Journal of Machine Learning Research 8 (2007) 2727-2754 Submitted 1/07; Revised 7/07; Published 12/07

A New Probabilistic Approach in Rank Regression with Optimal Bayesian Partitioning

Carine Hue CARINE.HUE@GMAIL.COM
Marc Boullé MARC.BOULLE@ORANGE-FTGROUP.COM
France Telecom R&D
2, avenue Pierre Marzin
22307 Lannion cedex, France

Editors: Isabelle Guyon and Amir Saffari

Abstract

In this paper, we consider the supervised learning task which consists in predicting the normalized rank of a numerical variable. We introduce a novel probabilistic approach to estimate the posterior distribution of the target rank conditionally to the predictors. We turn this learning task into a model selection problem. For that, we define a 2D partitioning family obtained by discretizing numerical variables and grouping categorical ones and we derive an analytical criterion to select the partition with the highest posterior probability. We show how these partitions can be used to build univariate predictors and multivariate ones under a naive Bayes assumption.

We also propose a new evaluation criterion for probabilistic rank estimators. Based on the logarithmic score, we show that such criterion presents the advantage to be minored, which is not the case of the logarithmic score computed for probabilistic value estimator.

A first set of experimentations on synthetic data shows the good properties of the proposed criterion and of our partitioning approach. A second set of experimentations on real data shows competitive performance of the univariate and selective naive Bayes rank estimators projected on the value range compared to methods submitted to a recent challenge on probabilistic metric regression tasks.

Our approach is applicable for all regression problems with categorical or numerical predictors. It is particularly interesting for those with a high number of predictors as it automatically detects the variables which contain predictive information. It builds pertinent predictors of the normalized rank of the numerical target from one or several predictors. As the criteria selection is regularized by the presence of a prior and a posterior term, it does not suffer from overfitting.

Keywords: rank regression, probabilistic approach, 2D partitioning, non parametric estimation, Bayesian model selection

1. Introduction

In this introduction, we precise the supervised learning task we address in this paper, that is the rank regression. We then show the interest of probabilistic learning approaches compared to deterministic ones. Finally, we outline our contribution which aims at selecting a probabilistic predictive model for the rank of a numerical target with a nonparametric Bayesian approach.

1.1 Value, Ordinal and Rank Regression

In supervised learning, classification tasks, where the target variable is categorical, are usually distinguished from regression tasks, where it is numerical. A less known task is the case where the target variable is ordinal, usually called ordinal regression (see Chu and Ghahramani, 2005, for a state of the art). In this case, there is a total order between the target values but no distance information. The practical problems studied in the machine learning community consider a low number of distinct integer ranks, roughly 5 or 10, fixed before the learning. The aim is then to predict the right quintile or decile an example belongs to. The algorithms are generally evaluated with the mean zero-one error or with the mean absolute error obtained by considering the ordinal scales as consecutive integers. Among the proposed methods, the principle of empirical risk minimization with a loss function measuring the probability of misclassification is applied by Herbrich et al. (2000), an online algorithm based on the perceptron algorithm is proposed in the work of Crammer and Singer (2001), support vector machines are used by Shashua and Levin (2002) and Chu and Keerthi (2005) and Gaussian processes by Chu and Ghahramani (2005).

The choice of a low number of distinct ranks is generally motivated by simplicity reasons. However, this predefined target discretization can separate values which form a pertinent prediction interval.

In this paper, we address rank regression tasks. More precisely, for a given numerical target variable, we aim at computing an estimator of its rank. During the estimation procedure we never take into account the distance between instances but only their order. Several reasons have guided that choice. First, considering ranks rather than values is a classical way to obtain models more robust to outliers and to heteroscedasticity. In linear regression for example, an estimator based on the centered ranks in the minimization of the least squared equation is proposed in the approach of Hettmansperger and McKean (1998). Secondly in some applications, predicting the rank of a target variable is more interesting than predicting its intrinsic value. For instance in information retrieval, some search engines use numerical scores to rank web pages but the score value has no other usefulness.

1.2 Deterministic and Probabilistic Regression

Whatever the learning task, the simpler approach is determinist in so far as its outputs is deterministic: the majority class in classification, the mean rank in ordinal regression and the conditional mean in metric regression. These punctual predictors turn out to be inefficient as soon as confidence intervals or prediction of extreme values are needed. In this context, quantile regression or density estimation aims at estimating the predictive density more accurately. Such a probabilistic approach is very useful as soon as the predictive model is used for decision-making. For instance, modelling predictive uncertainty is still an active research domain and has been the subject of two recent challenges: the evaluating predictive uncertainty challenge supported by the PASCAL network of excellence in 2004-2005 and the predictive uncertainty in environmental modelling competition (organized by Cawley et al., 2006).

Quantile regression consists in estimating some quantiles of the predictive law. For a real α in $[0,1]$, the conditional quantile $q_\alpha(x)$ is defined as the lowest real value such that the conditional cumulative distribution is higher than α. Quantile regression can be formulated as a minimization problem. Starting from that, different methods have been proposed according to the form assumed for the quantile function: the minimization problem is solved with splines in the approach

of Koenker (2005), with kernel functions in the work of Takeuchi et al. (2006) and neural networks in the work of White (1991). The approach proposed by Chaudhuri et al. (1994); Chaudhuri and Loh (2002) mixes a tree partitioning of the predictors space and a local polynomial assumption. Random forests have been extended to conditional quantile estimation in the work of Meinshausen (2006). For all these approaches, the reals α are known in advance and usually in a small number, and the estimation of each quantile is done and evaluated independently from the others.

Conditional density estimation aims at giving for any couple (x,y) an estimator of the predictive density $p(y|x)$. The parametric approaches assume that the predictive density belongs to a fixed parametric family and then reduce the density estimation problem to the estimation of a parameter vector. The non parametric approaches do not assume any fixed parametric form for the predictive density and are generally based on two ingredients: first, the estimator is computed on each point by using data contained in a point neighbourhood; then, an assumption is done on the local form of the estimator. Very popular, kernel methods weight the contribution of the data by convoluting the empirical law with a kernel density. The form and the width of the kernel remain tuning parameters. Once a neighbourhood is defined, methods differ on the estimator form: the local polynomial approach includes constant, linear and polynomial estimator (see Fan et al., 1996). Splines can also be used. Such a probabilistic approach has already been proposed in ordinal regression in the parametric context of the Gaussian processes with a Bayesian approach (see Chu and Keerthi, 2005).

1.3 Our Contribution

In this paper, we consider regression tasks, where the unknown target variable is numerical. Instead of looking for an estimator of the *value* of the target variable, we aim at building an estimator of the normalized *rank* (between 0 and 1) of this variable. Our second objective is to propose an evaluation criterion for these rank probabilistic estimators.

First, we propose a non parametric Bayesian method to build a probabilistic estimator specified by a set of quantiles of the rank cumulative distribution function. Unlike quantile regression approach, the choice of the quantiles is not made before the learning but is determined during this step. Moreover, in the case of several predictor variables, the multivariate estimator is obtained as a combination of univariate estimators under a naive Bayes assumption. Each univariate estimator is obtained from a 2D partition of the space (predictor, target) assuming that the rank density is constant on each cell of the partition. The optimal 2D partition is searched according to a model selection approach.

Secondly, we propose and evaluate a new criterion to evaluate probabilistic regression methods by comparing the rank predictive density and the true insertion rank of all test instances.

This paper is organized as follows: in the second section, we first describe the 2D-partitioning for numerical predictors. Compared to our previous work introduced in a conference paper (see Boullé and Hue, 2006), the approach is much more detailed, the selection criteria is completely explicited and we propose an estimator of the rank predictive density. We also propose the 2D-partitioning for categorical predictors that has never been published before.

In the third section we first expose how we can build a univariate estimator of the rank predictive density from each 2D partition. Using the naive Bayes assumption of conditional independence of the predictors, we then describe how to obtain multivariate predictors, with and without variable selection.

The fourth section focuses on the important topic of the evaluation of such rank regression models. We first give a brief overview of classical scores used in supervised learning. We then propose to use one of them, the logarithmic score, for rank probabilistic estimators without the shortcomings noted for probabilistic estimators based on values.

The last section is devoted to experimental evaluation. We begin with experiments on synthetic data to demonstrate on the one hand the relevance of the proposed evaluation criterion and on the other hand the performance of the 2D partitionings. We pursue with experimentations on real data sets to show the performance of the univariate and multivariate predictors. We end by a comparison of our approach with alternative methods proposed in a recent challenge dedicated to probabilistic metric regression.

2. A 2D-Partitioning Method for Probabilistic Regression

The 2D-partitioning method we present here comes from the extension of the so-called MODL approach to regression tasks. This approach has first been proposed for classification tasks in the work of Boullé (2005) and Boullé (2006).

For regression tasks, we present in the sequel two 2D partitioning methods depending on whether the considered predictor is numerical or categorical.

For numerical predictors, the 2D partition is the grid resulting from the discretization of both target and predictor. For categorical predictors, the 2D partition is the grid resulting from the discretization of the numerical target and from the grouping of the categorical predictor.

