ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

MAP Best Performances Prediction for Endurance Runners

Dimitri de Smet[1], Marc Francaux[2],
Laurent Baijot[3] and Michel Verleysen[1]

1- UCLouvain, ICTEAM, Louvain-la-Neuve - Belgium
2- UCLouvain, IoNS, Louvain-la-Neuve - Belgium
3- Formyfit - Enghien - Belgium

Abstract. The preparation of long-distance runners requires to estimate their potential race performances beforehand. Athlete performances can be modeled based on their past records, but the task is made difficult because of the high variability in runner race performances. This paper presents a *maximum a posteriori* (MAP) estimation that addresses the issues related to this high variability. The inclusion of athlete priors and a specific residual model are inferred with the help of a large set of race results.

1 Introduction

Long-distance runs involve aerobic efforts that require endurance as well as mental strength. They range from eight kilometers to hundreds of kilometers. Proper athlete preparation for a specific race must take into account his expected performance [2]. It is, therefore, useful to predict the latter accurately. This work aims at predicting the best performance an athlete can expect on any race based on his previous race results. This work focuses on modeling specifically best performances rather than any past performances because this is what is of primary interest in the scope of athlete preparation.

One of the most popular models for runner performances describes the relationship between a runner average speed and the total race distance as a law that contains two athlete-specific parameters [1]. Using such a model, a set of past race performances can be used to fit athlete parameters and predict expected performances on any race of a given length but a couple of issues need to be addressed.

First, races cannot be strictly summarized by their length: gradient of ascent, weather conditions, altitude, vegetation, uneven ground and ground firmness will affect athlete speeds. Nevertheless, fitting athlete parameters is still possible using the notion of *equivalent distances* described in [3]. This allows dealing with races fully characterized by only one parameter: all conditions being summarized by the *equivalent distance*.

Secondly, it is common, especially for casual runners, to experience high variability in performances. For instance, they may attend races in a group, barely prepared, or get hurt. If the task at hand is to unveil athletes potential, it is required to model how performances may deviate from it. For this purpose, this paper introduces an error model that reflects the distribution of athlete

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

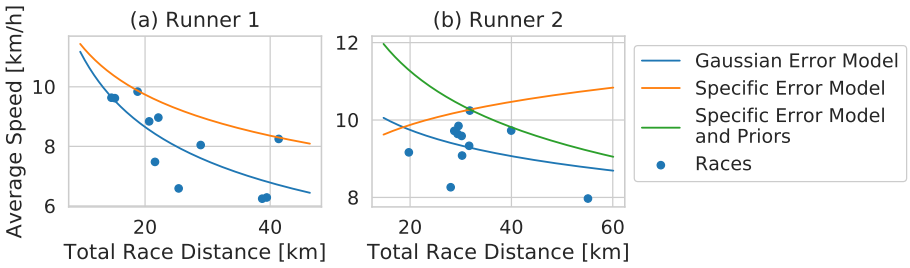

Fig. 1: Problem illustration with two examples.

performances with respect to the best ones. Fig. 1(a) shows in blue, a standard regression, and in orange, an example of best performances curve that would be obtained considering such a specific error model.

Thirdly, the high variability in observed performances can make the solution to the regression problem physiologically unlikely; for instance, average speeds that increase with the races total distances (as the orange curve in Fig. 1 (b)). This third issue is addressed by setting prior assumptions on athlete parameters that reflect what is more likely to be observed. These assumptions take the form of probability distributions characterized by parameters that are inferred from a large set of race results. Athletes curve fitting that considers the error model and athlete priors is performed using a *maximum a posteriori* estimation (MAP) (green curve in Fig. 1 (b)).

The methodology is applied to a set of 736 athletes who all attend to, at least, ten races in a period of two years. A special procedure is applied for the method validation because it needs to reflect the ability to predict athletes best race results and not the ability to predict any of their race results.

2 Data Source

The methodology is applied on official races results sampled on a large variety of running races organized in Belgium during a period of two years and ranging from 8 to 159 kilometers. For each of the races in the data set, the *equivalent distance* must first be computed. This is performed with the methodology briefly discussed in Section 4 using a set of 106,172 race times (264 races, 29,337 athletes). Subsequently, having a set of athletes and their race results on races for which *equivalent distances* are known, a MAP estimation is computed for a smaller set of athletes to model their performances. It was chosen to keep only runners with ten or more race results to ensure proper validation; this condition is met for 736 athletes.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

3 Best Performances Model

According to the model discussed in [1], an athlete a is expected to run race r with an average speed $s_{a,r}$ that depends on the race distance d_r and two athlete-specific parameters $(a_{a,0}, a_{a,1})$:

$$log(s_{a,r}) = a_{a,1} \cdot log(d_r) + a_{a,0} + \epsilon \tag{1}$$

where ϵ refers to an additive error term modeled as a random variable that is discussed in Section 5.1. The expression can be re-written as $s_{a,r} = e^{a_{a,0}} \cdot d^{a_{a,1}} \cdot e^{\epsilon}$ which shows that the additive error in the *log* domain becomes a multiplicative error term with transformed probability distribution in the speed domain.

