# NEURAL MODELLING OF RANKING DATA WITH AN APPLICATION TO STATED PREFERENCE DATA

C. Krier, M. Mouchart, A. Oulhaj

## 1. INTRODUCTION

Many business and social studies require modeling individual differences in choice behavior by asking respondents to rank alternatives. However, this kind of data present some particularities, related to their non-continuous and bounded character that should be taken into account by the models.

Neural Networks (NN) provide an approach that progressively attract more attention from statisticians working in a wide variety of problems. Some examples issued in *Statistica* give an interesting view of the variety of topics faced with a NN approach. Thus, Apolloni *et al.* (2001) considers a forecasting problem concerning quality characteristics of bovine; Biganzoli *et al.* (2000) proposes the automatic learning process of a NN for the study of complex phenomenon in biostatistic; also Pillati (2001) proposes to combine radial basis function networks and binary classification trees. The object of this work is to model with NN the firm's preferences, in particular the relative importance of each attribute, in the firm's ranking procedure.

The data used to illustrate the method consist of rankings of alternative solutions for freight transport provided by different companies through face-to-face interviews. These transport scenarios are defined by six attributes: frequency of service, transport time, reliability, carrier's flexibility, transport losses, and cost. Further details are given in section data and a more systematic presentation of the data may be found in Beuthe *et al.* (2005).

The paper presents first the data used to illustrate the method. Section 3 describes the assumptions made in connection with the firm's decision rule, and details the form considered for the underlying utility function. The estimation of the

The research underlying this paper is performed within the framework of the second Scientific Support Plan for a sustainable Development Policy (SPSP II) for the account of the Belgian State, Prime Minister's Office - Federal Science Policy Office (contract:CP/17/361). Financial support from the IAP research network nr P5/24 of the Belgian Government (Federal Office for Scientific, Technical and Cultural Affairs) is also acknowledged. Catherine Krier has also been funded by a F.R.I.A. grant. Comments by an anonymous referee and by several participants of seminars where this paper has been presented are gratefully acknowledged.

firm's decision rule is developed in Section 4, which is organized as follows: a general view of the perceptron structure is presented, followed by a short description of some traps which should be avoided. The last part of this section relates to the heuristic chosen to perform the minimization implied by the perceptron algorithm. Section 5 provides information on how the experiments were carried out, shows some results on the data and discusses them.

## 2. NEURAL NETWORKS APPROACH FOR RANKING DATA

Ranking data are obtained when $I$ objects $\mathfrak{x}_i \in \Re^J (i = 1...I)$ are ranked from 1 to $I$. A basic difference between "ordinal data" (*i.e.* data measured on an ordinal scale) and "ranking data" is that ordinal data are measured on a scale with far less degrees than the sample size; in contrast, the scale for ranking data has as many degrees as the sample size.

Thus ties – or *ex aequo* – are dominating in ordinal data, but scarce, and sometimes excluded, in ranking data.

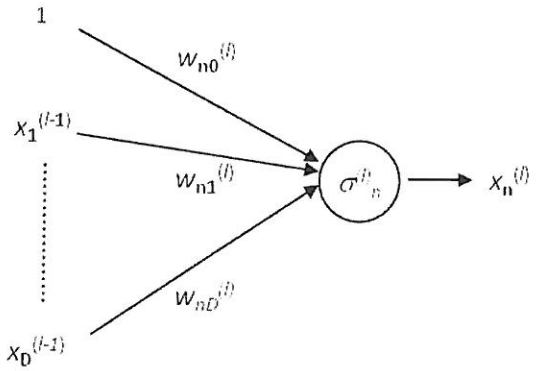


*Figure 1* – Neuron $n$ of layer $(l)$.

Beuthe *et al.* (2008) compares the analysis of ranking data under models adapted from models originally developed for ordinal data (such as ordered logit, conjoint analysis or UTA type models). In theoretical statistics, the distribution of rank statistics has been developed for the case of observable variables. This paper develops a model based on the idea of interpreting ranking data as rank statistics of a latent variable, namely the value of a latent utility function defined on the characteristics of the ranked objects.

The modeling strategy is based on a neural approach. For the sake of more specificity, suppose that we observe the ranking of $I$ objects identified by $J$ characteristics. The data consist therefore of an $(I \times J)$ – matrix $Z = [\mathfrak{x}_{ij}] = [\mathfrak{x}_1, \mathfrak{x}_2, ..., \mathfrak{x}_I]'$ where $\mathfrak{x}_i \in \Re^J$ represents the $J$ characteristics of the $i-th$ object. Furthermore, we have a vector of $I$ declared ranks

$R=(R_1,R_2,...,R_I)$, where $R_i$ denotes the rank of the $i-th$ object. To each object $i$ we associate a latent utility $u_i$ and therefore obtain an I-dimensional latent vector $u=(u_1,u_2,...,u_I)$.