2.1 The 2D Discretization for Probabilistic Regression with Numerical Predictor

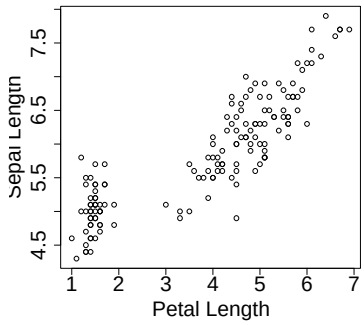


Figure 1: Scatter-plot of the iris data set of Fisher (1936) considered for a *regression* problem with the petal length variable as predictor and the sepal length variable as target.

In order to illustrate the regression problem with numerical predictor, we present in Figure 1 the scatter-plot of the iris data set considered for a *regression* problem with the petal length variable as predictor and the sepal length variable as target. The figure shows that iris plants with petal length below 2 cm always have a sepal length below 6 cm. We propose to exhibit the predictive information of the petal length variable by discretizing both the predictor and the target variables. For instance, the grid with six cells presented on the left of Figure 2 indicates that:

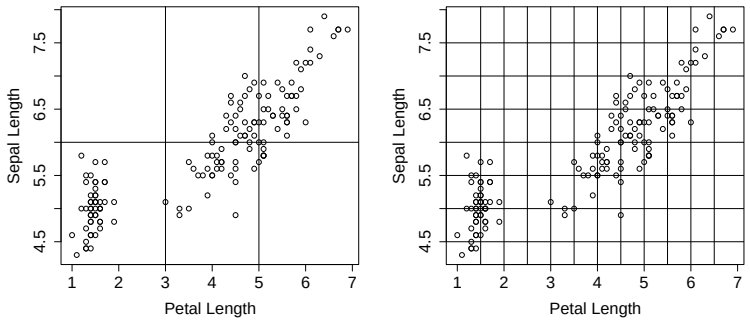


Figure 2: Two discretization grids with 6 or 96 cells, describing the correlation between the petal length and sepal length variables of the Iris data set.

- for a petal length lower than 3 cm, 100% of the training instances have a petal length higher than 6 cm;

- for a petal length between 3 and 5 cm, the two sepal length intervals are equiprobable (53% and 47%);

- for a petal length higher than 5 cm, 90% of the instances have a sepal length higher than 6 cm.

The 96 cells grid presented on the right of Figure 2 seems more accurate but may be less robust. These two examples illustrate that a compromise has to be found between the quality of the correlation information and the generalization ability, on the basis of the grain level of the discretization grid.

The issue is to describe the predictive distribution of the rank of the target value given the rank of the predictor value.

Let us now formalize this approach using a Bayesian model selection approach.

Definition 1 *A regression 2D discretization model is defined by:*

1. *a number of intervals for the target and predictor variables;*
2. *a partition of the predictor variable specified on the ranks of the predictor values;*
3. *for each predictor interval, the repartition of the instances among the target intervals specified by the instance counts locally to each predictor interval.*

Notations

N : the number of training instances

I : the number of predictor intervals

J : the number of target intervals

$N_{i.}$: the number of instances in the predictor interval i

$N_{.j}$: the number of instances in the target interval j

N_{ij} : the number of instances in the grid cell associated to predictor interval i and the target interval j.

A regression 2D discretization model is then entirely characterized by the parameters $\left\{I, J, \{N_{i.}\}_{1\leq i\leq I}, \left\{N_{ij}\right\}_{1\leq i\leq I, 1\leq j\leq J}\right\}$. The number of instances $N_{.j}$ can be deduced by adding the N_{ij} for each predictor interval, according to $N_{.j} = \sum_{i=1}^{I} N_{ij}$.

We now want to select the best model M given the available data, that is the most likely model given the data. Adopting a Bayesian approach, it comes to maximize :

$$p(M|D) = \frac{p(M)p(D|M)}{p(D)}.$$

The data distribution $p(D)$ being constant whatever the model M, it comes to maximize $p(M)p(D|M)$ which can be written:

$$p(M)p(D|M) = p(I,J)p(\{N_{i.}\}|I,J)p(\{N_{ij}\}|I,J,\{N_{i.}\})p(D|M).$$

We then add the restriction that the searched model is such that the conditional target distributions are independent. This assumption is first consistent with the objective to obtain a partition that discriminates distinct conditional distributions. Moreover, mathematically speaking, it enables to write the last factors in the precedent equation as products over the predictor intervals. It also reduces the complexity of the associated optimization algorithm.

Denoting by D_i the subset of D restricted to the interval i, one obtains:

$$p(M)p(D|M) = p(I,J)p(\{N_{i.}\}|I,J)\prod_{i=1}^{I} p(\{N_{ij}\}|I,J,\{N_{i.}\})\prod_{i=1}^{I} p(D_i|M).$$

To be able to evaluate a given model, we have to choose a prior distribution for the model parameters and a likelihood function. In Definition 2, we formalize our choices by using the assumption independence and proposing a uniform distribution at each stage of the prior parameter structure and of the likelihood function.

Definition 2 *The prior for the parameters of a regression 2D discretization model and the likelihood function of the data given a model are chosen hierarchically and uniformly at each level:*

1. *the numbers of intervals I and J are independent from each other, and uniformly distributed between* 1 *and N,*
2. *for a given number of predictor intervals I, every set of intervals is equiprobable,*
3. *for a given predictor interval, every distribution of the instances on the target intervals is equiprobable,*
4. *the distributions of the target intervals on each predictor interval are independent from each other,*
5. *for a given target interval, every distribution of the rank of the target values is equiprobable.*

Taking the negative log of the probabilities, this provides the evaluation criterion given in the following theorem:

Theorem 3 *A 2D-discretization model distributed according to a uniform hierarchical prior is Bayes optimal if its evaluation according to the following the criteria is minimal*

$$\begin{aligned} c(M) &= -\log(p(M)) - \log(p(D/M)) \\ &= 2\log N + \log\binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log\binom{N_{i.}+J-1}{J-1} \\ &+ \sum_{i=1}^{I} \log \frac{N_{i.}!}{N_{i1}!N_{i2}!\ldots N_{iJ}!} + \sum_{j=1}^{J} \log N_{.j}!. \end{aligned} \tag{1}$$

The first hypothesis introduced in Definition 2 gives that $p(I,J) = p(I)p(J) = \frac{1}{N}\frac{1}{N}$.

The second hypothesis is that all the divisions into I intervals are equiprobable for a given I. Computing the probability of one set of intervals turns into the combinatorial evaluation of the number of possible interval sets. Dividing the predictor values into I intervals is equivalent to decompose the number N as the sum of the $N_{i.}$ frequencies of the intervals. Using combinatorics, we can prove that this number of choices is equal to $\binom{N+I-1}{I-1}$. Using the equiprobability assumption, one finally obtains:

$$P(\{N_{i.}\}|I) = \frac{1}{\binom{N+I-1}{I-1}}.$$

The third hypothesis assumes that, for a given interval i of size $N_{i.}$, every distribution of the instances on the J target intervals are equiprobable. It remains to specify the parameters of a multinomial distribution of $N_{i.}$ instance over J values. Using combinatorics again, one obtains

$$P\big(\{N_{ij}\}|I,\{N_{i.}\}\big) = \frac{1}{\binom{N_{i.}+J-1}{J-1}}.$$

The prior terms being explicited, it remains to evaluate the likelihood on each predictor interval, that is the probability to observe the data restricted to each interval knowing the multinomial distribution model on each interval. The number of ways to observe $N_{i.}$ instances distributed according to such multinomial law is given by $\frac{N_{i.}!}{N_{i1}!N_{i2}!\ldots N_{iJ}!}$. To finish, according to the last hypothesis, for a given target interval, every distribution of the ranks of the target values are equiprobable which leads to the last terms.

By taking negative logarithms, one obtains the above Formula (1).□

To provide a first intuition, we can compute that for $I = J = 1$ the criterion value is $2\log(N) + \log(N!)$ (about 615 for $N = 150$) and for $I = J = N$ it gives $2\log(N) + \log\begin{pmatrix} 2N-1 \\ N-1 \end{pmatrix} + N\log(N)$ (about 966 for $N = 150$). This means that a discretization with one cell is always more likely that a 2D discretization with N^2 elementary cells.

We adopt a simple heuristic to optimize this criterion. We start with an initial random model and alternate the optimization on the predictor and target variables. For a given target distribution with fixed $J < \sqrt{(N)}$ and $N_{.j}$, we optimize the discretization of the predictor variable to determine the values of I, N_i and N_{ij}. Then, for this predictor discretization, we optimize the discretization of the target variable to determine new values of J, $N_{.j}$ and N_{ij}. The process is iterated until convergence, which usually takes between two and three steps in practice. The univariate discretization optimizations are performed using the MODL discretization algorithm. This process is repeated

several times, starting from different random initial solutions. The best solution is returned by the algorithm as described in Algorithm 1.

