# A note on the use of Bayesian Hierarchical Models for supervised classification *

Antunes, M.[†], Andreozzi, V. L. and Amaral Turkman, M. A.[‡]

**Abstract**: The task of supervised classification is to classify new cases into predefined classes. It seeks a rule for accurately predicting a categorical variable (class identity) based on measured variables.

This paper is motivated by the need of building a classification rule to assign individuals to one of several categories of a variable $T$ given the value of a variable $Y$, assumed to be correlated to variable $T$. In this paper, we propose a classification rule built using an hierarchical Bayesian model. The aim is to classify an individual in one of $k$ ($k = 3$) categories, based on an observation of a variable $Y$. Variable $Y$ is supposed to be easily observable and also to behave in a sufficiently different manner in each of the categories, so that an efficient classification rule can be built.

The classification rule is applied to data on Rheumatoid Arthritis, in order to predict the stage of the disease based on the X-ray exam result, namely on the SvdH score.

**Key words**: Bayesian Hierarchical Models; Supervised classification.

*Research supported by FCT/POCI 2010

[†]Faculty of Sciences, University of Lisbon, Lisbon, Portugal. e-mail:marilia.antunes@fc.ul.pt

[‡]Faculty of Sciences, University of Lisbon, Lisbon, Portugal. e-mail: antonia.turkman@fc.ul.pt

# 1 Introduction

In many practical situations, the individuals in a population belong to one of several categories. For example, patients suffering from some type of cancer can be separated into categories referring either to cancer subtypes or stages of the disease. Often, these categories are not easy to identify and hence individuals are not easily assigned to their group correctly. In such situations it is desirable to have an efficient and easy to use rule allowing to classify the individuals into the categories they belong to. This can be done using a variable $Y$ that can be measured in every individual and that behaves in a different manner according to the group the individual belongs to. The idea is to determine intervals such that the probability of an individual to belong to a certain category is maximum when $Y$ belongs to that interval. Like this, given the value $Y = y$, the individual is classified in class C if $P(\text{belong to class C}|Y = y)$ is maximum among the alike probabilities calculated for all the classes.

The paper is organized as follows: in section 2, we present the model and in section 3 we apply it to RA data.

# 2 The Model

We assume that

$$\mathcal{D} = \{(y_1, t_1), \ldots, (y_n, t_n)\} \tag{1}$$

is the available data, where $(y_i, t_i)$, corresponds the data for the $i-th$ individual and $n$ is the number of individuals in the study. Variable $Y$ represents a continuous characteristic. It is assumed that $Y$ is a characteristic that is present in every member of the population. Also, it is assumed that it is easy to measure and also that it behaves in a different manner, depending on the group of the population the individual belongs to. Hence, it is useful for the purpose of classifying a future individual given the observed measurement $y$ of $Y$.

Variable $T$ is a categorical variable, assuming values (labels) from 1 to $k$, where $k$ represents the number of groups the population is divided in. For example, it may represent presence/absence of a disease ($k$=2) or different types or stages of a disease ($k \geq 2$).

We assume also that both variables are known for a certain number $n$ of individuals, this is, $\mathcal{D}$ is a complete data set.

For practical convenience related to the application, and without any loss of generality we will consider $k = 3$, as the results are valid and easily extended for $k \neq 3$.

We build a Bayesian hierarchical model where the following distributions are considered:

- $T \sim \text{dcat}(\pi_1, \pi_2, \pi_3)$, that is, $T$ follows a categorical discrete distribution
$$T : \begin{cases} j, & j = 1, 2, 3 \\ \pi_j \end{cases},$$
where $\pi_j = P[T = j]$ and $\sum_{j=1}^{3} \pi_j = 1$.