Generally speaking, a multi-layer perceptron consists of several layers of weights and neurons which present the configuration illustrated in Figure 1. The output $x_n^{(l-1)}$ of one neuron can be used as an input for one or several neurons belonging to the next layer. Non-linear activation functions $\sigma_n^{(l)}$ are associated to each neuron. This makes the NN framework suitable for developing learning algorithms as a possible approach to iterative procedures used for complex statistical inferences as exemplified in Section 4. Let us call $w_{nd}^{(l)}$ the weights associated to the neuron $n$ and input $d$ of the layer $l$, the output of the layer l is

$$x_n^{(l)}=\sigma_n^{(l)}\left(\sum_{d=1}^{D} w_{nd}^{(l)}\, x_d^{(l-1)}\right). \tag{1}$$

From a neural perspective, we may therefore view the $u$- and $v$-vectors (see equations (5) and (6)) as hidden layers of a multi-layer perceptron (more details in Bishop (1995) and Haykin (1999)), the structure of which is detailed in Figure 2. Finally, the target function aggregates the squared differences between the observed ranks $R_i$ and the rank statistics of the estimated latent utilities.

## 3. STATISTICAL MODELLING

### 3.1 *The firm's decision rule*

The decision maker (*d.m.*) is assumed to make his choice as follows:

(*i*) To each scenario $z_i$ he associates a utility $U^*(z_i,\epsilon_i)$ depending on the relevant and known characteristics, or attributes, ($z_i$) and on characteristics of events which are uncontrolled and unknown and that also affect the decision maker's utility ($\epsilon_i$).

(*ii*) The utilities $U^*(z_i,\epsilon_i)$ are random for the decision maker because they depend on the unobservable vector $\epsilon=(\epsilon_1,\epsilon_2,...,\epsilon_I)$. Under an expected utility assumption, the *d.m.* computes for each scenario $z_i$ the expectation of these random utilities, namely:

$$U(z_i,\theta)=\mathbb{E}[U^*\mid z_i,\theta] \tag{2}$$

where $\theta$ contains the parameters of the utility function $U^*$ and of the distribution of $(\epsilon\mid z)$ (for further details, see in Varian (1992)). Thus, the function $U$ is a cardinal utility function, *i.e.* identified up to an arbitrary linear transformation only.

(*iii*) The observed ranking $R_i$ is interpreted as an ordering, over the $I$ scenarios, of the expected utilities $U(z_i,\theta)$, $1 \le i \le I$. Therefore, the theoretical rank $r_i(Z,\theta)$ is given by:

$$r_i(Z,\theta) = 1 + \sum_{i=1}^{I} \mathbf{1}_{\{U(z_i,\theta) < U(z_i,\theta)\}} \tag{3}$$

where $\mathbf{1}_{\{.\}}$ represents the indicator function. Thus $r_i(Z,\theta) = 1$ is given to the scenario with lowest utility.

Because the rank statistic $r(Z,\theta)$ is not sufficient for the utility vector $u$, the transformation (3) leads to an identification problem (see Oulhaj and Mouchart, 2003). More specifically, for a given set $Z$ of scenarios, the ranking function $r(Z,\theta)$ defined by

$$r(Z,\theta):(Z,\theta) \to (r_1(Z,\theta),...,r_I(Z,\theta)) \tag{4}$$

is not one-to-one. This means that different values of $\theta$ may correspond to a same ranking.

### 3.2 *A parametric utility*

The following parametric specification for $U$ is considered:

$$U(z_i,\theta) = \sum_{j=1}^{J} \omega_j v_j(z_{ij},\gamma) \quad 1 \le j \le I \tag{5}$$

where $v_j$, $1 \le j \le J$, are known functions and $\theta = (\gamma,\omega)$ is the parameter of interest. The parameter $\omega = (\omega_1,\omega_2,...,\omega_J)$ lies in the $(J-1)$ dimensional simplex $S_{[J-1]} = \{s \in \Re_+^J \mid \sum_{j=1}^{J} s_j = 1\}$, and $\gamma$ are parameters of $v_j$.

We pay a particular attention to a logistic specification of the utility function, namely:

$$v_j(z_{ij},\alpha_j,\beta_j) = \frac{e^{\alpha_j+\beta_j z_{ij}}}{1+e^{\alpha_j+\beta_j z_{ij}}} = \frac{1}{1+e^{-(\alpha_j+\beta_j z_{ij})}} \tag{6}$$

Here, $\gamma = (\alpha,\beta)$ where $\alpha = (\alpha_1,\alpha_2,...,\alpha_J), \beta = (\beta_1,\beta_2,...,\beta_J)$. It should be noticed that, if $\omega$ is not constrained to lie in the simplex, the minimization of the loss function (to come later on) provides uninterpretable results, namely negative values for most $\omega_j$ and meaningless signs for the coefficients $\beta_j$. This remark leads to the following reparametrization of the weights $\omega_j$:

$$\omega_j = \frac{e^{\lambda_j}}{1+\sum_{p=1}^{J-1} e^{\lambda_p}}, \quad 1 \le j \le J-1, \tag{7}$$

$$\omega_J = \frac{1}{1+\sum_{p=1}^{J-1} e^{\lambda_p}} \tag{8}$$

where $\lambda = (\lambda_1, \lambda_2, ..., \lambda_{J-1}) \in \Re^{J-1}$. The inverse transformation is

$$\lambda_j = ln\left(\frac{\omega_j}{1-\sum_{1 \le p \le j} \omega_p}\right), \quad 1 \le j \le J-1. \tag{9}$$

This relation characterizes a bijection between the interior of $S_{[J-1]}$ and $\Re^{J-1}$. The parameter vector to be estimated, $\theta = (\gamma, \lambda)$, has accordingly dimension $3J-1$.