Each 1D-discretization is implemented according to a bottom-up greedy heuristic followed by a post-optimisation whose time complexity is in $N\log(N)$ times the size of the fixed partition. This algorithm complexity is mainly obtained by using the criteria additivity for 1D discretization (see Boullé, 2006). Imposing a maximum iteration number $P = 10$ for instance, the worst case complexity is bounded by $PN\sqrt(N)log(N)$ without decreasing the quality of the search algorithm. Despite

Algorithm 1 Optimization of a MODL 2D-discretization for regression tasks

Ensure: $M^*; c(M^*) \leq c(M)$ {Final solution with minimal cost}
1: **for** $m = 1, \ldots, 10$ **do**
2: {Initialize with a random partition}
3: $M \leftarrow$ a random partition of size $\sqrt(N)$
4: **while** improved **do**
5: {Univariate optimal 1D-discretization of predictor variable X:}
6: freeze the univariate partition of target variable Y
7: $M \leftarrow$ call *univariate optimal 1D-discretization* (M) for predictor variable X
8: {Univariate optimal 1D-discretization of target variable Y:}
9: freeze the univariate partition of predictor variable X
10: $M \leftarrow$ call *univariate optimal 1D-discretization* (M) for target variable Y
11: **end while**
12: **if** $c(M) \leq c(M^*)$ **then**
13: $M^* \leftarrow M$
14: **end if**
15: **end for**

the fact that this optimization algorithm discretizes alternatively the predictor and target variables, it is important to notice that the criterion in (1) is not symmetrical in I and J. In other words, for a given target discretization, the criterion to minimize is not identical to the criterion to minimize given a predictor discretization.

The evaluation criterion $c(M)$ given in Formula (1) is related to the probability that a regression 2D discretization model M explains the target variable. In previous work (see Boullé and Hue, 2006), we propose to use it to build a relevance criterion for the predictor variables in a regression problem. The predictor variables can be sorted by decreasing probability of explaining the target variable. In order to provide a normalized indicator, we consider the following transformation of c:

$$g(M) = 1 - \frac{c(M)}{c(M_\emptyset)}, \tag{2}$$

where $M_\emptyset$ is the null model with only one interval for the predictor and target variables. This can be interpreted as a compression gain, since negative log of probabilities are no other than coding lengths (see Shannon, 1948). The compression gain $g(M)$ holds its values between 0 and 1, since the null model is always considered in our optimization algorithm. It has value 0 for the null model and is maximal when the best possible explanation of the target ranks conditionally to the predictor ranks is achieved.

Our method is non parametric both in the statistical and algorithmic sense: no statistical hypothesis needs to be done on the data distribution (like Gaussianity for instance) and, as the criterion is

regularized, there is no parameter to tune before minimizing it. This strong point enables to consider large data sets.

To our knowledge, few other works address the problem of discretization for regression problems. Nevertheless, we can cite a three-step approach proposed in the approach of Ludl and Widmer (2000): they first propose an equal width pre-discretization of the continuous predictors. These pre-discretizations are next projected onto the target values. A postprocessing consists in merging the split points found according to an algorithm inspired by edge detection concepts. The drawbacks of such an algorithm are the unsupervised way the predictor pre-discretizations are led and the tuning of the merging parameter needed in the postprocessing step.

2.2 The 2D Discretization-Grouping for Probabilistic Regression with Categorical Predictor

Figure 3: Equal-width histograms of the age variable according to each level of the workclass factor

In order to illustrate the regression problem with a categorical predictor, we present in Figure 3 the equal-width empirical histograms of the target age variable for each value of the predictor workclass variable from the 48842 instances of the adult data set from the UCI repository (see D.J. Newman and Merz, 1998). The workclass variable clearly influences the distribution of the age variable. The four values Federal-gov, Local-gov, Self-emp-inc and Self-emp-not-inc lead to similar histograms with a maximum for the $40-45$ interval. The Private value gives a distinct histogram with a maximum for the $20-25$ interval and the Stat-gov value leads to an histogram between theses two groups. The frequencies of the Never-worked and Without-pay values seems too low to constitute significant groups.

An example of discretization/grouping is shown for this data set on Figure 4 with 3 groups and 7 age brackets. Let us now formalize this approach using the MODL approach to explain how optimal Discretization/Grouping can be obtained.

Definition 4 *A regression discretization/grouping model is defined by:*

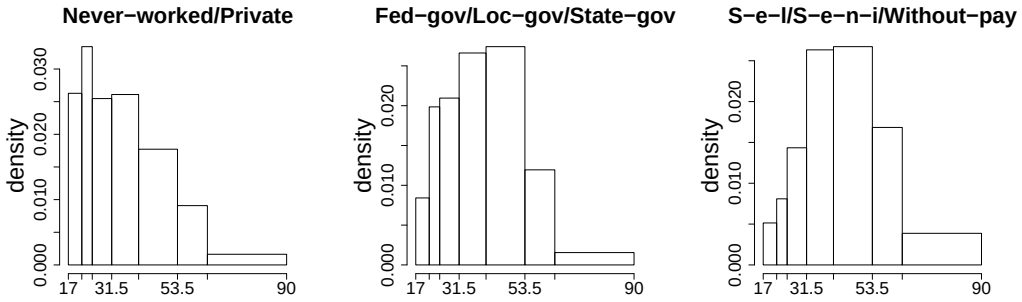


Figure 4: Histograms of the age variable for three groups of the workclass factor

1. *a number of intervals for the target variable and a number of groups for the predictor variable;*
2. *a partition of the predictor variable in a finite number of groups;*
3. *for each predictor group, the repartition of the instances among the target intervals specified by the instance counts locally to each predictor group.*

Notations

N : the number of training instances

V : the number of predictor values

I : the number of predictor groups

J : the number of target intervals

$\iota(\nu)$: the group index the ν value belongs to

$N_{i.}$: the number of instances in the predictor group i

$N_{.j}$: the number of instances in the target interval j

N_{ij} : the number of instances in the grid cell associated to predictor group i and the target interval j.

A regression discretization/grouping model is then entirely characterized by the parameters $\left\{I, J, \{\iota(\nu)\}_{1\leq i\leq V}, \{N_{ij}\}_{1\leq i\leq I, 1\leq j\leq J}\right\}$. The number of instances $N_{.j}$ can be deduced by adding the N_{ij} for each predictor group.

We adopt the following uniform hierarchical prior for the parameters of regression discretization/grouping models:

Definition 5 *The prior for the parameters of a regression discretization/grouping model is chosen hierarchically and uniformly at each level:*

1. *the number of groups I is uniformly distributed between* 1 *and* V,
2. *the numbers of intervals J is independent from the number of groups, and uniformly distributed between* 1 *and* N,
3. *for a given number of groups I, every partition of the predictor values into I groups is equiprobable,*
4. *for a given predictor group, every distribution of the instances on the target intervals is equiprobable,*
5. *the distributions of the target intervals on each predictor group are independent from each other,*
6. *for a given target interval, every distribution of the rank of the target values is equiprobable.*

The definition of the regression discretization/grouping model space and its prior distribution leads to the evaluation criterion given in Formula (3) for a discretization/grouping model M:

$$\begin{aligned} c(M) = \log(V) + \log(N) + \log B(V,I) + \sum_{i=1}^{I} \log \left(\begin{array}{c} N_{i.} + J - 1 \\ J - 1 \end{array} \right) \\ + \sum_{i=1}^{I} \log \frac{N_{i.}!}{N_{i,1}! N_{i,2}! \dots N_{i,J}!} + \sum_{j=1}^{J} \log N_{.j}!, \end{aligned} \tag{3}$$

where $B(V,I)$ is the number of ways to partition V values into I groups (possibly empty). For $I = V$, $B(V,I)$ corresponds to the Bell number. In general, $B(V,I)$ can be written as a sum of Stirling numbers of the second kind $S(V,i)$ (number of ways to partition a set of V values into i nonempty subsets) (see Abramowitz and Stegun, 1970):

$$B(V,I) = \sum_{i=1}^{I} S(V,i).$$

This criterion can be deduced from the grouping criterion in classification (see Boullé, 2005) and the 2D discretization criterion in regression presented in the previous section.

3. From 2D Partitioning to Rank Predictive Cumulative Distribution Estimate

In this section we expose how we can build a univariate estimator of the rank predictive density from each 2D partition and how to obtain multivariate predictors under the naive Bayes assumption.

3.1 From Values to Normalized Training Ranks and Vice-Versa

As seen in the precedent section, the MODL partitions are defined only with the ranks of the training instances and not with their values. Given N_T numerical training values $D^T = (y_1^T, \dots, y_{N_T}^T)$, the N_T ranked values are noted $y_{(1)}^T, \dots, y_{(N_T)}^T$ once the training values have been sorted.

A partition of the N_T ranked instances defined by J numbers $N_1, \dots, N_J$ such that $\sum_{j=1}^{J} N_j = N_T$ is associated to a partition of the values as follows: we define $J-1$ boundaries $b_1, \dots, b_{J-1}$ by $b_j =$

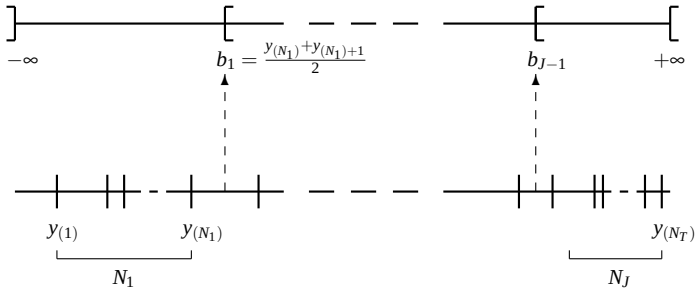


Figure 5: Value partition from the partition frequencies: for example, the upper bound of the first interval containing N_1 instances is the mean of the last value of this interval and of the first value of the second interval.