- Conditional on $T = j$, $Y$ follows a Gamma distribution with shape parameter $\alpha$ and scale parameter $\beta_j$, $Y_{|T=j} \sim Gamma(\alpha, \beta_j)$, $j = 1, 2, 3$, that is,
$$p(y|T = j, \alpha, \beta_j) = \frac{\beta_j^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta_j y}, \quad y > 0, \alpha > 0, \beta_j > 0$$

Let $\theta = (\pi_1, \pi_2, \alpha, \beta_1, \beta_2, \beta_3)$, be the vector containing all the model parameters. The prior distributions for the parameters are the following:

- $(\pi_1, \pi_2) \sim Dirichlet(a_1, a_2, a_3)$
- $\beta_j \sim Gamma(g, h)$, $j = 1, 2, 3$, all independent and independent from $\pi_1$ and $\pi_2$.

and $\alpha > 0$ is considered known, that is,

$$p(\pi_1, \pi_2) = \frac{\Gamma(a_1 + a_2 + a_3)}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)} \pi_1^{a_1-1} \pi_2^{a_2-1} \pi_3^{a_3-1}, \tag{2}$$

where $a_j > 0$, $\pi_3 = 1 - \pi_1 - \pi_2$, $\pi_j \geq 0, \quad j = 1, 2, 3$ and $\sum_{j=1}^{3} \pi_j = 1$, and the prior probability density function for the $\beta_j$ is given by

$$p(\beta_j) = \frac{h^g}{\Gamma(g)} \beta_j^{g-1} e^{-h\beta_j}, \quad j = 1, 2, 3. \tag{3}$$

The constants $a_1$, $a_2$, $a_3$, $g$ and $h$ are the hyperparameters of the model.

To write the likelihood, it is convenient to introduce the functions

$$I_j(t_i) = \begin{cases} 1, & t_i = j; \\ 0, & \text{otherwise}, \end{cases} \qquad \text{defined for } j = 1, 2, 3.$$

That is, for $i = 1, \ldots, n$,

$$I_j(t_i) = \begin{cases} 1, & \text{individual } i \text{ belongs to group } j; \\ 0, & \text{otherwise} \end{cases}$$

and hence, the number of individuals in each group $j$, is given by

$$n_j = \sum_{i=1}^{n} I_j(t_i), \quad j = 1, 2, 3.$$

For simplicity of notation, let $y_{jk}$ be the $k-th$ observation in the $j-th$ group, and let $t_{jk}$ be the correspondent value of $T$. Note that $t_{jk} = j$ for all $j = 1, 2, 3$ and $k = 1, \ldots, n_j$.

Since $p(y, t|\theta) = p(y|t, \theta)p(t|\theta)$, the likelihood is given by

$$\begin{aligned}
L(\theta|\mathcal{D}) &= p(\mathcal{D}|\theta) \\
&= \prod_{j=1}^{3}\prod_{k=1}^{n_j} p(y_{jk}|T = t_{jk}, \theta) \times P(T = t_{jk}|\theta) \\
&= \prod_{j=1}^{3}\prod_{k=1}^{n_j} \frac{\beta_j^{\alpha} y_{jk}^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta_k y_{jk}} \times P(T = t_{jk}|\theta) \\
&= \prod_{j=1}^{3}\prod_{k=1}^{n_j} \frac{\beta_j^{\alpha} y_{jk}^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta_k y_{jk}} \pi_1^{I_1(t_{jk})} \pi_2^{I_2(t_{jk})} \pi_3^{I_3(t_{jk})} \\
&= \frac{(\prod_{j=1}^{3}\prod_{k=1}^{n_j}\beta_j)^{\alpha}(\prod_{j=1}^{3}\prod_{k=1}^{n_j} y_{jk})^{\alpha-1}}{(\Gamma(\alpha))^n} \times \\
&\times \; e^{-\sum_{j=1}^{3}\sum_{k=1}^{n_j}\beta_j y_{jk}} \pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3} \\
&= \beta_1^{n_1\alpha}\beta_2^{n_2\alpha}\beta_3^{n_3\alpha}\left(\prod_{k=1}^{n_1} y_{1k}\right)^{\alpha-1}\left(\prod_{k=1}^{n_2} y_{2k}\right)^{\alpha-1}\left(\prod_{k=1}^{n_3} y_{3k}\right)^{\alpha-1} \times \\
&\times \; e^{-\beta_1\sum_{k=1}^{n_1} y_{1k}} e^{-\beta_2\sum_{k=1}^{n_2} y_{2k}} e^{-\beta_3\sum_{k=1}^{n_3} y_{3k}} \Gamma(\alpha)^{-n} \qquad (4)
\end{aligned}$$