## 4. NEURAL ESTIMATION

The objective of this section is to build an estimator of $\theta$ minimizing a loss function $L(\theta)$ which aggregates a loss $L_i(\theta)$ associated to each scenario $i = 1, ..., I$, *viz.*

$$L(\theta) = \sum_{i=1}^{I} L_i(\theta). \tag{10}$$

Under a neural approach, the iterative algorithm generates a sequence of estimates $\hat{\theta}_q (q \ge 0)$ following the structure illustrated in Figure 2. This algorithm, repeated independently for each firm, presents the structure of a perceptron with two hidden layers and proceeds as follows:

0. Input the $I$ scenarios $Z = [z_i]$, the observed ranks $R = (R_1, ..., R_I)$ and an initial value $\hat{\theta}_0$.
1. Compute $U(z_i, \hat{\theta}_q)$ $1 \le i \le I$ (from (5) and (6)).
2. From equation (3), compute the estimated ranking $r_i(Z, \hat{\theta}_q)$.
3. Knowing $r_i(Z, \hat{\theta}_q)$ and the observed rank of $z_i$ (*i.e.* $R_i$) for each scenario, evaluate the loss associated to the ranking error of the scenario $i$, namely: $L_i(\hat{\theta}_q)$. Then compute the total loss function $L(\hat{\theta}_q)$ (from (10)).
4. Update the parameter $\hat{\theta}_q$. The update is based on the minimization of the total loss function $L(\hat{\theta})$.
5. Iterate steps 1 to 4 until convergence.

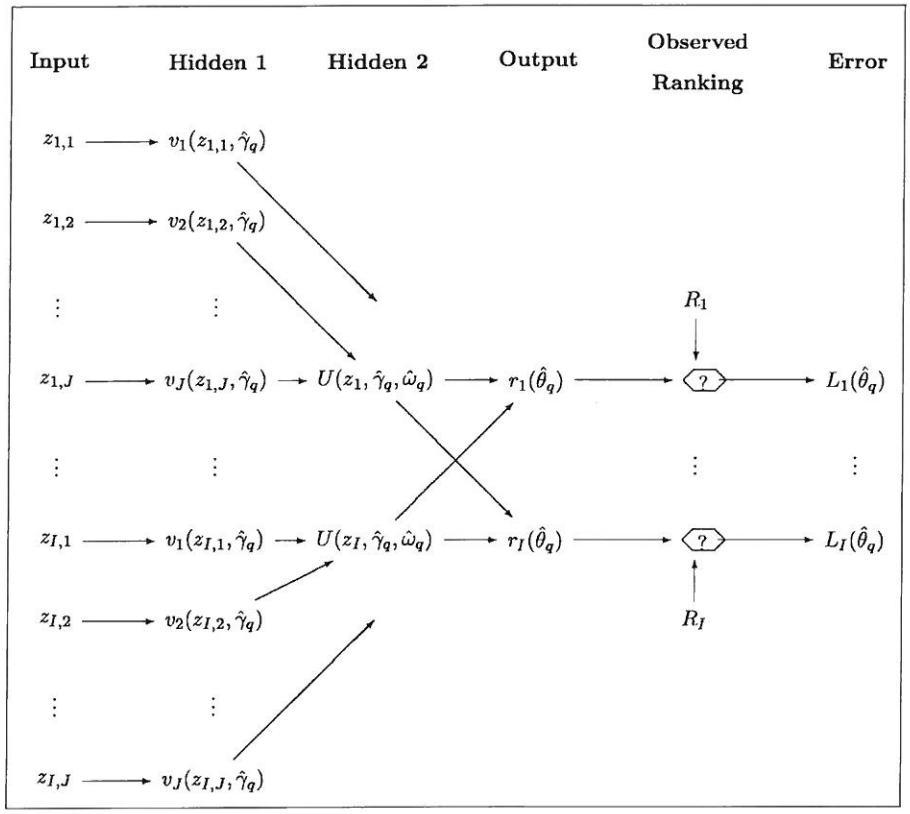


*Figure 2* – Perceptron structure (for $\hat{\theta} = \hat{\theta}_q = (\hat{\gamma}_q, \hat{\lambda}_q)$ at the $q$-th iteration).

