Analyzing cycling sensors data through ordinal logistic regression with functional covariates

Julien Jacques, Sanja Samardžić

Analyzing cycling sensors data through ordinal logistic regression with functional covariates

Julien Jacques
Université de Lyon, Lyon 2, ERIC UR 3083
and
Sanja Samardžić
Université de Lyon, Lyon 2, ERIC UR 3083

December 15, 2020

Abstract

With the emergence of numerical sensors in sports, all cyclists can now measure many parameters during their effort, such as the speed, the slope, the altitude, their heart rate or their pedaling cadence. The present work studies the effect of these parameters on the average developed power, which is the best indicator of the cyclist performance. For this, a cumulative logistic model for ordinal response with functional covariate is proposed. This model is shown to outperform the competitors on a benchmark study, and its application on cyclist data confirms that the pedaling cadence is a key performance indicator. But maintaining a high cadence during long effort is a typical characteristic of high level cyclists, which is something on which amateur cyclists can work on to increase their performance.

Keywords: cycling sensor data; ordinal data; functional data; cumulative logistic regression

1 Data and motivation

With the emergence of numerical sensors in sports, there is an increasing need for tools and methods to analyse the produced data. Cycling is not an exception, with its many professional but low costly devices available to amateur cyclists Bini et al. (2014). In cycling, the most frequently produced and used data are the speed, the slope, the altitude, the heart rate of the cyclist and its pedaling cadence. Left panel of Figure 1 illustrates such data for a one-hour bike session. At a slightly higher cost, power sensors also make it possible to measure the instantaneous power developed during the activity (right panel of Figure 1). Thus, any cyclist can contemplate the data produced following a training or a competition. But apart from contemplating this data and publishing them on social media, what else can be done with it? Is it possible to use them to improve performance? There is a large literature on the subject Grappe (2018); van Dijk et al. (2017), not always accessible to uninitiated amateur cyclists, who may nevertheless be interested in improving their performance.

If the reference indicator was the heart rate in the last century, the latter has been dethroned in recent years by the power developed by the cyclist. Indeed, if the heart rate can be distorted by external elements such as the weather, the heart rate being positively correlated with the temperature, this is not the case with the power which is to date the best indicator of the performance of the cyclist Beattie et al. (2016); Grappe (2012). Right panel of Figure 1 plots the power developed by the cyclist during the same bike session as the one corresponding to the left panel. We can notice in this figure that power data are highly irregular, and cyclists rarely use the precise value of the power developed during the effort. Cyclists are used to working with *power zones*, defined as a set range of watts, calculated on the basis of percentages of the Functional Threshold Power (FTP)

Borszcz et al. (2018); Grappe (2012). Several definitions of theses ranges exists, and those automatically calculated by the device used to collect the data are used in this work. The limits of the 7 power zones for the cyclist whose data are plotted on Figure 1 are represented by the horizontal lines on the right panel of the figure.

If power is the best cyclist's performance indicator, it is necessary to seek to optimize it during the effort. For a cyclist with a fixed and limited capacity, several parameters could help him to optimize the power during the effort. In particular, the cyclist can easily act on its pedaling cadence, which is known to be a parameter influencing significantly the developed power Faria et al. (2005). But here again, knowledge in terms of cadence to be developed has changed a lot in recent years. You just have to watch videos of climbing a pass in the 1980s and now to see that the pedaling cadence is absolutely not the same. If in the 1980s cyclists sought to use the biggest gear, by working essentially on the force that they were able to develop, the paradigm evolved during the 2000s. Indeed, using a smaller force ally with a higher pedaling cadence results in a much better final performance. And the final performance indicator that brought this to light is the developed power Abbiss et al. (2009). Thus, all professional cyclists now use very high pedaling cadences, especially during long climbs where it is essential to maintain the highest possible power during the entire climb Nimmerichter et al. (2011). But what about the amateur cyclists? Is it also possible for them to reproduce these cadences? Will an higher cadence increase the power that a cyclist is able to develop, and therefore his performance? And whatever is the effort length?

The goal of the present study is to give some answers to these questions, at least for the cyclist who produced the analyzed data. In particular, we want to exhibit which levers of action are available for the cyclist in his practice in order to optimize the average power that

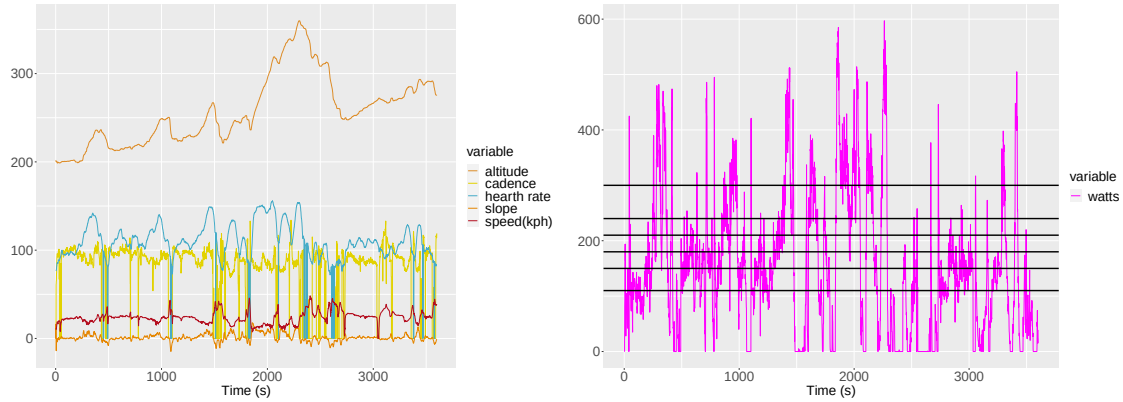

Figure 1: Raw cycling data (left) and power (right)