$\frac{y^T_{(\beta_j)}+y^T_{(\beta_{j+1})}}{2}$ where $\beta_j = \sum_{l=1}^{j} N_l$ and the J value partition intervals are $]-\infty, b_1[, [b_1, b_2[, \ldots, [b_{J-1}, +\infty[$. For a numerical value, its rank interval index is equal to its value interval index. This value partition from the partition frequencies is illustrated in Fig 5.

As presented in the introduction, ordinal regression aims at predicting an ordinal variable which takes a finite number of ordered values, most of the time already known in advance. In our case, we aim at giving a finer grain prediction by considering the set of the N_T possible ranks of a training data set of size N_T. In order to manipulate normalized values in $[0,1]$, we consider the N_T elementary equal-width rank intervals of $[0,1]$ denoted by Te_n for $n = 1, \ldots, N_T$ and equal to $Te_1 = [0, \frac{1}{N_T}[, Te_2 = [\frac{1}{N_T}, \frac{2}{N_T}[, \ldots, Te_{N_T} = [\frac{N_T-1}{N_T}, 1]$. These intervals are centered on the normalized ranks $R_{D^T}(y^T_{(n)}) = \frac{1}{2N_T} + \frac{n}{N_T}$ of the training instances obtained by projection on $[0,1]$ of the rank of $y^T_{(n)}$ among D^T. This normalization is illustrated in Fig 6.

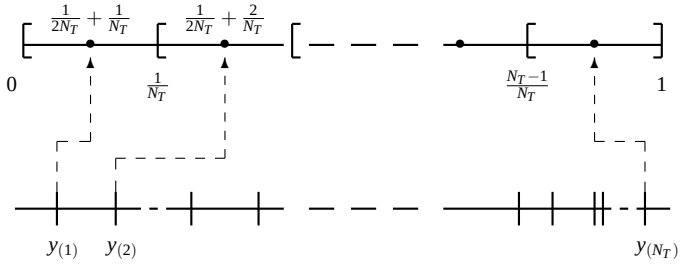


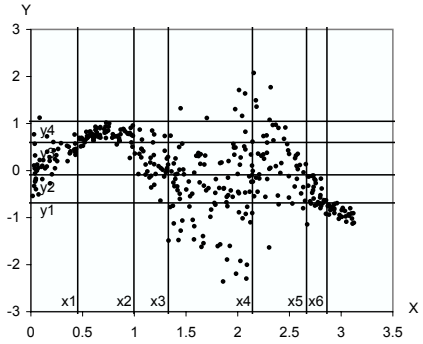
Figure 6: From sorted values to normalized ranks: for example, the normalized rank related to the first value $y^T_{(1)}$ is $R_{D^T}(y^T_{(1)}) = \frac{1}{2N_T} + \frac{1}{N_T}$, which is the center of the first elementary interval $Te_1 = [0, \frac{1}{N_T}[$.

In practice, there may be equal values in D^T. In this case, we affect the averaged rank to the concerned instances. For a new value y unseen during training, we define its rank $R_{D^T}(y)$ as the average of the normalized ranks of $y^T_{(n_1)}$ and $y^T_{(n_2)}$ such that $y^T_{(n_1)} \leq y < y^T_{(n_2)}$. The integers n_1 and n_2 may not be consecutive if one of them is associated with several equal values. In the rest of the paper, we use either ranks or values depending on the context.

We now detail how to build an estimate of the predictive cumulative distribution function of the target standardized rank from a univariate MODL 2D discretization in a first section and from multiple univariate partitionings in a second section.

3.2 Univariate Case

We illustrate the construction of the univariate estimator from the 2D partitioning on a synthetic data set proposed during the recent *predictive uncertainty in environmental modelling* competition (see Cawley et al., 2006). This data set, called synthetic, contains $N_T = 384$ training instances and one numerical predictor. The scatter plot and the optimal MODL partition are presented in Figure 7. The optimal MODL rank intervals are denoted P_i for $i = 1,\ldots,7$ for predictor and T_j, $j = 1,\ldots,5$ for the target. The first and the last rank intervals are of the form $[0, k_f/N_T[$ and $[k_l/N_T, 1]$ respectively and the other intervals are of the form $[k_1/N_T, k_2/N_T[$. The *value* interval bounds $x_1,\ldots,x_6$ and $y_1,\ldots,y_4$ are obtained by projecting the frequencies partition on the value partition as described in 3.1. For the predictor component x of a new instance whose rank range is $P_i(x)$, the number of



	P1	P2	P3	P4	P5	P6	P7	Total
T5	1	0	0	6	5	0	0	12
T4	4	68	5	2	9	0	0	88
T3	38	5	26	32	24	0	0	125
T2	9	0	6	33	15	21	0	84
T1	0	0	0	34	4	9	28	75
Total	52	73	37	107	57	30	28	384

Figure 7: Scatter plot, 2D partitioning and numbers of the MODL grid for the synthetic data set.

examples in each grid cell give us an estimator of the probability that the target standardized rank belongs to a given range T_j:

$$P_{Modl}\left(R_{D^T}(y) \in T_j \mid R_{D^T}(x) \in P_i(x)\right) = \frac{N_{ij}}{N_{i.}}.$$

Assuming that the conditional rank density is constant over each rank interval, the probabilities of the elementary intervals Te_n for $n = 1,\ldots,N_T$ are given by:

$$P_{Modl}\left(R_{D^T}(y) \in Te_n \mid R_{D^T}(x) \in P_i(x)\right) = \frac{N_{ij}}{N_{i.}N_{.j}} \text{ for } j \text{ such that } Te_n \subset T_j. \tag{4}$$

We obtain an estimate of the n^{th} N_T-quantile n/N_T of the conditional cumulative distribution for $n = 1,\ldots,N_T$ by summing these elementary probabilities.

The MODL estimators are plotted for each of the seven rank predictor ranges on Figure 8 for the synthetic data set. The marks on the $x-$axis correspond to the normalized target ranks of the boundaries training instances exhibited by the target partition: $\frac{75}{384} \approx 0.19$, $\frac{75+84}{384} \approx 0.41$, $\frac{75+84+125}{384} \approx 0.74$ and $\frac{75+84+125+88}{384} \approx 0.97$ which correspond to the projection of the value bounds y_1, y_2, y_3 and y_4

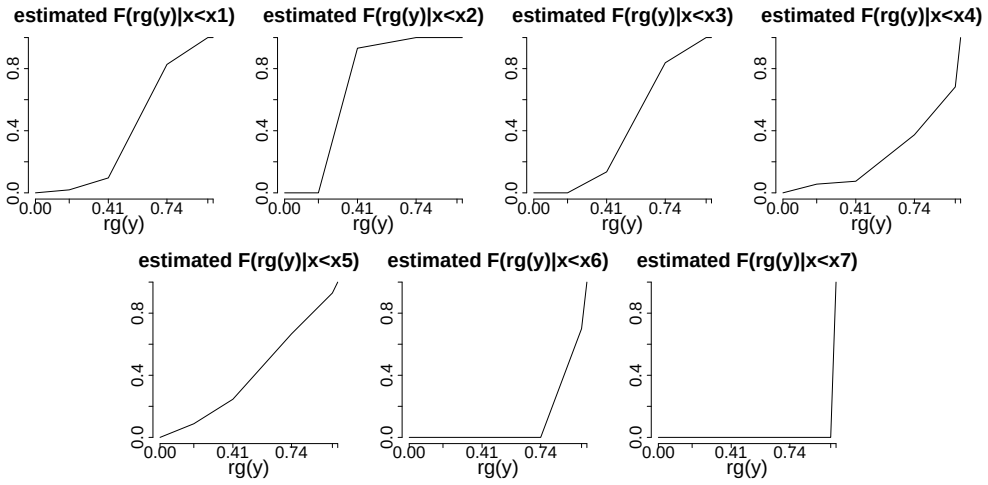


Figure 8: MODL estimators of the conditional univariate standardized rank cumulative distribution for the seven predictor rank ranges.

on the normalized ranks. The shape differences illustrate the seven distinct zones characterized by the MODL optimal partition.

In the case of categorical predictors, the rank predictive cumulative distribution estimator can be obtained in the same way by replacing the predictor intervals by predictor groups.