The error criterion to be minimized takes into account the discrete character of the observed ranking, and the continuous character of the hidden (latent) utility function (5) and (6). A natural solution is to use the quadratic error between the stated ranking $R_i$ and the estimated ranking $r_i(Z,\theta)$ produced by the model namely:

$$L = L_D(\theta) = \sum_{i=1}^{I} (r_i(Z,\theta) - R_i)^2. \tag{11}$$

**Remark**: When fitting ordinal data, minimizing the quadratic error (11) might seem less attractive than maximizing the Kendall coefficient, namely

$$\tau_K(\mathrm{R},\hat{\mathrm{R}}) = \frac{S}{\frac{1}{2}I(I-1)}, \tag{12}$$

where $I$ is the number of scenarios and $S$ the observed sum of the +1 and -1 scores for all possible pairs of scenarios and where the scores are calculated as:

$$S_{ij} = 2 \cdot \mathbf{1}_{\{(R_i - R_j)(\hat{R}_i - \hat{R}_j) > 0\}} - 1$$

$$S = \sum_{1 \le i \le I-1,\, j \ge i} S_{ij}.$$

Both criteria correspond to closely related ideas and may be expected to produce similar results. In particular, in the case of perfect fit (*i.e.* $R_i = \hat{R}_i, \forall i$), $L_D(\hat{\theta}) = 0$ and $\tau_K(\hat{\theta}) = 1$. In the application, we systematically optimize $L_D$ for being numerically more stable than $\tau_K$.

## 5. APPLICATION

### 5.1 *The data*

For a given firm, we observe $I = 25$ scenarios and a corresponding ranking. Each scenario is represented by a vector of $J = 6$ attributes. By convention, $z_1$ stands for a reference scenario. We denote by $Z$ the $(25 \times 6)$ matrix containing all the 25 scenarios. The ranking of these scenarios is represented by a vector $R = (R_1, R_2, ..., R_{25})$ where $R_i$ denotes the rank of $z_i$ according to the firm's preference. The ranking $R_i$ of each scenario lies eventually between 1 and 25. Thus, for each firm we have a $(25 \times 7)$ data matrix $(Z, R)$. The rankings of 9 firms have been treated independently of each others.

### 5.2 *Pitfalls in minimization*

Equation (3) makes clear that $r_i(z, \theta)$ and therefore $L_D(\theta)$ (where $D$ stands for discrete), are not continuously differentiable in $\theta$. Its minimization cannot be carried out by classical algorithms such as gradient methods. In order to circumvent this difficulty, one might think that the rankings behave as a discrete approximation of a utility scaled to lay in $[1, 25]$. This may be achieved through the following transformation of $U(z_i, \theta)$:

$$U^s(z_i, \theta) = 1 + 24 \frac{U(z_i, \theta) - m(\theta)}{M(\theta) - m(\theta)} \quad \in [1, 25] \tag{13}$$

where $m(\theta) = min_i U(z_i, \theta)$ and $M(\theta) = max_i U(z_i, \theta)$. Thus, $m(\theta)$ (*resp.* $M(\theta)$) is the lowest (*resp.* highest) utility. In this case, the loss function can be written as follows:

$$L = L_C(\theta) = \sum_{i=1}^{25} (U^s(Z, \theta) - R_i)^2 \tag{14}$$

where $C$ stands for continuous. The loss function $L_C(\theta)$ is differentiable and can be minimized by a gradient method.

Experience has however revealed that such an approach raises substantial problems. For two selected companies, we observed the following difficulties. In Figures 3 and 4, we plot, for 2 different firms, in plain line the value of $L_C(\hat{\theta}_q)$ and in dashed line, the corresponding value of $L_D(\hat{\theta}_q)$, as functions of the number of iterations. Figure 3 shows that minimizing $L_C$ may be conflicting with minimizing $L_D$. Figure 4 reveals that, for another company, assessing whether the algorithm has achieved a reasonable neighborhood of the true minimum may be problematic because the decrease of the objective function may have an unusual behavior: will the steep decrease around the 1000-th iteration repeated later on? after the iteration 100 000? and is the value of the objective function, namely 26, far or close to the true minimum? The maximum ranking error $L_D$ among $J$ alternatives, say $M(J)$, is equal to

$$M(J) = 2\sum^{\lfloor J/2 \rfloor}(J - 2j + 1)^2 \tag{15}$$

where $\lfloor a \rfloor$ stands for the integer part of $a$. Thus, with 25 alternatives, we know that $0 \leq L_D \leq 5200$.

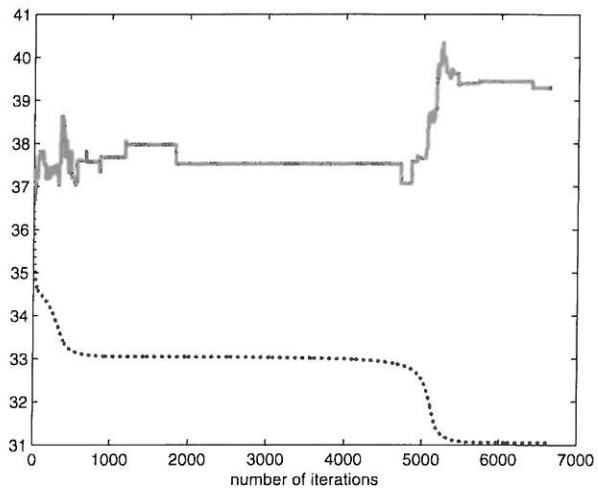


*Figure 3* – Evolution of the discrete loss function $L_D$ (plain line) and the continuous loss function $L_C$ (dashed line) in function of the number of iterations for **company 4**.