he is able to maintain during the activity. And this, for different duration of the activity. For this, the proposed approach rely on a modeling of the mean power zone according to the speed, the slope, the altitude, the heart rate, the pedaling cadence and the outdoor temperature. Different models will be established according to the activity duration, from very short efforts, involving the cyclists' lactic capacities, to longer efforts, involving aerobic channels: 2 minutes, 10 minutes and 30 minutes. For each duration of effort, a variable selection is carried out in order to select which features are the most discriminant for the power zone. This would exhibit what are the levers of action on which the cyclist could act in order to optimize the mean power zone.

From a statistical point of view, the power zones are ordinal data, whereas the other features (speed, slope, altitude, heart rate, cadence and temperature) are functional data (quantitative measures evolving over a continuum). The next section presents the existing models for ordinal and functional data and concludes that none model exists in the literature for predicting an ordinal variable from functional ones. The Functional Ordinal Logistic

Regression (FOLR) model is thus proposed in Section 3, as well as its maximum likelihood inference. Section 4 presents a comparison of FOLR with alternative approaches on the basis of a real data set from the literature. Cycling data are then analyzed in Section 5. Some conclusions and perspectives are given in Section 6

2 Related work

Ordinal data is one particular type of categorical data, occurring when the categories are ordered. Such data are very frequent in practice, as for instance in marketing studies where people are asked through questionnaires to evaluate some products or services on an ordinal scale. Nevertheless, this is not seldom that practitioners either consider them as quantitative integer data, assimilating the indexes of categories to integers, or even as nominal data, neglecting the order among the categories. In supervised learning, when the task is to predict an ordinal response variable, historical models are based on the modeling of cumulative probabilities that the ordinal variable is lower than a given category Agresti (2010). More recent research on ordinal data are essentially in classification or clustering, without using covariate. In the classification context, Cardoso and Pinto da Costa (2007) convert the problem of ordinal prediction into a binary classification problem, whereas Chu and Keerthi (2007) adapt the support vector machine paradigm to the ordinal case. In the clustering context, Jacques and Biernacki (2018) propose a mixture model based on a new distribution for ordinal data, whereas McParland and Gormley (2013) propose a latent variable approach. But when the goal is to predict an ordinal variable using covariate, the reference models remain those modeling the cumulative probabilities with a *link*-linear model Agresti (2010).

In these latter, the ordinal response is predicted from the observations of scalar covariates. In the present work, we are interested in functional covariates, occurring when covariates are curves. Functional data Ramsay and Silverman (2005) become ubiquitous since the modern technologies ease the collection of high frequency data. The cycling sport devices discussed in the introduction are a good example.

In the literature, regression with functional covariates has been developed for many types of responses. The most usual is the regression model for continuous scalar response, which has been proposed either in a parametric Ramsay and Silverman (2005) or non-parametric way Ferraty and Vieu (2006). Several models have been proposed for categorical nominal response: Ratcliffe et al. (2002) proposes a binary logistic regression model for functional covariate, whereas Escabias et al. (2005) propose a model based on functional principal components. A Partial Least Square approach has also been considered in Preda et al. (2007). A model for a functional response is also available in Ramsay and Silverman (2005). However, to the best of our knowledge, none functional regression model has been developed for ordinal response. In Preda et al. (2007), the PLS model is proposed to predict the quality of cookies from observation of the resistance of dough during the kneading process. The quality, Good, Adjustable or Bad, is clearly expressed on an ordinal scale, but has been considered as a nominal one, more precisely as a binary one removing the Adjustable category.

The present work aims to provide a prediction model for an ordered categorical response variable on the basis of functional covariates. The next section presents the Functional Ordinal Logistic Regression (FOLR) model. Section 3.2 focuses on the specificity of functional data and their modeling, whereas Section 3.3 proposes an estimation algorithm for the FOLR model.

3 The Functional Ordinal Logistic Regression model

3.1 The model

Let Y be an ordinal categorical variable, with C categories, quoted by 1 to C. Let X_j be a functional random variable ($1 \leq j \leq p$) with values in $L_2[0,T]$, $T > 0$, and assume that X_j is a L_2-continuous stochastic process, $X_j = \{X_j(t),\ t \in [0,T]\}$. Let $\pi_c(x) = p(Y = c|X = x)$. Cumulative logit models aims to model

$$\text{logit } p(Y \leq c|X = x) = \log \frac{p(Y \leq c|X = x)}{p(Y > c|X = x)} = \frac{\pi_1 + \ldots + \pi_c}{\pi_{c+1} + \ldots + \pi_C}$$

for $c = 1, \ldots, C-1$, with a linear combination of predictors. For functional predictors, the Functional Ordinal Logistic Functional Regression (FOLR) model proposed in this paper can be written:

$$\text{logit } p(Y \leq c|X = x) = \alpha_c - \sum_j^p \int_0^T \beta_j(t) x_j(t) dt, \tag{1}$$

where $\beta_j(t)$, $t \in [0,T]$, are the functional regression coefficients, $\alpha_1 \leq \ldots \leq \alpha_{C-1}$ and $1 \leq c \leq C-1$. With this model, each cumulative logit (1) has its own intercept, whereas the effect of the covariates $X_j(t)$ is shared by all of them. The minus sign for the covariates effect is chosen in order that, for small values of $\sum_j^p \int_0^T \beta_j(t) x_j(t) dt$ the response is likely to fall in the first category and for large values the response is likely to fall in the last category. Figure 2 illustrated the corresponding probabilities (for one covariate, $p = 1$).