3.3 Multivariate Case

In the case of several predictors (K with $K > 1$), a first approach is to build an estimator under the naive Bayesian assumption that the predictors are independent given the target. Let $x = (x^1, \ldots, x^K)$ be the coordinates of a new instance in the predictors space and $P^k_{i_k}(x)$ the discretization interval (or group of values) to which belongs each component x^k and $R^k_{D^T}(x)$ its rank. Under the naive Bayesian assumption, the elementary probability can be written :

$$\begin{aligned} & P\left(R_{D^T}(y) \in Te_n \mid (R^1_{D^T}(x), \ldots, R^K_{D^T}(x)) \in (P^1_{i_1}(x), \ldots, P^K_{i_K}(x))\right) \\ & \propto P\left(R_{D^T}(y) \in Te_n\right) \prod_{k=1}^{K} P\left(R^k_{D^T}(x) \in P^k_{i_k}(x) | R_{D^T}(y) \in Te_n\right) \\ & = P\left(R_{D^T}(y) \in Te_n\right) \prod_{k=1}^{K} \frac{P(R_{D^T}(y) \in Te_n | R^k_{D^T}(x) \in P^k_{i_k}(x)) P(R^k_{D^T}(x) \in P^k_{i_k}(x))}{P(R_{D^T}(y) \in Te_n)}. \end{aligned} \tag{5}$$

This last expression can be estimated with the instance numbers in the 2D grid cells. The first factor $P(R_{D^T}(y) \in Te_n)$ can be estimated by the empirical probability $1/N_T$. Each factor of the product can be computed from the numbers of the 2D partitioning of the target and of the k^{th} predictor, denoted $(I_k, J_k, N^k_{i_k.}, N^k_{i_k j_k})$:

$$P(R_{D^T}(y) \in Te_n | R^k_{D^T}(x) \in P^k_{i_k}(x)) = \frac{N^k_{i_k j_k}}{N^k_{i_k.} N^k_{.j_k}} \text{ according to (4).}$$

$$P(R^k_{D^T}(x) \in P^k_{i_k}(x)) = \frac{N^k_{i_k\cdot}}{N_T} \text{ and } P(R_{D^T}(y) \in Te_n) = \frac{1}{N_T}.$$

Each factor reduces to the fraction $\frac{N^k_{i_k j_k}}{N^k_{\cdot j_k}}$ and the elementary probabilities in Formula 5 reduce to $\frac{1}{N_T}\prod_{k=1}^{K}\frac{N^k_{i_k j_k}}{N^k_{\cdot j_k}}$.

However, the independence hypothesis assumed in the naive Bayes predictor is usually violated for real data sets. In this case, estimates of the conditional probabilities are deteriorated as already noticed in the work of Frank et al. (1998). For classification tasks, variable selection has been employed to build selective naive Bayes classifiers. This procedure reduces the strong bias of the naive independence assumption. The objective is to search among all the subsets of variables, in order to find the best possible classifier, compliant with the naive Bayes assumption. Several selection criteria have been tested, such as the accuracy criterion (see Langley and Sage, 1994), the area under receiver operating characteristic (ROC) curve (see Provost et al., 1998) or the posterior probability of the model given the data proposed in the work of Boullé (2007). In this last case, the posterior probability is written as the sum of the prior probability of the model and of the likelihood of the data given the model. The prior is chosen such that each specific small subset of variables has a greater probability than each specific large subset of variables in order to favour small models. For a given subset of variables, the likelihood is computed using the naive Bayes assumption.

We propose here to build selective naive Bayes rank predictors using a MAP approach as in Boullé (2007). The extension to rank regression tasks is straightforward: the prior law remains unchanged and, for a given subset of variables, the likelihood of the ranks of the instances are computed assuming the naive Bayes assumption according to (5).

To summarize this section, we have exposed how the 2D partitionings give us estimators of quantiles of the univariate or multivariate rank cdf. The choice of the estimated quantiles are given by the univariate partitionings. Let us see, in the next section, on which criterion such probabilistic rank models can be evaluated.

4. Performance Evaluation for Probabilistic Rank Regression

For deterministic predictive models, performance evaluation consists in evaluating the distance between the predicted class or value and the true class or value. Depending on the metric used, several performance measures can be used such as the mean absolute error or the mean squared error.

For probabilistic predictive models, performance evaluation consists in comparing the true value or class with an estimate of its conditional cumulative distribution function (cdf) or an estimate of its conditional probability density function (pdf). It is measured through a score function which can be negatively oriented (large values imply poor performance) or positively oriented (large values imply high performance). A score function is said to be proper if its expectation is maximized (or minimized) for the true predictive distribution. It is said strictly proper if this optimum is unique.

Among the strictly proper scoring functions, the two more commonly used are the logarithmic and the quadratic scores (see Gneiting and Raftery, 2004). They take different forms depending on the learning task. In the sequel, we first describe the quadratic and the logarithmic scores and their use in classification and regression. We then present an interesting way to build the logarithmic score function from ranks rather than from values.

4.1 Performance Evaluation for Probabilistic Classification and Regression

The quadratic score is called the Brier score for binary classification (see Brier, 1950). For classification with a finite number of ordered classes $j = 1,\ldots,J$ (ordinal regression), the discrete ranked probability score, DRPS, has been proposed in the work of Epstein (1969):

$$S_{DRPS}(\hat{p},(x,y)) = 1 - \frac{1}{I-1}\sum_{j=1}^{J}(\hat{P}(Y \leq j \mid x) - 1_{\{Y \geq j\}}(y))^2.$$

Its extension to the continuous case (metric regression) is the continuous ranked probability score, CRPS (see Matheson and Winkler, 1976):

$$S_{CRPS}(\hat{p},(x,y)) = -\int_{-\infty}^{+\infty}(\hat{P}(Y \leq u \mid x) - 1_{\{Y \geq u\}}(y))^2 du.$$

The CRPS is a bounded global score which manipulates the cumulative distribution function. Its main disadvantage is that it is generally not a closed form. Nevertheless, a closed form has been derived in the Gaussian case and in the case of ensemble prediction systems where the cdf is piecewise constant (see Hersbach, 2000).

The logarithmic score is commonly called the ignorance score for classification and has been introduced in the work of Good (1952). For value regression it is called the negative log-likelihood (NLL) or negative log predictive density (NLPD). It takes the negative logarithm of the posterior class probabilities for classification and of the predictive density for regression:

$$S_{NLPD}(\hat{p},(x,y)) = -\log(\hat{p}(y|x)). \tag{6}$$

The NLPD is a local score as it only depends on the predictive density on the example. It is clearly negatively oriented.

Practically, for regression, this criterion requires the definition of the probability density function on any point. If the target distribution is described by a sample, its pdf is not defined. A common choice to specify such predictive distribution is to describe the cdf by a set of N quantiles of the form:

$$\alpha^x(n) = \hat{P}(y < q^x(n) \mid x) \text{ for } n = 1,\ldots,N.$$

Moreover, assuming that the cdf is constant on each interval $[q^x(n);q^x(n+1)[$, one obtains the following expression for the logarithmic score:

$$S_{NLPD}(\hat{p}(\{q^x(n)\}),(x,y)) = -\log\left(\frac{\alpha^x(n_y) - \alpha^x(n_y-1)}{q^x(n_y) - q^x(n_y-1)}\right), \tag{7}$$

where $q^x(n_y-1) \leq y < q^x(n_y)$ and $\hat{p}(\{q^x(n)\})$ designs the estimator specified by the quantiles $\{q^x(n)\}$.

Such as defined in Equation (6), the NLPD presents the main drawback *not to be minored*: the more the density is peaked around the sample values, the smaller the score values. It is not a problem in classification as the posterior probabilities are bounded but, for regression, it encourages dishonest predictions concentrated around some specific values, as noticed in recent challenges (see Kohonen and Suomela, 2005). In order to achieve a fair evaluation, if the predictive density is expressed as a set of quantiles as exposed before, a last resort is to impose a minimum width for

scoring rule	quadratic score	logarithmic score
binary classification	Brier score	Ignorance score
ordinal regression	DRPS	
value regression	CRPS	NLPD

Table 1: Quadratic and logarithmic scores for performance evaluation of probabilistic predictive models

the intervals $[q^x(n-1);q^x(n)[$. Anyway, such a score function contributes to confusing performance prediction since arbitrary small values can be obtained fortunately.

The two scores mentioned above are presented in Table 1.

Let now study what happens if we compute the logarithmic score for the rank predictive density rather than for the value predictive density.

4.2 A Robust Logarithmic Score Defined on the Rank Predictive Density

Given a training data set set $D^T = (x_n^T, y_n^T)_{n=1,\ldots,N_T}$, we assume that we have at our disposal a rank probabilistic estimator specified for any given x by the N_T estimated quantiles $\alpha^x(n)$ for $n = 1,\ldots,N_T$ of the standardized rank cumulative distribution function:

$$\alpha^x(n) = \hat{P}(R_{D^T}(y) < \frac{n}{N_T}|x) \text{ for } n = 1,\ldots,N_T.$$

We can immediately see that this assumption is not restrictive: knowing the values associated to the ranks, each *value* estimator gives us an estimator of the cdf on the N_T target normalized *ranks* of the training data set. If we denote by $y_{(n)}$ the n^{th} target value of D_T, we have :

$$\alpha^x(n) = \hat{P}(R_{D^T}(y) < \frac{n}{N_T}|x) = \hat{P}(y < \frac{y_{(n)} + y_{(n+1)}}{2}|x). \tag{8}$$

By appropriate integration of the cdf, each *value* estimator specified by N given quantiles gives us the N_T quantiles estimates defined in (8).