### 5.3 *A heuristic Approach*

The unsatisfactory results and the problems related to the approximation of the discrete loss function $L_D$ by the continuous loss function $L_C$ lead us to look for an alternative approach. Minimizing the discrete error $L_D$ fosters the use of non classical techniques able to deal with a discontinuous criterion.

In the present case, the heuristic allowing the minimization is based on the Pocket algorithm Gallant (1996). Similarly to a gradient method, the Pocket algorithm generates a sequence of estimates $\hat{\theta}_q$. One major difference is that the computation of $\hat{\theta}_{q+1}$, as a transformation of $\hat{\theta}_q$, is obtained through an iterative procedure with steps indexed by, say, $t$. In this application, there are 17 parameters (see equations (6) to (9)), but, as each step $t$ may require a positive or a negative variation, the Pocket algorithm considers 34 possible variations to be evaluated. Thus the 17-dimensional vector $\theta = (\theta_f)$ is replaced, in the Pocket algorithm, by a 34-dimensional vector $(\theta_{f,s})$ with $f = 1,...,17$ and $s = +1,-1$.

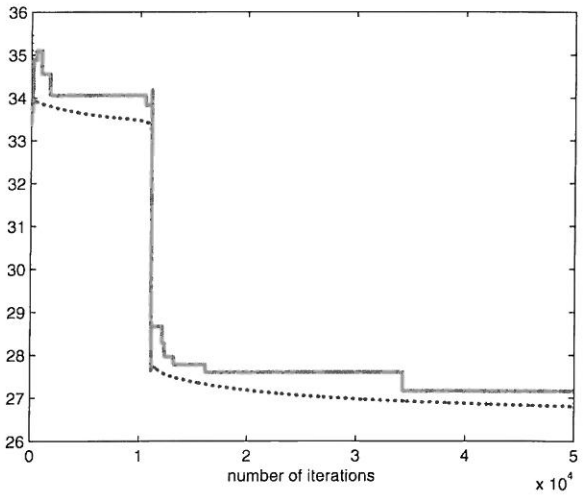


*Figure 4* – Evolution of the discrete loss function $L_D$ (plain line) and the continuous loss function $L_C$ (dashed line) in function of the number of iterations for **company 9**.

Two types of parameters characterize a Pocket algorithm, namely a fixed number $(D)$ of coordinates $(f,s)$ to be updated at each step $t$ and a length of adaptation $(\Delta_f)$, kept constant at each step $t$. For simplicity, let us consider a particular iteration $q$ and write $\hat{\theta}$ instead of $\hat{\theta}_q$, where $\hat{\theta}$ has coordinates $\hat{\theta}_{f,s}$ The sequence $\hat{\theta}_{f,s}(t)$ is generated as:

$$\hat{\theta}_{f,s}(t+1) = \hat{\theta}_{f,s}(t) + \mathbf{1}_{\{(f,s)\in \mathcal{A}_t\}}\, s\Delta_f, \tag{16}$$

where $\mathcal{A}_t$ selects, among all possible $(f,s)$, the $D$ most favorable ones, *i.e.* the updates corresponding to the steepest decrease of squared error $L_D$ eliminating those coordinates for which $s\Delta_f$ increases $L_D$. Thus the step $t$ is final once $\mathcal{A}_{t+1}$ becomes empty. Obviously $D \leq 2F$, where $F$ stands for the number of parameters to be optimized; here $F = 3J - 1 = 17$.

### 5.4 *Choice of the parameters for the iterative procedures*

For the initialization of the *perceptron* procedure, the weights $\omega_j$ are set equal to values declared in the interviews when available, this is the case for companies 2, 3, 4, 6, 7, 8 and 9. For the other companies, namely companies 1 and 5, the initial weights are set equal to those obtained in a UTA model developed in Beuthe *et al.* (2008). The other parameters, $\alpha$ and $\beta$, are initialized to 1. The number of iterations is fixed at 200; the same number of iterations is used for all simulations.

A simple run of the *Pocket algorithm* described above is quick but the reliability of the results crucially depends on the specification of $F+1$ parameters required by the working of the algorithm, namely $D$ and $\Delta_f$ with $f = 1,...,F$. It is therefore compelling to input several trial values for these parameters. In the present application, the optimization for each firm is organized as follows:

- $D$ varies between 1 and 17 by steps of 1
- the length of adaptation for the $\alpha$ and $\beta$ parameters varies from 0.1 to 1 by steps of 0.1
- the length of adaptation for $\lambda$ varies from 0.0005 to 0.002 by steps of 0.00025.