In the sequel, only one functional covariate is considered ($p = 1$) for simplicity, but extension is straightforward. Cycling data analyzed at the end of the paper are multivariate ($p = 6$).

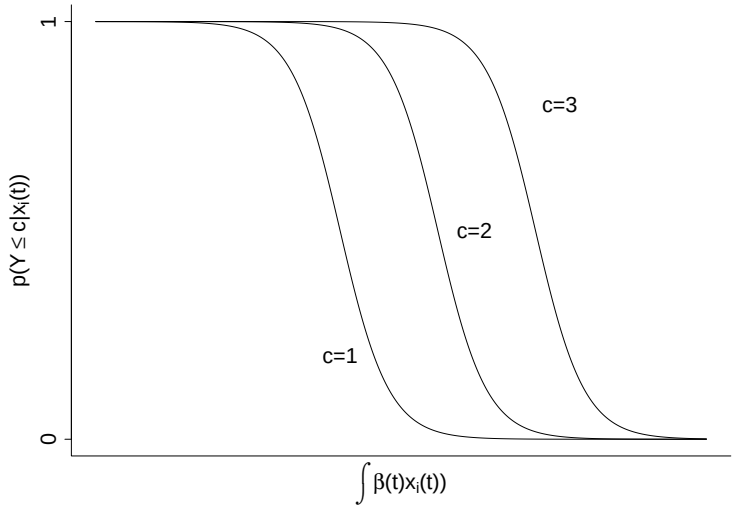

Figure 2: Illustration of the FOLR model probabilities

3.2 From discrete observation to functional data

Let consider a data set $(y_i, x_i(t))_{1\leq i\leq n}$ of joint observations of the ordinal response Y and the functional covariate X. In practice, the functional expression of the $x_i(t)$ are not known, and we only have access to their observation at some discrete time points $0 \leq t_1 \leq \ldots \leq t_S \leq T$. For simplicity of presentation, the same number of time points is considered for every $x_i(t)$, but the contrary case can in practice easily be considered. The first task, when working with functional data, is therefore to convert these discretely observed values to a function $x_i(t)$, computable for any desired argument value $t \in [0, T]$. One way to do that is interpolation, which is used if the observed values are assumed to be errorless. However, if there is some noise that needs to be removed, a common way to reconstruct the functional form is to assume that the curves $x_i(t)$ can be decomposed into a finite dimensional space, spanned

by a basis of functions Ramsay and Silverman (2005):

$$x_i(t) = \sum_{r=1}^{R} a_{ir}\phi_r(t) = \boldsymbol{a}_i'\boldsymbol{\phi}(t) \tag{2}$$

where $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_R(t))'$ is the basis of functions, R the number of basis functions, and $\boldsymbol{a}_i = (a_{i1}, \ldots, a_{iR})'$ the basis expansion coefficients.

The choice of the basis functions $\boldsymbol{\phi}(t)$, has to be made by the user. There is no straight rules about how to choose the appropriate ones Jacques and Preda (2014). We can nevertheless recommend the use of a Fourier basis in the case of data with a repetitive pattern, and B-spline functions in most other cases.

The estimation of the coefficients $\boldsymbol{a}_i$ is usually done through least square smoothing (see Ramsay and Silverman (2005)), as a preliminary step of the estimation of Model (1). If $\boldsymbol{x}_i = (x_i(t_1), \ldots, x_i(t_s))'$ is the vector of discrete observations of $x_i(t)$, and $\boldsymbol{\Phi}$ the $S \times p$ matrix containing the $\phi_j(t_s)$, the least square estimation of $\boldsymbol{a}_i$ are:

$$\hat{\boldsymbol{a}}_i = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{x}_i.$$

Similarly, the functional regression coefficient $\beta(t)$ are also assumed to be decomposed into a finite basis of functions. For simplicity, it is assumed to be the same basis as for X:

$$\beta(t) = \sum_{r=1}^{R} b_r\phi_r(t) = \boldsymbol{b}'\boldsymbol{\phi}(t) \tag{3}$$

with $\boldsymbol{b} = (b_1, \ldots, b_R)'$.

Under these basis expansion assumptions, the FOLR model is:

$$\begin{aligned}
\text{logit } p(y_i \leq c | X = x_i) &= \alpha_c - \int_0^T \sum_{r=1}^{R} b_r \phi_r(t) \sum_{r'=1}^{R} a_{ir'} \phi_{r'}(t) dt \\
&= \alpha_c - \sum_{r=1}^{R} \sum_{r'=1}^{R} b_r a_{ir'} \int_0^T \phi_r(t) \phi_{r'}(t) dt \\
&= \alpha_c - \boldsymbol{b}' \Psi \boldsymbol{a}_i \\
&= [1 \quad -\Psi \boldsymbol{a}_i] * \begin{bmatrix} \alpha_c \\ \boldsymbol{b}' \end{bmatrix}
\end{aligned}$$

where Ψ is the $R \times R$ matrix of inner products between basis functions $\int_0^T \phi_r(t)\phi_{r'}(t)dt$.