Let us come back to the evaluation of such a rank probabilistic estimator with the logarithmic score function. It consists in comparing it with the standardized insertion rank $R_{D^T}(y)$ of the true value y among the training data set D^T. We propose to estimate the predictive density on the insertion rank by the ratio:

$$\hat{p}(R_{D^T}(y)|x) \approx \frac{\hat{P}(R_{D^T}(y) \in Te_{n_y} \mid x)}{1/N_T},$$

where Te_{n_y} is the elementary interval to which the insertion rank $R_{D^T}(y)$ belongs.

This approximation enables us to define the Negative Log Rank Predictive Density as follows:

Definition 6 *Let a training data set set $D^T = (x_n^T, y_n^T)_{n=1,\ldots,N_T}$ and $\alpha^x(n)$ for $n = 1,\ldots,N_T$ some estimates of the N_T quantiles of the standardized rank cumulative distribution function:*

$$\alpha^x(n) = \hat{P}(R_{D^T}(y) < \frac{n}{N_T}|x) \text{ for } n = 1,\ldots,N_T.$$

Rank predictive distribution specification: Let $D^T = (x_n^T, y_n^T)_{n=1,\ldots,N_T}$ be a training data set with Y a numerical target and $X = (X^1, \ldots, X^K)$ K numerical or categorical predictors. For any given x, the rank predictive distribution is specified by the N_T quantiles of the rank cdf:

$$\alpha^x(n) = \hat{P}(R_{D^T}(y) < \frac{n}{N_T} \mid x) \text{ for } n = 1, \ldots, N_T.$$

Rank predictive distribution evaluation: Let $D^V = (x_n^V, y_n^V)_{n=1,\ldots,N_V}$ be a validation data set. Compute the logarithmic score of the rank predictive distribution on data set D^V as

$$NLRPD = -\frac{1}{N_V} \sum_{n=1}^{N_V} \log \frac{\alpha^{x_n^V}(n_y) - \alpha^{x_n^V}(n_y - 1)}{1/N_T},$$

where $R_{D^T}(y_v)$, the standardized insertion rank of the true value y_n^V among the training data set D^T, is included in the elementary interval Te_{n_y}.

Table 2: Summary of the proposed approach for probabilistic standardized rank regression and its evaluation.

The score function for the negative log rank predictive density is then defined by:

$$\begin{aligned} S_{NLRPD}(\hat{P}(N_T), (x,y)) &= -\log \frac{\hat{P}(R_{D^T}(y) \in Te_{n_y} \mid x)}{1/N_T} \\ &= -\log(\alpha^x(n_y) - \alpha^x(n_y - 1)) - \log(N_T), \end{aligned} \tag{9}$$

where $\hat{P}(N_T)$ is the rank predictive density estimator specified by the quantiles $\{\frac{n}{N_T}\}$. We now present two interesting properties of the NLRPD.

Theorem 7 *In absence of predictive information the NLRPD is equal to zero.*

Without any predictive information, the probability that the insertion rank of a new instance belongs to a given interval is simply equal to $1/N_T$. By construction, we have then for the uniform predictor $\hat{P}^{unif}(N_T)$:

$$S_{NLRPD}(\hat{P}^{unif}(N_T), (x,y)) = -\log 1 = 0.$$

□

Moreover, unlike the NLPD score, this score function on ranks has the great advantage to be minored as it is precised in the following property:

Theorem 8 *Given a training data set set D^T of size N_T, the NLRPD score function is minored by* $-\log(N_T)$.

This bound is directly obtained by considering that the difference $\alpha^x(n_y) - \alpha^x(n_y - 1)$ belongs to $]0;1]$.□

By construction, the proposed NLRPD score is then bounded. However, it depends on the training data set through its size. We will see in next section its relative insensitivity to this size.

Let us now examine the link between the score function on ranks and the score function on values. Using the expression (7) with the $N_T - 1$ quantiles $b_1 = \frac{y_{(1)}+y_{(2)}}{2}$, $b_2 = \frac{y_{(2)}+y_{(3)}}{2}$, ..., $b_{N_T-1} = \frac{y_{(N_T-1)}+y_{(N_T)}}{2}$, we obtain the relation :

$$S_{NLRPD}(\hat{P}(N_T),(x,y)) = S_{NLPD}(\hat{p}(\{b_{n_y}\}),(x,y)) - \log(N_T) - \log(b_{n_y} - b_{n_y-1}).$$

As precised at the beginning of the section, the NLRPD score function can be used by converting any value predictive density estimator to rank predictive density estimator specified or using the above relation between NLRPD and NLPD score functions.

The framework of our approach is summarized in Table 2.

5. Experimental Evaluation

We first present experiments about the criteria proposed in the precedent section. Then, we focus on the quality of the 2D-partitioning with experiments on synthetic data. We finish by experiments with the univariate and multivariate predictors presented on five real data sets.

5.1 Experiments on the NLRPD

We focus here on the properties of the proposed criterion, the NLRPD. We consider the data generated according to the following heteroscedastic model (used for the synthetic data set of the *predictive uncertainty in environmental modeling competition*):

$$\begin{cases} x_n \sim U_{[0\pi]} \\ y_n \sim \mathcal{N}\left(\sin\left(\frac{5x}{2}\right)\sin\left(\frac{3x}{2}\right), \frac{1}{100} + \frac{1}{4}\left(1-\sin\left(\frac{5x}{2}\right)\right)^2\right) \end{cases}$$

This data set has been used in Section 3 to describe the building of the rank conditional densities. Knowing the true cdf of the synthetic data set, we can compute the true NLRPD by using the true probabilities instead of their estimates in (9).

The contentious aspect of the NLRPD is that it depends on the training data set size. The objective of this experiment is then to study the sensitivity of the NLRPD with respect to the training size in a first time and to the validation data set size in a second time. For that, we have generated $m = 100$ training data sets of size $N_T = 2^n$ for $n = 1, \ldots, 12$. The true NLRPD has been computed given each of the $m * 12$ training quantile vectors for a test data set of size $N_V = 1000$ and another one of size $N_V = 10000$. The mean NLRPD over the m training data sets is plotted versus the training data set size N_T on left of Figure 9 for $N_V = 1000$ and on right for $N_V = 10000$.

First, this plot shows a threshold around a training data set size of $N_T = 100$ instances. Below this threshold, the true NLRPD decreases when the training set size increases. Above this threshold, the optimal NLRPD seems insensitive to the training set size. Secondly, we can notice that both curves for $N_V = 1000$ and $N_V = 10000$ are very similar. This allows us to think the NLRPD is not very sensitive to the test data set size. To confirm this fact, we have fixed a training data set of size 384 and we have computed the true NLRPD for $m = 100$ test data sets of size $N_V = 1024$. The standard-deviation obtained is around 3%. Our criterion looks robust with respect to the training and validation data set size.

5.2 Experiments on the 2D-Partitioning

In this section, we focus on the quality of the 2D-partitioning with three experiments.

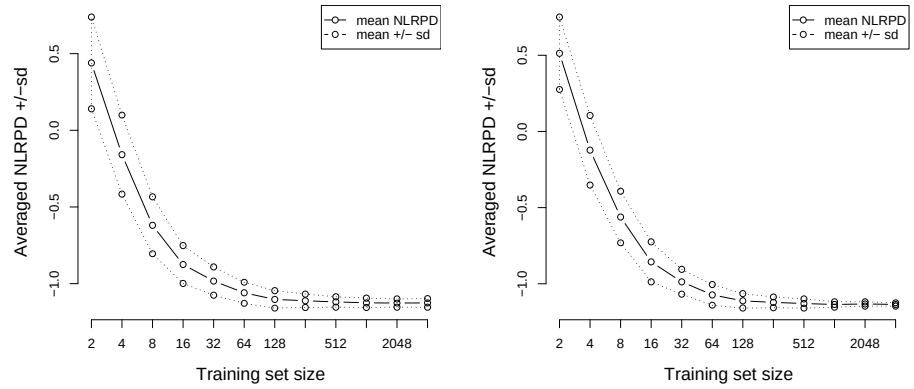


Figure 9: Mean NLRPD over $m = 100$ training data sets of size $N_T = 2^n$, $n = 1, \ldots, 12$ and for a validation data set of size $N_V = 1000$ on the left and $N_V = 10000$ on the right.

A strong point of the MODL approach is that it does not presuppose the existence of a relation between the predictor and the target variables. In the case of numerical predictors, the absence of relevant information should conduct to elementary 2D-partitions consisting of one single cell. To check the quality of the MODL 2D discretizations, we test it on several *noise pattern* data sets: we generate 10^5 training data sets of size equal to 2^n for $n = 2, \ldots, 10$, for which the predictor and target variables are generated independently according to a uniform law in $[0, 1]$. As the predictor variable contains in fact no relevant information to predict the target variable, we expect that the number of predictor intervals I and of target intervals J would be equal to one. Mean target interval number versus the data set size is plotted in Figure 10. For the two smallest sizes 4 and 8, the number of target intervals is always equal to one as there are not enough instances to constitute any pattern. For larger sample sizes, the number of intervals is sometimes equal to two with an exponential decreasing frequency after a peakly increasing for sizes 16 and 32. In the worst case for $N_T = 32$, some predictive information is wrongly detected in 0.8% of the realizations and it falls to 0.2% for $N_T = 128$ and 0.006% for $N_T = 1024$. The absence of predictive information is then almost always detected.