The posterior distribution is given by

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_{\Theta} p(\mathcal{D}|\theta)p(\theta)d\theta}. \qquad (5)$$

For simplicity , we consider only the numerator of (5) and hence, obtain

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta).$$

Since the parameters are considered independent, $p(\theta)$, the prior distribution of the parameters, factorizes as $p(\theta) = \left(\prod_{j=1}^{3} p(\beta_j)\right) p(\pi_1, \pi_2)$ and hence,

$$p(\theta|y, t) \quad \propto \quad p(y, t|\theta)\left(\prod_{j=1}^{3} p(\beta_j)\right) p(\pi_1, \pi_2)$$

$$= \frac{\Gamma(a_1+a_2+a_3)}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)\Gamma(\alpha)^n} \tag{6}$$

$$\times \left(\prod_{k=1}^{n_1} y_{1k}\right)^{\alpha-1}\left(\prod_{k=1}^{n_2} y_{2k}\right)^{\alpha-1}\left(\prod_{k=1}^{n_3} y_{3k}\right)^{\alpha-1} \tag{7}$$

$$\times \pi_1^{n_1+a_1-1}\pi_2^{n_2+a_2-1}\pi_3^{n_3+a_3-1}\beta_1^{n_1\alpha+g-1}\beta_2^{n_2\alpha+g-1}\beta_3^{n_3\alpha+g-1} \tag{8}$$

$$\times e^{-\beta_1(\sum_{k=1}^{n_1} y_{1k}+h)-\beta_2(\sum_{k=1}^{n_2} y_{2k}+h)-\beta_3(\sum_{k=1}^{n_3} y_{1k}+h)} \tag{9}$$

The terms in (6) and (7) are constants whereas (8) and (9) correspond to the kernels of $Dirichlet(n_1+a_1, n_2+a_2, n_3+a_3)$, $Gamma(n_1\alpha+g, \sum_{k=1}^{n_1} y_{1k}+h)$, $Gamma(n_2\alpha+g, \sum_{k=1}^{n_2} y_{2k}+h)$ and $Gamma(n_3\alpha+g, \sum_{k=1}^{n_3} y_{3k}+h)$ distributions respectively.

Calculating $\int_\Theta p(y,t|\theta)p(\theta)d\theta$, we conclude that the posterior distribution of $\theta$ can be written as a product of a Dirichlet and three Gamma probability density functions,

$$p(\theta|\mathcal{D}) = p(\pi_1,\pi_2|\mathcal{D})p(\beta_1|\mathcal{D})p(\beta_2|\mathcal{D})p(\beta_3|\mathcal{D}) \tag{10}$$

that is,

$$(\pi_1,\pi_2|\mathcal{D}) \sim Dirichelet(n_1+a_1, n_2+a_2, n_3+a_3)$$

and

$$\beta_j|\mathcal{D} \sim Gamma(G_j, H_j),$$

where $G_j = n_j\alpha + g$ and $H_j = \sum_{k=1}^{n_j} y_{jk} + h$, for $j = 1,2,3$. This means that, a posteriori, the $\beta_j$ are also all independent and independent of $\pi_i$ for all $i = 1,2,3$ and $j = 1,2,3$.