4 Equivalent Distances

The hypothesis that is made here is that race parameters that affect the running speed can be summarized with only one parameter per race; namely the *equivalent distance*. The problem of identifying the *equivalent distance* for each races was addressed in [3]. Shortly, under the above hypothesis, equivalent distances can be assigned to races so that the average squared deviation to Equation (1) is minimized. It follows that races *equivalent distances* and athlete parameters can both be identified by solving the optimization problem

$$\underset{\mathbf{A},\mathbf{D}}{\arg\min} \sum_{(a,r)\in\Omega} (a_{a,1} \cdot log(d_r^{eq}) + a_{a,0} - log(s_{a,r}))^2, \tag{2}$$

$\mathbf{A}$ being the matrix of all athlete parameters $(a_{a,0}, a_{a,1})\forall a$, $\mathbf{D}$ the set of all races *equivalent distances* d_r^{eq} and Ω the set of index pairs (a, r) for which race performances $(s_{a,r})$ are observed (athlete a attended race r).

The solution to this equation provides athlete parameters and equivalent distances but only the equivalent distances are further used. As mentioned in the introduction, the athlete parameters that are discussed in the next sections have a different purpose: predict athlete best performances using a MAP estimation. The athlete parameters discussed in this section are more rough approximation but are sufficient to compute equivalent distances thanks to the very large number of athlete that are used (several tens of thousands).

5 MAP Athlete Parameters Estimation

The problem at hand is to find the two parameters $\mathbf{a}_a = (a_{a,0}, a_{a,1})$ in Equation (1) that characterize the potential performances of athlete a, knowing all his race results. Athlete performances are summarized for each race event r by the athlete average speed $s_{a,r}$ and the race equivalent distance d_r^{eq}. In the following, $data_a = \{s_{a,r}\}$ denotes the data of athlete a.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

A first approach can assign athlete parameters $\mathbf{a}_a$ so that they maximize the likelihood to observe what is actually observed for the set of races Ω_a attended by athlete a. This is known as the *maximum likelihood*(ML) estimation:

$$\begin{aligned}\hat{\mathbf{a}}_a^{ML} &= \arg\max_{\mathbf{a}_a} p(data_a|\mathbf{a}_a) = \arg\max_{\mathbf{a}_a} \prod_{r\in\Omega_a} p(log(s_{a,r})|\mathbf{a}_a) \\ &= \arg\max_{\mathbf{a}_a} \prod_{r\in\Omega_a} f_\epsilon(a_{a,1}\cdot log(d_r^{eq}) + a_{a,0} - log(s_{a,r}))\end{aligned} \tag{3}$$

with f_ϵ being the error probability density function associated with the random variable ϵ in Equation(1). It reflects how athlete performances deviate from their best performance curve. If normal distribution is assumed, solving (3) is equivalent to solve a linear regression with the least square error criteria. The linear regression displayed in blue in Fig. 1 (a) does not reveal athlete best performances. This motivates the need for a specific probability density function f_ϵ that is presented in Section 5.1.

As mentioned in the introduction, the high variability in observed performances can make the regression model nonsensical. This issue is addressed by including prior assumptions on athlete parameters. The Bayesian theorem can be used to reformulate our optimization problem to include prior assumptions $f_{\mathbf{a}}$ on athlete parameters. This second approach selects athlete parameters $\mathbf{a}_a$ at the maximum of their probability density function (the mode) knowing what is observed. This is known as the *maximum a posteriori* (MAP) estimation:

$$\begin{aligned}\hat{\mathbf{a}}_a^{MAP} &= \arg\max_{\mathbf{a}_a} p(\mathbf{a}_a|data_a) = \arg\max_{\mathbf{a}_a} \frac{p(data_a|\mathbf{a}_a)\cdot f_{\mathbf{a}}(\mathbf{a}_a)}{p(data_a)} \\ &= \arg\max_{\mathbf{a}_a} \prod_{r\in\Omega_a} f_\epsilon(a_{a,1}\cdot log(d_{eq,r}) + a_{a,0} - log(s_{a,r}))\cdot f_{\mathbf{a}}(\mathbf{a}_a).\end{aligned} \tag{4}$$

The second probability distribution function $f_{\mathbf{a}}$ is discussed in Section 5.2.