The variation for the length of adaptation of $\lambda$ is chosen lower than that of $\alpha$ and $\beta$ because of the high impact of a variation of $\lambda$ in the value of $\omega$. As a matter of fact, the problem is quite sensitive when the number of updates per step $(D)$ is high. The best model can be selected according to the loss function $L_D$ or the $\tau_K$. Because the evaluation of $\tau_K$ is not computationally convenient, we systematically minimize $L_D$ and report the results and the $\tau_K$ for the models reaching the smallest value of $L_D$ and the highest value of $\tau_K$ respectively; when the two models coincide, we write $L_D \sim \tau_K$.

### 5.5 *Results for freight transport data*

The weights associated to each attribute, as well as the corresponding Kendall coefficients are presented in Table 1 for the models related to nine firms. The criterion used to choose the Pocket parameters is also given. Tables 2 and 3 give the values taken by the $\alpha$ and $\beta$ parameters for each model.

Let us first examine Table 1. For Companies 1, 2, 4 and 5, the models with lowest $L_D$ and highest $\tau_K$ are different but the corresponding $\tau_K$ 's are close together (for instance, .9067 and .9133 for company 1, .8867 and .9000 for company 2) and the estimates of the weights are also similar. For the other companies, namely companies 3, 6, 7, 8 and 9, the two models are the same. The fact that all methods, estimated in Beuthe *et al.* (2006), lead to models whose $\tau_K$ is one for company 9 suggests that the behavior of this firm is quite simple and that all models overfit. It is likely that this phenomenon of overfitting occurs in several cases. Indeed, it is questionable that a model could approximate the behavior of a company in terms of choice of transportation mode in such way that the Kendall coefficient would reach 0.9000. The fact that the NN method counts less parameters than UTA and leads to lower $\tau_K$ is reassuring from this point of view.

TABLE 1

*Weights for each model and Kendall coefficient*

| *Company* | *Criterion* | *Frequency* | *Time* | *Reliability* | *Flexibility* | *Loss* | *Cost* | $L_D, \tau_K$ |
|---|---|---|---|---|---|---|---|---|
| 1 | $L_D$ | 0.0076 | 0.0353 | 0.1074 | 0.0426 | 0.0625 | 0.7446 | **0.9067** |
| | $\tau_K$ | 0.0077 | 0.0361 | 0.1128 | 0.0454 | 0.0666 | 0.7314 | **0.9133** |
| 2 | $L_D$ | 0.1426 | 0.1426 | 0.1407 | 0.1397 | 0.1388 | 0.2966 | **0.8867** |
| | $\tau_K$ | 0.1406 | 0.1372 | 0.1341 | 0.1316 | 0.1296 | 0.3266 | **0.9000** |
| 3 | $L_D \sim \tau_K$ | 0.0997 | 0.1471 | 0.3397 | 0.1425 | 0.0000 | 0.2707 | **0.7200** |
| 4 | $L_D$ | 0.4319 | 0.3678 | 0.0486 | 0.1129 | 0.0092 | 0.0296 | **0.8200** |
| | $\tau_K$ | 0.4388 | 0.3643 | 0.0478 | 0.1088 | 0.0086 | 0.0316 | **0.8400** |
| 5 | $L_D$ | 0.0028 | 0.0000 | 0.0007 | 0.0049 | 0.0007 | 0.9908 | **0.8200** |
| | $\tau_K$ | 0.0030 | 0.0000 | 0.0008 | 0.0054 | 0.0008 | 0.9900 | **0.8600** |
| 6 | $L_D \sim \tau_K$ | 0.0001 | 0.0201 | 0.0001 | 0.1650 | 0.0001 | 0.8146 | **0.7000** |
| 7 | $L_D \sim \tau_K$ | 0.0493 | 0.0489 | 0.1953 | 0.1940 | 0.0476 | 0.4649 | **0.9400** |
| 8 | $L_D \sim \tau_K$ | 0.3321 | 0.0815 | 0.5058 | 0.0262 | 0.0520 | 0.0023 | **0.7600** |
| 9 | $L_D \sim \tau_K$ | 0.0000 | 0.0000 | 0.4001 | 0.0000 | 0.0000 | 0.5998 | **1.0000** |

Table 2 shows that the estimation of the $\alpha$ 's is reasonably robust with respect to the choice between the two criteria $L_D$ or $\tau_k$: they keep the same sign and the same order of magnitude, with however one noticeable exception for company 2 where the estimations differ substantially, in sign and in order of magnitude, for Reliability and Flexibility. This may be taken as a signal numerical sensitivity due to the discreteness of the rankings.

Equation (5) shows that the weights $\omega_j$ provide some insight about the relative importance of each feature of the freight transport. According to the weights, Cost is the main or the second main attribute for 7 out of the 9 companies. The

corresponding values of the $\beta$ parameters support this observation. On the contrary, the Loss attribute seems unimportant with respect to the values of the weights, which are less or equal to 7%, except for company 2. Reliability has also an impact for some companies. Company 4 presents an atypical behavior: its main attribute is Frequency, followed by Time.