3.3 Model inference

For a data set $(y_i, \boldsymbol{a}_i)_{1 \leq i \leq n}$ of joint observation of the response and the basis expansion coefficients, there is a need to estimate FOLR model parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{b})$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{C-1})$. This is done by maximizing the following log-likelihood:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) = & \sum_{i=1}^{n} \left(1_{y_i=1} \log g(\alpha_1 - \boldsymbol{b}' \Psi \boldsymbol{a}_i) + 1_{y_i=C} \log[1 - g(\alpha_{C-1} - \boldsymbol{b}' \Psi \boldsymbol{a}_i)] \right) \\
+ & \sum_{i=1}^{n} \sum_{c=2}^{C-1} 1_{y_i=c} \log[g(\alpha_c - \boldsymbol{b}' \Psi \boldsymbol{a}_i) - g(\alpha_{c-1} - \boldsymbol{b}' \Psi \boldsymbol{a}_i)] \qquad (4)
\end{aligned}$$

where $g(t) = 1/(1 + \exp(-t))$ is the standard logistic cumulative density function.

In order to compute the maximum likelihood estimator, the derivative according to $\boldsymbol{b}$ and $\boldsymbol{\alpha}$ are computed. By denoting $h(t) = \exp(-t)/(1 + \exp(-t))^2$ the derivative of $g(t)$, we have:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{b}} = \sum_{i=1}^{n} \left(1_{y_{i=1}} \Psi \boldsymbol{a}_i \frac{h(\alpha_1 - \boldsymbol{b}'\Psi\boldsymbol{a}_i)}{g(\alpha_1 - \boldsymbol{b}'\Psi\boldsymbol{a}_i)} + 1_{y_{i=C}} \Psi \boldsymbol{a}_i \frac{-h(\alpha_{C-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)}{1 - g(\alpha_{C-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)} \right) + \sum_{i=1}^{n} \sum_{c=2}^{C-1} 1_{y_{i=c}} \Psi \boldsymbol{a}_i \frac{h(\alpha_c - \boldsymbol{b}'\Psi\boldsymbol{a}_i) - h(\alpha_{c-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)}{g(\alpha_c - \boldsymbol{b}'\Psi\boldsymbol{a}_i) - g(\alpha_{c-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)} \tag{5}$$

and

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \alpha_c} = \sum_{i=1}^{n} \left(1_{y_{i=1}} \frac{\delta_{c,1} h(\alpha_1 - \boldsymbol{b}'\Psi\boldsymbol{a}_i)}{g(\alpha_1 - \boldsymbol{b}'\Psi\boldsymbol{a}_i)} + 1_{y_{i=C}} \frac{-\delta_{c,C} h(\alpha_{C-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)}{1 - g(\alpha_{C-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)} \right) + \sum_{i=1}^{n} \sum_{k=2}^{C-1} 1_{y_{i=c}} \frac{\delta_{c,k} h(\alpha_k - \boldsymbol{b}'\Psi\boldsymbol{a}_i) - \delta_{c,k} h(\alpha_{k-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)}{g(\alpha_k - \boldsymbol{b}'\Psi\boldsymbol{a}_i) - g(\alpha_{k-1} - \boldsymbol{b}'\Psi\boldsymbol{a}_i)} \tag{6}$$

where $\delta_{c,k}$ is the Kronecker delta, equal to 1 if category c is the same as category k, 0 otherwise.

Since the maximum likelihood equations deriving from these derivatives have no closed form solutions, an iterative optimization algorithm has to be applied. Here, we have opted for the Fisher scoring algorithm Osborne (1992). Let $\mathcal{V}(\boldsymbol{\theta})$ be the gradient of $\ell(\boldsymbol{\theta})$, composed of terms given in equations (5) and (6), and the $\mathcal{I}(\boldsymbol{\theta})$ be the Fisher Information matrix. Starting from a initialization $\boldsymbol{\theta}^{(0)}$ of $\boldsymbol{\theta}$, the Fisher scoring algorithm update the parameter by:

$$\boldsymbol{\theta}^{(q+1)} = \boldsymbol{\theta}^{(q)} + \mathcal{I}(\boldsymbol{\theta}^{(q)})^{-1} \mathcal{V}(\boldsymbol{\theta}^{(q)})$$

until convergence of the parameter values, i.e. when $|\boldsymbol{\theta}^{(q+1)} - \boldsymbol{\theta}^{(q)}| < \epsilon$.

4 Comparison with competitors

In this section we show that the proposed FOLR model is competitive compared to the closest competitors, which are the multinomial functional logistic regression and the follow-

ing non-functional methods: random forest, ordinal logistic regression and support vector machine. Rather than choosing a subjective simulated data set, we choose to base the comparison on a real data set from the literature, with an ordinal response to forecast from functional features.

4.1 The Kneading data set

The Kneading data set is a well-known benchmark in functional data analysis, described in details in Lévéder et al. (2004). It concerns the quality of cookies and the relationship with the flour kneading process. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. One obtains 115 kneading curves observed at 241 equispaced instants of time in the interval $[0, 480]$. The 115 flours produce cookies of different quality: 50 of them have produced cookies of *good* quality, 25 produced *medium* quality and 40 *low* quality. This data, have been already studied in a supervised classification context Lévéder et al. (2004); Preda et al. (2007). They are known to be hard to discriminate, even for supervised classifiers, partly because of the medium quality class. Taking into account that the resistance of dough is a smooth curve measured with error, and following previous works on this data Lévéder et al. (2004); Preda et al. (2007), least squares approximation on a basis of cubic B-spline functions is used to reconstruct the true functional form of each sample curve. If Lévéder et al. (2004); Preda et al. (2007) used cubic B-spline with 18 knots (22 basis functions), we propose to select it according to our prediction purpose using cross validation. Raw and smoothed data are plotted on Figure 3.