5.1 Error Model

The error model is chosen such that the Equation (1) describes the athlete best performances. The error model must, therefore, follow a probability distribution that reflects how athlete performances deviate from expected best ones. It follows that the random variable ϵ must be allowed to take only negative values because an athlete can only under-perform his best possible race results. The gamma distribution (taken with a minus sign) shown in Fig. 2 provides the expected properties. The multiplicative error term, in the speeds domain, ranges from 0 to 1 and peaks close to 1 as shown in the same figure. The gamma distribution has two parameters: a shape parameter k and a scale parameter θ. They will be chosen to maximize the prediction accuracy as discussed in Section 5.4.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

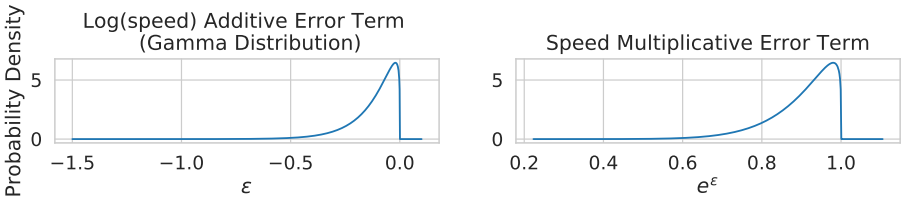

Fig. 2: The error probability distribution function shows how athletes deviate from their best performance curve.

5.2 Athlete Priors

Athlete parameters have a limited range of allowed values by nature: for instance, most athletes are not likely to outperform world records and average speeds should decrease as the total race distance increases ($a_{a,0}$ must be negative for all athletes). Setting athlete priors aims at *guiding* the solution towards the more likely ones. Gaussian priors are assumed for both athlete parameters $(a_{a,0}, a_{a,1})$. The assumption on athlete parameters distribution are, therefore, conditioned by four parameters (mean and standard deviation for both parameters: (μ_0, σ_0) and (μ_1, σ_1)).

5.3 Athlete Parameters Fitting

Assuming that the priors and the error distributions are known, solving Equation (4) gives athlete parameters that will allow to predict his best performance for any given long-distance running race. As the second line of the equation can be evaluated for any set of athlete parameters, a simplex optimization method can be used to identify the ones that maximize the probability density. This optimization problem is solved for each athlete independently with the same distribution parameters.

5.4 Validation and Distribution Parameters Fitting

The validation of the methodology requires to evaluate the ability to predict *best* race performances. For this purpose, *best* race performances are defined in terms of the ratio between the race average speed and the world best performance expected on the same distance. World best performances are modeled using a simple linear regression described by Equation (1) with official track running world records that are ratified by the International Association of Athletics Federations [4]. With the given definition of *best* performances, for each athlete, one of his two best performances is randomly selected and kept aside for validation. The other one stays in the data-set for the MAP regression. The accuracy of the methodology is computed with the root mean squared error on the validation races that were not used for athlete parameters fitting.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Athlete parameters fitting requires that both error distribution and athlete priors are known: formally it means that 6 distribution parameters have to be identified (two per athlete parameter and two for the gamma distribution of the error model). As a large set of athlete race results is at disposal, this operation is performed by iterating in the distribution parameters space and keeping the set of distribution parameters that gives the best accuracy on the validation set (race results that were not used to fit athlete parameters).

6 Results and Conclusion

The methodology allows predicting best performances with an accuracy that is sufficient to plan casual runners workout sessions.

This paper presents a methodology that aims at predicting best running performances for an athlete on any given long-distance race. Two issues, both related to the high variability of the athlete race records are addressed. The first one is that athlete performances deviate a lot from their potential and only negatively. This is addressed by the inclusion of a gamma-distributed error term in the log-speed domain that leads to a multiplicative error term ranging from 0 to 1 in the speed domain. The second one is that the variability in the performances can be such that sound regression cannot be inferred from the athlete data alone. This is addressed by the inclusion of Gaussian prior on athlete parameters derived from 736 athletes: both distribution parameters (error and prior) were empirically selected by maximizing the prediction accuracy.

This work is of interest for web companies that provide running guidance inferred from massive runner data. Such companies own the necessary data to estimate distribution parameters. Moreover, athlete parameters prior distribution could be individualized based on athlete data (age, sex, weight, etc.).

References

[1] García-Manso, J., Martín-González, J., Vaamonde, D., Da Silva-Grigoletto, M.: The limitations of scaling laws in the prediction of performance in endurance events. Journal of theoretical biology **300**, 324–329 (2012)

[2] Midgley, A.W., McNaughton, L.R., Jones, A.M.: Training to enhance the physiological determinants of long-distance running performance. Sports Medicine **37**(10), 857–880 (2007)

[3] de Smet, D., Verleysen, M., Francaux, M., Baijot, L.: Long-distance running routes' flat equivalent distances from race results and elevation profiles. In: 6th International Congress on Sport Sciences Research and Technology Support (2018)

[4] The International Association of Athletics Federations: World Records. `https://www.iaaf.org/records/by-category/world-records` (2018), [Online; accessed 27-July-2018]