Secondly we test the capacity of our method to detect predictive information contained in a noisy XOR pattern. A XOR pattern is obtained by generating a predictor variable uniformly in $[0, 1]$ and the target variable uniformly in $[0, 0.5]$ if the predictor variable is less than 0.5 and uniformly in $[0.5, 1]$ otherwise. In this case, the expected partition is the one with $I = J = 2$ intervals which splits the predictor and the target variable on 0.5. Some noise instances generated as in the previous noise pattern are added to constitute what we call a noisy XOR pattern. We can study the robustness of the discretization to noise. In Figure 11, the mean number of target intervals on 100 samples are plotted versus the data set size for noise rates varying from 0 to 1 by 0.2.

Each curve shows a sharp threshold above which the pattern is correctly detected. First, a XOR pattern without noise is correctly detected from $N_T = 16$. The resultat is similar for low noise rate

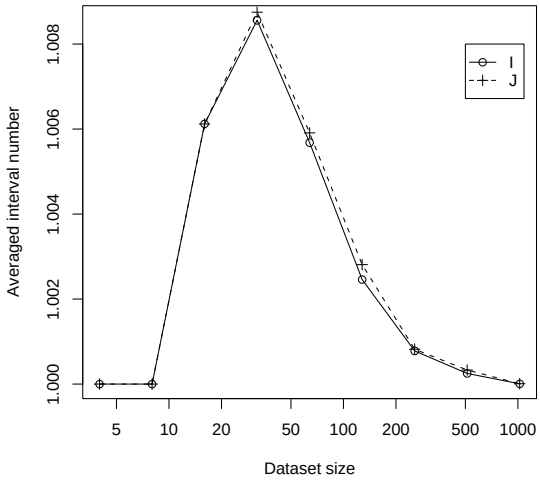


Figure 10: Mean target and predictor interval numbers for 100000 noise pattern data sets of size 4 to 1024

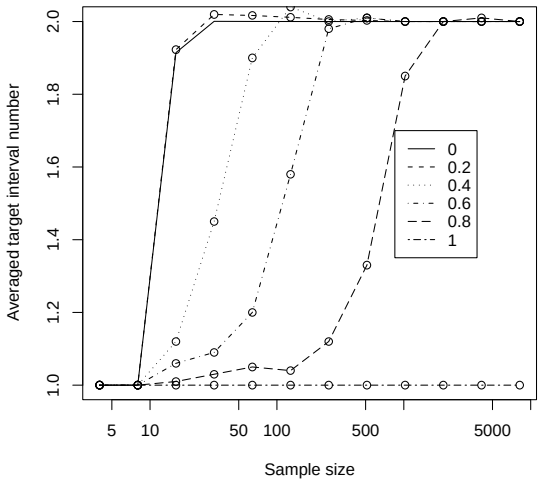


Figure 11: Mean target interval number for 100 noisy XOR pattern data sets of size 4 to 5096 for noise rate $= 0, 0.2, 0.4, 0.6, 0.8, 1$.

equal to 0.2. This threshold increases with the noise rate and reachs respectively 128, 256 and 1024 for a rate equal to 0.4, 0.6 and 0.8. Our discretization is then robust to noise rate.

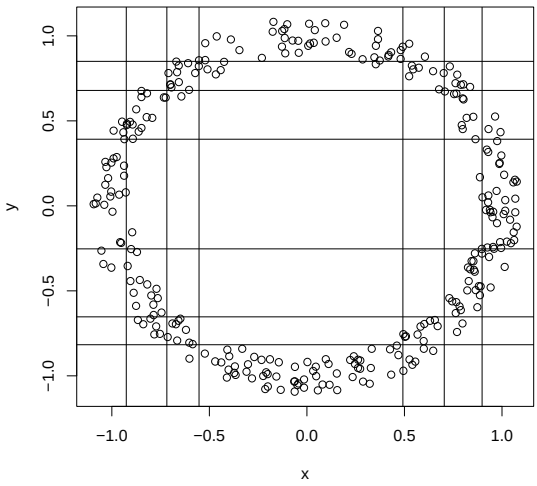


Figure 12: Optimal MODL 2D-partition for data on a noisy circle.

Thirdly, we test the capacity of our method to detect multimodality. For that we generate 300 instances on a noisy circle as showed in Figure 12. The density of y conditionally to x is bimodal for most values of x. As there is no assumption in the MODL approach about the form of the conditional distribution, the 2D-partitioning can produce multimodal conditional densities. For the *noisy circle* data set, the optimal MODL 2D-partition plotted in Figure 12 clearly shows the two modes of the law.

5.3 Experiments on Real Data Sets

In this section, we test the estimators proposed in Section 3 on real data . We have chosen the following five regression data sets, detailed in Table 3:

- SO2, precip and temp, available from the predictive uncertainty competition website (http://theoval.cmp.uea.ac.uk/competition/).
- Adult and housing (Boston), available from the UCI machine learning repository (http://www.ics.uci.edu/ mlearn/MLSummary.html);

Data Set	SO2	Precip	Temp	Adult	Housing
Numerical predictors	27	106	106	6	13
Categorical predictors	0	0	0	7	0
Number of patterns	22956	10546	10675	48842	506

Table 3: Summary of the dimensions of five data sets chosen to evaluate the probabilistic rank predictive density estimators.

For each data set, we have performed a five-fold cross validation for the three following estimators:

SO2	Level	I	J	Precip	Level	I	J	Temp	Level	I	J
V7	0.01467	9	7	V3	0.01986	5	6	V102	0.1234	13	14
V25	0.00993	7	6	V35	0.01858	5	6	V104	0.1095	12	13
V27	0.009658	7	7	V4	0.01823	5	6	V101	0.1069	13	12
V2	0.009483	6	28	V81	0.01729	4	6	V103	0.1039	12	13
V26	0.008302	6	7	V69	0.01716	4	7	V100	0.06891	9	12
V1	0.003991	5	127	V36	0.01674	4	7	V98	0.06288	9	10
V11	0.001366	5	5	V82	0.01653	4	6	V106	0.05691	8	8
V8	0.0009511	3	4	V70	0.01633	4	6	V99	0.05614	8	9
V12	0.0008027	4	3	V57	0.01632	4	6	V97	0.05566	9	10
V18	0.0007759	3	4	V45	0.01629	4	6	V6	0.03347	7	7

Adult	Level	I	J	Housing	Level	I	J
marital-status	0.0244	11	6	LSTAT	0.0945	5	5
relationship	0.01936	12	5	RM	0.06378	5	5
hours-per-week	0.009682	10	6	NOX	0.04466	5	5
education	0.009164	10	9	INDUS	0.04350	3	5
education-num	0.009016	9	9	PTRATIO	0.03779	5	4
class	0.006555	10	2	CRIM	0.03451	4	5
occupation	0.003556	7	7	TAX	0.03352	4	5
workclass	0.002672	6	5	AGE	0.03208	4	3
capital-gain	0.002257	4	18	DIS	0.02929	5	3
sex	0.0007349	3	2	RAD	0.01977	3	2

Table 4: Compression gains (levels) and size (I,J) of the 2D partitions for the ten most informative variables of the 1-fold of the five data sets.

- the univariate estimator built with the most MODL informative variable;
- the multivariate estimator built under the naive Bayes assumption with all the predictor variables (NB);
- the multivariate estimator built under the naive Bayes assumption with the best selected subset of predictor variables (SNB).

For each fold, the estimators are built from the optimal MODL 2D-discretizations for each couple (predictor,target). The compression gain (or level) defined in (2) enables us to rank the predictors. For illustration, Table 4 presents the level and the size (I,J) of the 2D partitions for the best predictors, for the first training set of each data set.

Each estimator is evaluated with the NLRPD criteria proposed in the previous section. Table 5 presents the mean and standard-deviation of the NLRPD for each data set and each estimator.

First, we can see the poor performance of the naive Bayes estimator which exploits all the univariate predictors: for all data sets except for the adult data set, the NLRPD for the NB estimator is positive that is to say it performs not as good as the predictor built without any predictive information. This phenomenon is due to the violation of the naive Bayes assumption. For example, in

	Univariate	SNB	NB
SO2	-0.136 +/- 0.003	-0.162 +/- 0.0076	0.24 +/- 0.036
Precip	-0.165 +/- 0.005	-0.220 +/- 0.023	9.0 +/- 1.36
Temp	-1.018 +/- 0.012	-1.046 +/- 0.0173	5.46 +/- 0.99
Adult	-0.237+/- 0.0048	-0.378+/- 0.03	-0.287 +/- 0.029
Housing	-0.393+/- 0.128	-0.26 +/- 0.26	0.849+/- 0.49

Table 5: Mean and standard-deviation of the NLRPD for the 3 predictors and the five real data sets.

the case of the temp data set, Figure 13 presents the scatter plot of the two most informative variables. The very high linear correlation clearly deteriorates the naive Bayes predictor based on these variables.