Figure 3: Raw and smoothed kneading data

4.2 Models in competition

All the following models are compared on the basis of the correct classification rate (accuracy ratio) evaluated by 5-folds cross validation.

Functional Ordinal Logistic Regression (FOLR) As previously mentioned, we have to select how many cubic B-spline basis functions we have to use. The main idea was to test multiple values to determine the optimal number of basis functions. Different number of B-spline basis functions have been considered in a range of values from 5 (single interior knot) to 15 (11 interior knots).

The global evaluation scheme, used to compare the different models, is 5-folds cross validation. Here, we have implemented nested cross-validation, which consists of two cross-validations. First of all, we have divided our data set on 5 folds, in a manner that each of them contains equal amount of data (20%). Next, at each step, we have retained one fold for testing which would leave 80% of data for training. In order to test different number of basis functions, 10-folds cross validation is implemented on this training data set. Using

this nested cross-validation scheme lead to select 6 cubic B-spline basis functions.

Functional Multinomial Logistic Regression (FMLR) FMLR is the multinomial (non ordinal) version of functional logistic regression, introduced in its binary version in Ratcliffe et al. (2002). We extended this method to more than two categories and used our own implementation in R. This implementation is based on the `multinom` function of the `nnet` package. The same nested cross-validation as for FOLR indicated that 9 cubic B-spline basis functions should be selected.

LASSO-Ordinal Logistic Regression (OLR) We also used as a competitor the non functional version of ordinal logistic regression, applied directly on the raw data (241 features corresponding to the 241 time points). Due to the fact that the number of predictors was larger than the number of observations, we performed the LASSO penalized version of the ordinal logistic regression model, with the usage of the `ordinalNet` R package. The choice of the penalty parameter λ is done with the same nested cross-validation scheme as before.

Random Forest (RF) The next competitor is Random Forest, applied through the `caret` R package. The *mtry* parameter which determines the optimal number of variables that will be used at each random split of the decision tree, has also been selected by nested cross-validation.

Support Vector Machine (SVM) Last but not least, support vector machine was also considered. The cost parameter C is selected by nested cross-validation. The cost parameter determines the width of the margin of classification. For small values of the cost

parameter C, observations inside the margin are not penalized and we obtain better fit but larger estimation error. The otherwise is true for larger values of cost parameter, where the estimation error is minimized but the model may overfit the data.

4.3 Results

Table 1 summarize the obtained results from the nested cross-validation technique for all considered models. Standard deviations across folds are in parenthesis. The highest accuracy ratio is achieved with FOLR model, which can be expected because FMLR is not able to take into account the ordinal nature of the response, whereas RF, OLR and SVM do not consider the functional nature of the covariate.

Functional models	AR	Non-functional models	AR
FOLR	**0.829 (0.070)**	RF	0.776 (0.097)
FMLR	0.770 (0.087)	OLR	0.760 (0.123)
		SVM	0.80 (0.090)

Table 1: Cross-validated Accuracy Ratio (AR) for Kneadning data set (with standard deviation across folds).

5 Determination of the factors influencing the average power during cycling session

5.1 The data set

The data set is composed of 216 one hour bike sessions, during which are measured every second: the speed, the slope, the altitude, the heart rate of the cyclist, the pedaling cadence and the outdoor temperature. These data are measured with a Garmin Edge 520. The power is also recorded every second with a powermeter ROTOR INpower ROAD. These cycling sessions were carried out during the same year (2019), by the same amateur cyclist, and combine training and competition sessions. These sessions were cut into 3 different lengths: 2 minutes, 10 minutes and 30 minutes. Consequently, we have 6480 sessions of 2 minutes, 1296 sessions of 10 minutes and 432 sessions of 30 minutes. If these data came from a panel of different cyclists, the independence assumption requested for performing maximum likelihood estimation (Section 3.3) of the FOLR model would not hold. But in the present data set, all the data have been provided by the same individual, who have been cycling for a long time and whose level of performance is assumed to be broadly constant throughout the year. Consequently, each session of this cyclist can be reasonably assumed to be independent from each other.

The power is averaged during the session and ranked into the 7 power zone. Table 2 describes the distribution of the sample over the power zone in function of the session duration.

Figure 4 plots a sample of data, corresponding to 30 minutes session in power zone 1 (lowest one), 3, 5 and 7 (highest one).

Functional data reconstruction is performed with cubic spline basis with 20 basis func-

power zone	1	2	3	4	5	6	7
2’	1618	1431	927	631	493	785	595
10’	163	404	246	159	115	175	34
30’	27	148	117	46	31	59	4