TABLE 2

*Values taken by $\alpha$ parameters for each company*

| *Company* | *Criterion* | *Frequency* | *Time* | *Reliability* | *Flexibility* | *Loss* | *Cost* |
|---|---|---|---|---|---|---|---|
| 1 | $L_D$ | 0.0000 | -1.2000 | 0.0000 | 1.2000 | 1.8000 | 4.4000 |
| | $\tau_K$ | 0.2000 | -2.0000 | 0.2000 | 0.2000 | 1.2000 | 3.0000 |
| 2 | $L_D$ | 3.0000 | 2.0000 | -4.0000 | -8.5000 | 3.5000 | 2.0000 |
| | $\tau_K$ | 5.0000 | 4.2000 | 1.8000 | 2.6000 | 8.2000 | 2.6000 |
| 3 | $L_D \sim \tau_K$ | 3.4000 | 1.0000 | 2.6000 | 8.2000 | 1.0000 | -1.4000 |
| 4 | $L_D$ | 2.6000 | 3.0000 | 2.2000 | 1.8000 | 2.2000 | -0.6000 |
| | $\tau_K$ | 3.000 | 2.6000 | 2.2000 | 4.2000 | 0.2000 | -0.6000 |
| 5 | $L_D$ | 4.2000 | 0.2000 | 2.6000 | 3.4000 | 7.4000 | 2.6000 |
| | $\tau_K$ | 3.4000 | 1.6000 | 1.000 | 2.8000 | 13.000 | 2.2000 |
| 6 | $L_D \sim \tau_K$ | 1.0000 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | 2.000 |
| 7 | $L_D \sim \tau_K$ | 5.0000 | 4.2000 | 1.8000 | 2.6000 | 8.2000 | 2.6000 |
| 8 | $L_D \sim \tau_K$ | 8.0000 | -5.0000 | 7.0000 | -10.0000 | 6.0000 | 5.0000 |
| 9 | $L_D \sim \tau_K$ | 1.0000 | 0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

The sign of the $\beta$ parameters expresses the favorable (when positive) or unfavorable impact (when negative) of the related attribute. Attributes whose weights in the model are close to zero are not significant and should not be taken into account when interpreting the corresponding values of $\alpha$ and $\beta$.

The comparison of Tables 1 and 3 shows therefore that signs of $\beta$ are intuitive for Frequency, Time, Loss and Cost, in the case of significant attributes. For instance, the signs of $\beta$ parameters related the Cost (main factor for most firms) are negative in all models; this means that an increase of this attribute leads to a lower utility. The negative impact of Time is also correctly expressed by 8 of the 13 models, while the 5 models left present a small weight (less or equal to 0.0361) for this transport feature. The signs of $\beta$ for these 5 models are consequently not significant. In the case of Reliability, all signs are intuitively correct, except for one model. A $\beta$ parameter not significantly different of zero means that the corresponding attribute has no impact in the model (even if the weight is nonzero). In the case of Flexibility, three significant $\beta$ parameters are non positive, but one of these is not significant. This therefore suggests that Flexibility plays no significant role in this model.

We may notice in general that the results for both criteria of selection (loss function and Kendall) lead to the same model or to models rather similar in terms of performances and relative importance of the attributes. The other parameters seem however less stable.

TABLE 3

*Values taken by $\beta$ parameters for each company*

| *Company* | *Criterion* | *Frequency* | *Time* | *Reliability* | *Flexibility* | *Loss* | *Cost* |
|---|---|---|---|---|---|---|---|
| 1 | $L_D$ | 3.4000 | 0.4000 | 1.4000 | 0.6000 | -1.4000 | -8.8000 |
| | $\tau_K$ | 3.8000 | 1.2000 | 3.4000 | 0.6000 | -2.6000 | -6.8000 |
| 2 | $L_D$ | 0.5000 | -1.5000 | -4.5000 | -22.5000 | -32.0000 | -17.0000 |
| | $\tau_K$ | 4.2000 | -0.6000 | 9.8000 | 1.8000 | -35.8000 | -38.2000 |
| 3 | $L_D \sim \tau_K$ | 12.2000 | -31.0000 | 5.0000 | -23.0000 | 49.4000 | -13.4000 |
| 4 | $L_D$ | 1.4000 | -3.4000 | 7.0000 | 1.0000 | -1.8000 | -13.4000 |
| | $\tau_K$ | 2.2000 | -3.8000 | 8.2000 | -1.0000 | 0.2000 | -20.2000 |
| 5 | $L_D$ | 2.6000 | -0.6000 | 14.6000 | 3.4000 | -31.000 | -95.000 |
| | $\tau_K$ | 0.4000 | 0.4000 | 10.000 | 1.6000 | -53.6000 | -64.4000 |
| 6 | $L_D \sim \tau_K$ | 0.0000 | 0.0000 | 2.0000 | 0.0000 | 0.0000 | -2.0000 |
| 7 | $L_D \sim \tau_K$ | 4.2000 | -0.6000 | 9.8000 | 1.8000 | -35.8000 | -38.2000 |
| 8 | $L_D \sim \tau_K$ | 10.0000 | -2.0000 | 10.0000 | -10.0000 | -14.0000 | -39.0000 |
| 9 | $L_D \sim \tau_K$ | 1.0000 | 1.0000 | 0.4000 | 1.0000 | 1.0000 | -1.3000 |