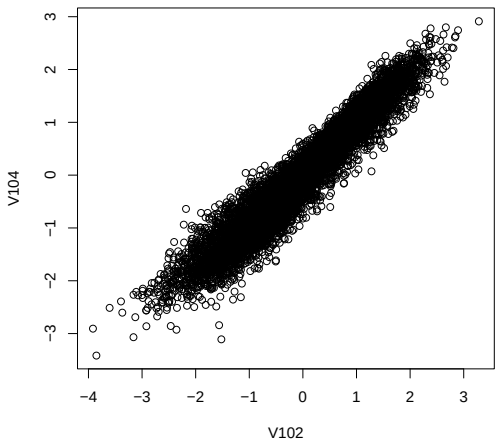


Figure 13: Scatter plot of the two most informative variables for the temp data set.

The second point from these results is the relative good performance of the univariate predictor. For the SO2, precip and temp data sets, it performs nearly as well as the SNB, which selects around five predictors, even if the NLRPD mean for the SNB is significantly lower than for the univariate predictor according to a Student's t-test. These three data sets seem a bit specific in the sense that the most informative variable contains a lot of the predictive information. The good 2D partitioning quality enables to build a very performant univariate predictor. For the adult data set, the SNB, which selects around 9 predictors, performs better than the NB which performs better than the Univariate predictor. The predictive information is shared by several variables and the selection procedure enables to eliminate redundant predictors like education and education-num, or marital-status and relationship. For the housing data set, the univariate predictor performs as well as the SNB which selects around five predictors. The variance of the results is important which certainly explains that the equality hypothesis of the means is not rejected according to a Student t-test. It

NLPD on test set	SO2	Precip	Temp
Organizer's method	4.25 (1st)	-0.509 (1st)	0.053 (2nd)
Best submitted method	4.37 (3th)	-0.279 (3th)	0.034 (1st)
MODL SNB	4.31 (2nd)	-0.437 (2nd)	0.259 (8th)
MODL Univariate	4.33 (2nd)	-0.361 (2nd)	0.284 (8th)
Reference method	4.5 (4th)	-0.177 (4th)	1.30 (9th)

Table 6: NLPD values for each *test* data set for the univariate estimator using the best MODL predictor (MODL univariate), the selective naive Bayes predictor (MODL SNB), the best method in competition, the organizer's submission and the reference method. The ranking of each method is between brackets after each NLPD value.

may seem surprising that the univariate sometimes performs better than the SNB. The reason may be that the selection procedure in the SNB assumes that the univariate predictors are perfect and focuses on the choice of the number of variables. In other words, the uncertainty of the univariate predictive model is not taken into account at this stage. It also suggests that the SNB predictor could be improved. Contrary to the classification case, the target partition is different for each predictor considered. This aspect could be taken into account in the selection of the predictors. Model averaging could also improve our multivariate predictors.

The objective of our last experiment is to compare our approach with other regression methods. To our knowledge, there is no alternative rank regression method available in the literature. We therefore compare it to *value* predictive density estimators. Such estimators being still an active subject of research, we decide to compare our approach to the methods proposed very recently in the predictive uncertainty in environmental modelling competition organized in 2006 by Gavin Cawley. Since these methods are hard to re-implement and tune, we project our rank estimator to a value estimator and we compare them with the NLPD criteria. Knowing the values associated to the ranks, each *rank* estimator gives us an estimator of the cdf on the N_T target values of the training data set using (8). To compute the predictive pdf on any point from the conditional quantiles, we adopt the same assumptions as those used in the challenge, that is that the pdf is assumed uniform between two successive values and that the distribution tails are exponential.[1]

As our approach is implicitly regularized and needs no tuning parameter, we use the training and validation data sets to compute the optimal 2D partitionings. Given the poor performance of the NB in the previous experiments, we only train the univariate and the SNB predictors. Table 6 indicates the NLPD on the test data set for these two MODL estimators, the best method in competition and the reference method which computes the empirical estimator of the marginal law $p(y)$. For the three data sets, the MODL estimators are better than the reference method, that is far from being the case for all submitted methods. Secondly, we observe good performance for the MODL estimators, in particular for the SO2 and Precip data sets where the SNB estimator is at the front after the organizer's method. The good performance of the univariate predictor demonstrates the 2D partitioning quality despite the use of the ranks and not of the values for this step. Moreover, the SNB estimator is always better than the univariate estimator. This proves the presence of additional information and the interest of the selection procedure.

1. For that we affect an $\varepsilon = 1/2N$ probability mass at each tail.

6. Conclusion

We have first proposed a non parametric Bayesian approach for the estimation of the conditional distribution of the normalized rank of a numerical target. Our approach is based on an optimal 2D partitioning of each couple (target, predictor). These partitionings are used to build univariate estimators and multivariate ones under the naive Bayesian assumption of predictors conditional independence, with and without variable selection.

Our approach is applicable for all regression problems with categorical or numerical predictors. It is particularly interesting for those with a high number of predictors as it automatically detects the variables which contain predictive information. As the criteria selection is regularized by the presence of a prior and a posterior term, it does not suffer from overfitting.

Secondly we have proposed a new criterion to evaluate a probabilistic estimator of the rank predictive density. It uses the logarithmic score and presents the main advantage to be minored contrary to the logarithmic score computed for probabilistic estimators of the target *value*. As a value estimator can be projected on a rank estimator, this criterion provides a reliable evaluation criterion for all probabilistic regression estimators on values or on ranks.

Experiments on synthetic data sets show the validity of the proposed evaluation criterion and the quality of the 2D partitioning. Experiments on real data sets show the failure of the naive Bayes but the potential of the selective naive Bayes estimator. A comparison with methods proposed in a recent challenge dedicated to probabilistic metric regression methods evaluates the competitivity of our approach after the projection of our rank estimators on the value range. The very good performance of our best univariate and selective naive Bayes estimators encourages us to work in the future to improve the SNB approach and to evaluate the potential benefit of model averaging.

Acknowledgments

We are grateful to the editor and the anonymous reviewers for their useful comments.

References

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publications Inc., New York, 1970.

M. Boullé. A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research*, 2005.

M. Boullé. MODL: A Bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131–165, 2006.

M. Boullé. Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research*, To appear, 2007.

M. Boullé and C. Hue. Optimal Bayesian 2d-discretization for variable ranking in regression. In *Ninth International Conference on Discovery Science (DS 2006)*, 2006.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

G.C. Cawley, M.R. Haylock, and S.R. Dorling. Predictive uncertainty in environmental modelling. In *2006 International Joint Conference on Neural Networks*, pages 11096–11103, 2006.

P. Chaudhuri and W.-Y. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernouilli*, 8, 2002.

P. Chaudhuri, M.-C. Huang, W.-Y. Loh, and R. Yao. Piecewise-polynomial regression trees. *Statistica Sinica*, 4, 1994.

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

W. Chu and S. Keerthi. New approaches to support vector ordinal regression. In *ICML '05: Proceedings of the 22nd international conference on Machine Learning*, 2005.

K. Crammer and Y. Singer. Pranking with ranking. In *Proceedings of the Fourteenth Annual Conference on Neural Information Processing Systems (NIPS)*, 2001.

C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

E. S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987, December 1969.

J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83:189–196, 1996.

R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, 7, 1936.

E. Frank, L. Trigg, G. Holmes, and I. Witten. Naive Bayes for regression, 1998. URL `citeseer.ist.psu.edu/article/frank98naive.html`. Working Paper 98/15. Hamilton, NZ: Waikato University, Department of Computer Science.

T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction and estimation. Technical report, Department of Statistics, University of Washington, 2004.

I. Good. Rational decisions. *Journal of the Royal Statistical Society*, 14(1):107–114, 1952.

R. Herbrich, T. Graepel, and K. Obermayer. *Large margin rank boundaries for ordinal regression*, chapter 7, pages 115–132. 2000.

H. Hersbach. Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

T. P. Hettmansperger and J. W. McKean. *Robust Nonparametric Statistical Methods*. Arnold, London, 1998.

R. Koenker. *Quantile Regression*. Econometric Society Monograph Series. Cambridge University Press, 2005.

J. Kohonen and J. Suomela. Lessons learned in the challenge: making predictions and scoring them. In *Revised Selected Papers of the 1st PASCAL Machine Learning Challenges Workshop (MLCW, Southampton, UK, April 2005)*, Lecture Notes in Artificial Intelligence 3944, pages 95–116, 2005.

P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406, 1994. URL citeseer.ist.psu.edu/langley94induction.html.

M.-C. Ludl and G. Widmer. Relative unsupervised discretization for regression problems. In *Eleventh European Conference on Machine Learning (ECML-2000)*, pages 246–254, 2000.

J. Matheson and R. Winkler. Scoring rules for continuous probability distributions. *Management Sci.*, 22:1087–1096, 1976.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *In Proc. Fifteenth Intl. Conf. Machine Learning*, pages 445–453, 1998. URL citeseer.ist.psu.edu/provost97case.html.

C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 1948.

A. Shashua and A. Levin. Ranking with large margin principles : two approaches. In *Proceedings of the Fiveteenth Annual Conference on Neural Information Processing Systems (NIPS)*, 2002.

I. Takeuchi, Q.V. Le, T.D. Sears, and Smola A.J. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

H White. Nonparametric estimation of conditional quantiles using neural networks. In *Proceedings of the 1991 Interface Conference*, 1991.