Table 2: Distribution over the power zones in function of the session duration

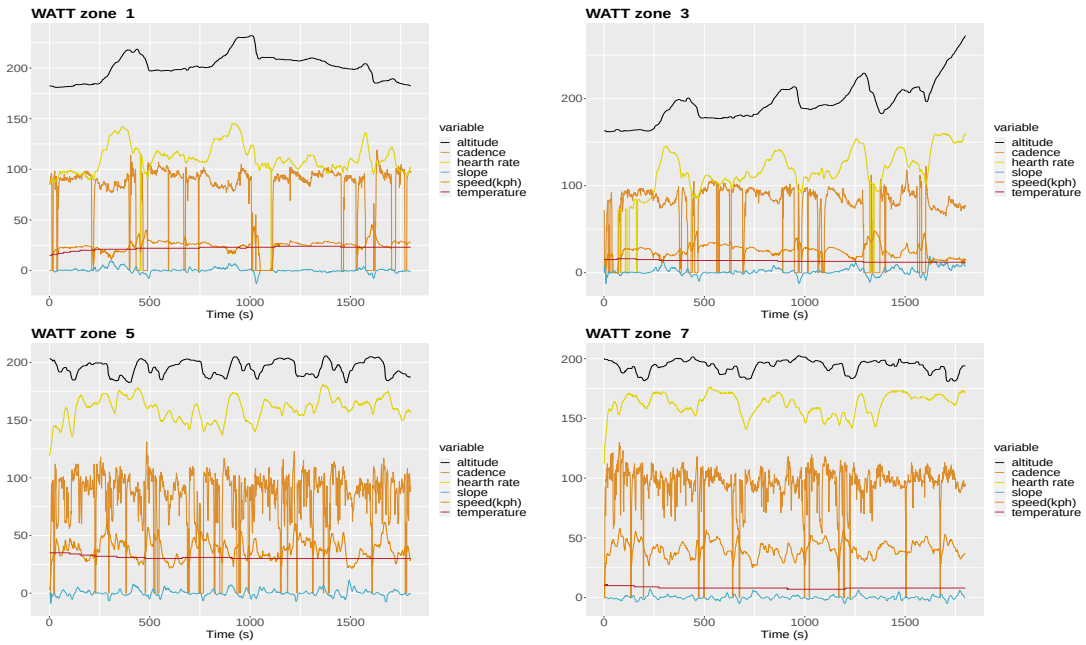

Figure 4: Cycling data for 30 minutes bike session in power zone 1 (top left), 3, 5 and 7 (bottom right)

tions. This choice was made empirically, so that the main variations of the curves are taken into account. For simplicity, the same number of basis functions is used for every session duration, although we could have used a smaller number for the shorter sessions.

5.2 Experimental setting

The goal is to model the mean power zone of the bike session according to the six functional covariates: the speed, the slope, the altitude, the heart rate of the cyclist, the pedaling cadence and the outdoor temperature. Since these covariates are not necessary relevant for modeling the power zone, a variable selection is performed. For this, all the possible subsets of covariates are considered and evaluated by 10-fold cross-validation. Even if this strategy is often avoided because of the exponential combinatorics of the number of subgroups of variables, it is quite feasible here because 6 variables imply 63 subgroups. Due to the ordinal nature of the power zone, the models are evaluated by the Root Mean Square Error (RMSE), as suggested in Gaudette and Japkowicz (2009). This criterion is preferred to the classification accuracy which does not take into account the proximity between two ordinal categories.

Model estimation is implemented in the R package `FRM`. This package provide several models for functional predictors (linear regression, logistic regression FMLR and the FOLR model) as well as the cycling data set. The `FRM` package is available from the authors upon request, and will be submitted to CRAN after publication of the present paper. With this package, one FRM model estimation is about 10 seconds on a 3,5 GHz Intel Core i7 processor with 16 Go of memory.

5.3 Results

Variables selection In the proposed approach, the discriminant variables are selected through a model selection approach. There is a large literature on model selection, and a recurring question is whether we should choose the best model or a set of good models.

In order to illustrate this question, Figure 5 plots the cross-validated RMSE value for the 63 models (ordered by increasing RMSE), for the 3 bike session duration (2, 10 and 30 minutes). On this Figure, the lowest is the RMSE, the better is the model. On the one hand, we can see that there is not one model which is clearly better than the others. On the other hand, we notice that a group of model stands out slightly from the others. We arbitrarily select this set of models, denoted in the sequel as *good models*, by stopping at the first break in the RMSE values (red vertical lines). This leads to the selection of 24 models for 2 and 10 minutes bike session, and 26 models for 30 minutes length.

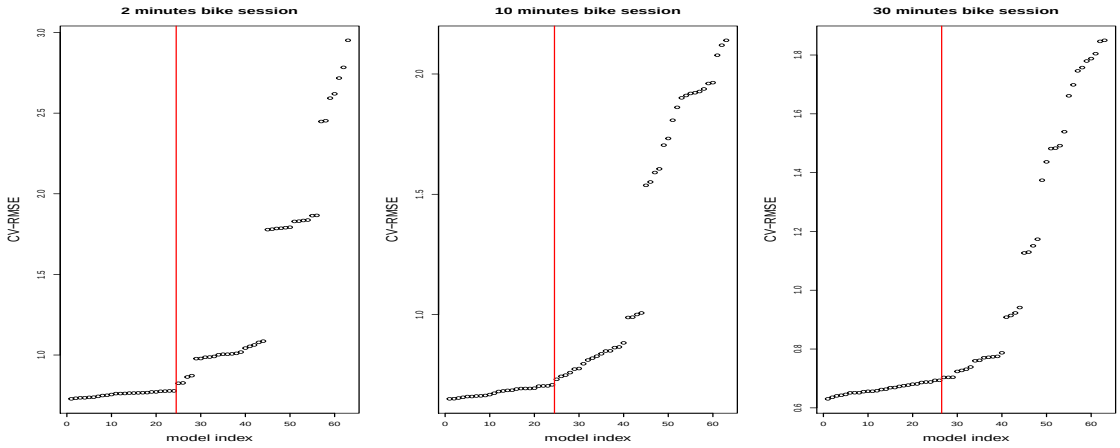

Figure 5: Cross-validated RMSE values for the 63 models and for the 3 bike session duration

In order to estimate the importance of each variable in this set of good models, we give a score to each variable occurring in each model. This score depends on the ranking of the model. For instance, for 2 minutes bike sessions, variables occurring in the best model obtain a score of 24 (the number of good models), those in the second best model a score of 23, and so on. Then, the score of the variables are summed over all the good models.