## 6. CONCLUSION

This paper shows that a perceptron model can lead to a good prediction of the ranking. In addition, the parameters of the model express correctly the negative or positive impact of an increase in an attribute level, in most cases. The weights of the model give an insight about the relative importance of each freight transport attribute. In particular, it is shown that Cost and Reliability are often the most important features.

A continuous approximation of the quadratic error between the ranking and its estimate is not always suitable. Indeed, the minimization of the quadratic error and the continuous approximation may be conflicting. Also, the behavior of the continuous approximation make the minimization by a gradient method difficult.

The performances in terms of Kendall coefficients is lower than the UTA model ($\tau_K$ is always higher than 0.9, against 0.7 for the perceptron). However, the UTA model counts 23 parameters for 25 alternatives: the UTA model most probably overfits. Moreover, achieving this $\tau_K$ in such complex problem seems unrealistic. The Neural Network model outperforms the non-metric conjoint analysis and rank-ordered logit models (in simple and nested versions). The Kendall coefficients are however slightly lower than Quasi-UTA, a simplified ver-

sion of UTA with less parameters. The description and comparison of these models can be found in Beuthe *et al.* (2008). However, those methods do not provide results easily interpretable, contrarily to the perceptron model.

*OS Engineer, KPN Group Belgium* CATHERINE KRIER

*CORE and ISBA, Université catholique de Louvain* MICHEL MOUCHART

*DTU, Nuffield Department of Clinical Medecine* ABDERRAHIM OULHAJ
*University of Oxford*

## REFERENCES

B. APOLLONI, I. ZOPPIS, R. ALEANDRI, A. GALLI, (2001), *A neural network based procedure for forecasting bovine semen motility*, "Statistica", LXI (2), 301-313.

M. BEUTHE, CH. BOUFFIOUX, J. DE MAEYER, G. SANTAMARIA, M. VANDRESSE, E. VANDAELE, F. WITLOX, (2005), *A multi-criteria methodology for stated preferences among freight transport alternatives*, in AURA REGGIANI and LAURIE SCHINTLER (ed.) *Methods and Models in Transport and Communications: Cross Atlantic Perspectives*, Berlin: Springer Verlag.

M. BEUTHE M., CH. BOUFFIOUX, C. KRIER, M. MOUCHART, (2008), *A comparison of conjoint, multicriteria, conditional logit and neural network analyses for rank-ordered preference data*, chap. 9 in M. BEN-AKIVA, H. MEERSMAN and E. VAN DE VOORDE (ed.) *Recent Developments in Freight Transport Modelling: Lessons from the Freight Sector*, Emerald Group Publishing Limited, 157-178.

M. BEUTHE, G. SCANNELLA, (2001), *Comparative analysis of UTA multi-criteria methods*, 'European Journal of operational research', 130 (2): 246-262.

E. BIGANZOLI, P. BORACCHI, I. POLI, (2000), *Reti neurali artificiali per lo studio di fenomeni complessi: limiti e vantaggi delle applicazioni in biostatistica*, 'Statistica', LX (4), 723-734

C.M. BISHOP, (1995), *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

S.I. GALLANT, (1996), *Perceptron based learning algorithms*, 'IEEE Transactions on Neural Networks', 1:179-191.

S. HAYKIN, (1999), *Neural Networks: a comprehensive foundation*, 2nd ed., Prentice-Hall, Inc.

A. OULHAJ, M. MOUCHART, (2003), *Partial sufficiency with connection to the identification problem*, 'Metron', LXI (2), 267-683.

M. PILLATI, (2001), *Le reti neurali in problemi di classificazione: una soluzione basata sui metodi di segmentazione binaria*, 'Statistica', LXI (3), 407-421

H.R. VARIAN, (1992), *Microeconomic Analysis*, 3rd ed., Norton & Company, New-York.

## SUMMARY

*Neural modelling of ranking data with an application to stated preference data*

Although neural networks are commonly encountered to solve classification problems, ranking data present specificities which require adapting the model. Based on a latent utility function defined on the characteristics of the objects to be ranked, the approach sug-

gested in this paper leads to a perceptron-based algorithm for a highly non linear model. Data on stated preferences obtained through a survey by face-to-face interviews, in the field of freight transport, are used to illustrate the method. Numerical difficulties are pinpointed and a Pocket type algorithm is shown to provide an efficient heuristic to minimize the discrete error criterion. A substantial merit of this approach is to provide a workable estimation of contextually interpretable parameters along with a statistical evaluation of the goodness of fit.