A., Swinnen, W., Hoogkamer, W.: Energetics and biomechanics of running footwear with increased longitudinal bending stiffness: a narrative review. Sports Medicine **51**(5), 873–894 (2021)
10. Ren, B., Pueyo, L., Zhu, G.B., Debes, J., Duchêne, G.: Non-negative matrix factorization: Robust extraction of extended structures. The Astrophysical Journal **852**(2), 104 (jan 2018)
11. de Smet, D., Verleysen, M., Francaux, M.: Running race times prediction and runner performances comparison using a matrix factorization approach. In: Proceedings of the 5th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2017). pp. 96–101 (01 2017)
12. de Smet, D., Verleysen, M., Francaux, M., Baijot, L.: Long-distance running routes' flat equivalent distances from race results and elevation profiles. In: Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support - Volume 1: icSPORTS. pp. 56–62. INSTICC, SciTePress (2018)
13. de Smet, D., Verleysen, M., Francaux, M., Baijot, L.: Long-distance running routes' flat equivalent distances from race results and elevation profiles. In: Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2018). pp. 56–62 (09 2018)
14. Snyder, K.L., Hoogkamer, W., Triska, C., Taboga, P., Arellano, C.J., Kram, R.: Effects of course design (curves and elevation undulations) on marathon running performance: a comparison of breaking 2 in monza and the ineos 1:59 challenge in vienna. Journal of Sports Sciences **39**(7), 754–759 (2021)
15. Srebro, N., Jaakkola, T.: Weighted low-rank approximations. In: Proceedings of the 20th international conference on machine learning (ICML). vol. 3, pp. 720–727 (2003)
16. Wang, S., Gao, M., Xiao, X., Jiang, X., Luo, J.: Wasted efforts of elite marathon runners under a warming climate primarily due to atmospheric oxygen reduction. npj Climate and Atmospheric Science **7**(1), 97 (2024)
17. Zavorsky, G.S., Tomko, K.A., Smoliga, J.M.: Declines in marathon performance: Sex differences in elite and recreational athletes. PLOS ONE **12**(2), 1–17 (02 2017)