Figure 6 plots the resulting variables importance, according to the bike session duration. On this figure, the more outward the indicator, the more important the variable.

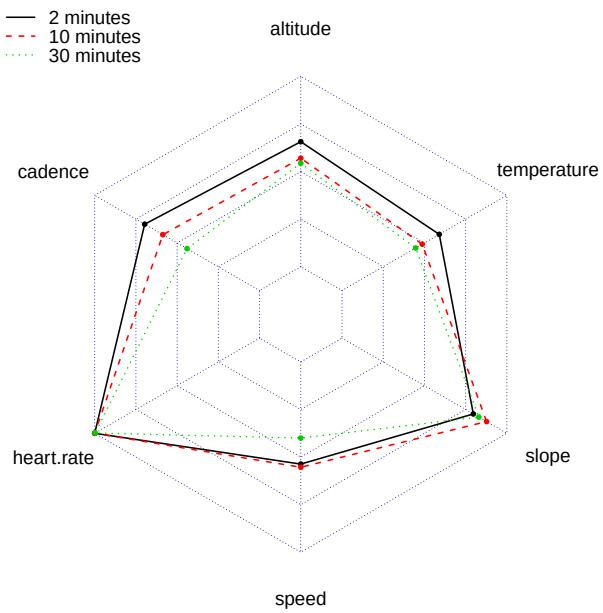

Figure 6: Variable importance according to the bike session duration

Prediction accuracy Table 3 presents the cross-validated prediction results on whole data set. These results are obtained by the best model according to CV-RMSE, for each bike session duration. For 2 minutes sessions, the best model uses all variables except the altitude, and has a cross-validated RMSE of 0.728. For 10 minutes sessions, the best model uses the heart rate, the cadence, the slope and the speed, and has a cross-validated RMSE of 0.650. For 30 minutes sessions, the best model uses the heart rate, the cadence and the

slope, and has a cross-validated RMSE of 0.630.

5.4 Analysis

First of all, we can note that the quality of the prediction is relatively correct, with a majority of elements on the diagonal of the confusion matrices (bold numbers in Table 3). Even when a power zone is wrongly predicted, the error is small since this is generally a contiguous power zone which is being predicted (numbers on the subdiagonal or the superdiagonal). This means that the studied variables reflect the average developed power, i.e. the cyclist's performance. Then, we can notice that the quality of the prediction increases with the duration of the exercise, as RMSE decreases.. It seems easier to predict the average power of a longer effort, which is subject to less irregularities, than the shorter efforts.

Regarding the importance of the variables, Figure 6 shows that all variables are important. The most important variable, regardless of the duration of the exercise, is the cyclist's heart rate: it is selected in all the good models. This is not surprising, because developing significant power requires significant physical effort. And this confirms that the heart rate, when power was not easily measurable during an exercise, was a good indicator of performance, or at least of the developed power. Then comes the slope, which is slightly more important when the duration of the exercise is short. This is certainly due to the fact that amateur cyclists naturally develop high powers to climb steep slopes with short efforts, but it becomes more complicated when the effort is prolonged. The third variable in order of importance is the pedaling cadence. This variable is of particular interest because it is the one on which the cyclist can act. Indeed, the slope that we have just discussed, is an external parameter, like the temperature and the altitude, on which the cyclist cannot

act. Let us note that the temperature and the altitude are also important, and that their importance decrease when the effort is prolonged. Let's come back to the cadence: it is very interesting to notice that this is the variable for which the difference in importance according to the length of the effort is the most obvious. Moreover, the importance of the pedaling cadence decreases significantly with the length of the effort. The interpretation is as follows: if the studied amateur cyclist is able to hold high cadences to maintain high power during short efforts, it becomes much more complicated during longer efforts. Maintaining high power using high cadence over a long period of time is the prerogative of high level cyclists, such as the professional cyclists mentioned in the introduction. Finally, the least important variable is the speed. Even if this might seem surprising at first glance, it is because it is totally linked to the nature of the terrain: we will develop significant powers to climb steep climbs, without going very fast; and on the contrary, we will go very quickly downhill at very low or even null power when the cyclist is not pedaling.

6 Conclusion

This work proposes a study of the data that are commonly produced by cyclists during the practice of their sport. In particular, this study is interested in the factors allowing to discriminate the average power developed during the effort. One of the most interesting conclusions from a sports practice's point of view is that the pedaling cadence is indeed a lever for optimizing the developed power. Nevertheless, this study shows that for the amateur cyclist who provided these data, maintaining a high cadence over a long time is difficult. This maintenance of a high cadence over a long time is one of the typical characteristics of high level cyclists, and is a factor on which the amateur cyclist can work

on to increase his performance. From a statistical modeling point of view, this study has needed the development of an ordinal logistic regression model with functional predictors (FOLR). Experimental study on a benchmark has shown the efficiency of this model in comparison to the competitors which either omit the ordinal nature of the response or the functional nature of the covariate.

The perspectives from a cycling data point of view would be to complete the data set with data from other cyclists, of various profiles and levels, in order to build a more heterogeneous database and thus to be able to draw more general conclusions. But that will require adapting the FOLR model to take into account an individual effect, or at least to incorporate the effect of the age, the weight, etc. This requires to develop a new cumulative logit model with functional and non functional covariate. Similarly, it would also be interesting to be able to take into account the period of the year at which the data is measured. Cyclists have a level that evolves throughout the year and it would be nice to be able to take that into account. But again, that would require the development of a new model.

SUPPLEMENTARY MATERIAL

R-package for FOLR routine: R-package FRM containing code to perform FOLR inference and prediction described in the article. The package also contains the cycling data set, which are currently under copyright and could be shared after publication of the present paper. (GNU zipped tar file)

Cycling analysis: R-code Cycling-Analysis containing code to perform analyses described in Section 5. (Cycling-Analysis.Rmd)

resultXminute: Cross-validation results of the cycling data analysis used for model selection. (resultXminute.Rdata)

References

Abbiss, C., J. Peiffer, and P. Laursen (2009, 03). Optimal cadence selection during cycling. *ECU Publications 10*.

Agresti, A. (2010). *Analysis of ordinal categorical data*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York.

Beattie, K., B. Carson, M. Lyons, and I. Kenny (2016, 09). The effect of maximal and explosive strength training on performance indicators in cyclists. *International Journal of Sports Physiology and Performance 12*, 1–25.

Bini, R., F. Diefenthaeler, and F. Carpes (2014). Determining force and power in cycling: A review of methods and instruments for pedal force and crank torque measurements. *International Sportmed Journal 15*(1), 96–112.

Borszcz, F., A. Tramontin, A. Bossi, L. Carminatti, and V. Costa (2018, 05). Functional threshold power in cyclists: Validity of the concept and physiological responses. *International Journal of Sports Medicine 39*.

Cardoso, J. and J. Pinto da Costa (2007). Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research 8*, 1393–1429.

Chu, W. and S. Keerthi (2007). Support vector ordinal regression. *Neural Computation 19*(3), 792–815.

Escabias, M., A. Aguilera, and M. Valderrama (2005). Modeling environmental data by functional principal component logistic regression. *Environmetrics 16*, 95–107.

Faria, E., D. Parker, and I. Faria (2005, 02). The science of cycling: Factors affecting performance ??? part 2. *Sports medicine (Auckland, N.Z.) 35*, 313–37.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis*. Springer Series in Statistics. New York: Springer.

Gaudette, L. and N. Japkowicz (2009). Evaluation methods for ordinal classification. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, Canadian AI '09, Berlin, Heidelberg, pp. 207–210. Springer-Verlag.

Grappe, F. (2012). *Puissance et performance en cyclisme*. De Boeck.

Grappe, F. (2018). *Cyclisme et opitmisation de la performance*. De Boeck.

Jacques, J. and C. Biernacki (2018). Model-based co-clustering for ordinal data. *Computational Statistics and Data Analysis 123*, 101–115.

Jacques, J. and C. Preda (2014). Model based clustering for multivariate functional data. *Computational Statistics and Data Analysis 71*, 92–106.

Lévéder, C., P. Abraham, E. Cornillon, E. Matzner-Lober, and N. Molinari (2004). Discrimination de courbes de pï¿œtrissage. In *Chimiomï¿œtrie 2004*, Paris, pp. 37–43.

McParland, D. and C. Gormley (2013). *Algorithms from and for Nature and Life: Studies in Classification, Data Analysis, and Knowledge Organization*, Chapter Clustering Ordinal Data via Latent Variable Models, pp. 127–135. Switzerland: Springer.

Nimmerichter, A., R. Eston, N. Bachl, and C. Williams (2011, 05). Longitudinal monitoring of power output and heart rate profiles in elite cyclists. *Journal of sports sciences 29*, 831–40.

Osborne, M. R. (1992). Fisher's method of scoring. *International Statistical Review 60*(1), 99–117.

Preda, C., G. Saporta, and C. Lévéder (2007). PLS classification of functional data. *Computational Statistics 22*(2), 223–235.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second ed.). Springer Series in Statistics. New York: Springer.

Ratcliffe, S. J., G. Z. Heller, and L. R. Leader (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in medicine 21*(8), 1115–1127.

van Dijk, H., R. van Megen, and G. Vroemen (2017). *The Secret of Cycling: Maximum Performance Gains Through Effective Power Metering and Training Analysis*. Meyer & Meyer Sport.

pred \ true	1	2	3	4	5	6	7
1	**1260**	250	19	1	1	0	0
2	324	**939**	329	40	1	0	0
3	33	207	**412**	229	38	3	0
4	1	32	129	**224**	149	32	0
5	0	3	34	94	**142**	99	4
6	0	0	4	43	161	**530**	195
7	0	0	0	0	1	121	**396**

pred \ true	1	2	3	4	5	6	7
1	**113**	25	1	0	0	0	0
2	50	**304**	92	6	0	0	0
3	0	68	**120**	50	1	0	0
4	0	7	31	**80**	22	3	0
5	0	0	2	21	**55**	31	0
6	0	0	0	2	37	**135**	21
7	0	0	0	0	0	6	**13**

pred \ true	1	2	3	4	5	6	7
1	**19**	7	0	0	0	0	0
2	8	**108**	34	3	0	0	0
3	0	27	**76**	17	0	0	0
4	0	6	7	**22**	4	1	0
5	0	0	0	3	**17**	8	0
6	0	0	0	1	10	**49**	4
7	0	0	0	0	0	1	**0**

Table 3: True power zones versus cross-validated predicted ones by the best model (according to CV-RMSE), for 2 minutes (top), 10 minutes (middle) and 30 minutes (bottom) bike session. In bold are the true predictions.