The conditional predictive distribution of a future $Y$ given $T = j$ is given by

$$\begin{aligned} p(y|\mathcal{D}, T=j) &= \int_\Theta p(y|T=t_j,\theta)p(\theta|\mathcal{D})d\theta \\ &= \int_0^\infty p(y|T=t_j,\beta_j)p(\beta_j|\mathcal{D})d\beta_j \\ &= \int_0^\infty \frac{\beta_j^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta_j y} \times \frac{H_j^{G_j}}{\Gamma(G_j)}\beta_j^{G_j-1}e^{-(y+H_j)\beta_j}d\beta_j \\ &= \frac{\Gamma(\alpha+G_j)}{\Gamma(\alpha)\Gamma(G_j)} \times \frac{H_j^{G_j}y^{\alpha-1}}{(y+H_j)^{G_j+\alpha}}, \quad 0<y<\infty, \end{aligned} \tag{11}$$

which is a Gamma-gamma distribution. The marginal predictive distribution of $Y$ is obtained calculating

$$p(y|\mathcal{D}) = \sum_{j=1}^{3} p(y|\mathcal{D}, T=j)p(T=j|\mathcal{D}) \tag{12}$$

where $p(y|\mathcal{D}, T = j)$ is the conditional predictive distribution of $Y$ given $T = j$, and $p(T = j|\mathcal{D})$ is the predictive distribution of $T$. These are respectively given by

$$p(y|\mathcal{D}, T = j) = \frac{\Gamma(\alpha + G_j)}{\Gamma(\alpha)\Gamma(G_J)} \times \frac{H_j^{G_j} y^{\alpha-1}}{(y + H_j)^{G_j+\alpha}} \tag{13}$$

and

$$\begin{aligned} P(T = j|\mathcal{D}) &= \int_\Theta P(T = j|\theta)p(\theta|\mathcal{D})d\theta \\ &= \int_\mathcal{P} \pi_j \times \frac{\Gamma(n_1 + n_2 + n_3 + a_1 + a_2 + a_3)}{\Gamma(n_1 + a_1)\Gamma(n_2 + a_2)\Gamma(n_3 + a_3)} \\ &\times \pi_1^{n_1+a_1-1}\pi_2^{n_2+a_2-1}\pi_3^{n_3+a_3-1} \\ &= \frac{n_j + a_j}{n + a_1 + a_2 + a_3} \end{aligned} \tag{14}$$

where $\mathcal{P} = [0, 1] \times [0, 1] \times [0, 1]$.

Now, the marginal predictive distribution of $Y$ can be written:

$$\begin{aligned} p(y|\mathcal{D}) &= \sum_{j=1}^{3} p(y|\mathcal{D}, T = j)P(T = j|\mathcal{D}) \\ &= \sum_{j=1}^{3} \frac{\Gamma(\alpha + G_j)}{\Gamma(\alpha)\Gamma(G_J)} \times \frac{H_j^{G_j} y^{\alpha-1}}{(y + H_j)^{G_j+\alpha}} \times \frac{n_j + a_j}{n + a_1 + a_2 + a_3}. \end{aligned} \tag{15}$$

The conditional predictive distribution of $T$ given $y$ can also be written:

$$\begin{aligned} P(T = j|\mathcal{D}, y) &= \frac{p(y|\mathcal{D}, T = j)P(T = j|\mathcal{D})}{p(y|\mathcal{D})} \\ &= \frac{\frac{\Gamma(\alpha+G_j)}{\Gamma(\alpha)\Gamma(G_J)} \times \frac{H_j^{G_j} y^{\alpha-1}}{(y+H_j)^{G_j+\alpha}} \times \frac{n_j+a_j}{n+a_1+a_2+a_3}}{\sum_{j=1}^{3} \frac{\Gamma(\alpha+G_j)}{\Gamma(\alpha)\Gamma(G_J)} \times \frac{H_j^{G_j} y^{\alpha-1}}{(y+H_j)^{G_j+\alpha}} \times \frac{n_j+a_j}{n+a_1+a_2+a_3}} \end{aligned} \tag{16}$$

The predictive distribution of $Y$ given the value of $T$, $p(y|\mathcal{D}, T = j)$, can be seen as a way to estimate the distribution of $Y$ within each group $j$. Plotting these three curves together allow us to compare the the groups as they are useful for descriptive purposes. Namely, they give a clear picture of the way the values of $Y$ distribute within and along the groups. The points where these curves intersect define the limits of the regions where elements from different groups coexist.

Conversely, the conditional predictive distribution of $T$ given $Y = y$ calculated for each group $j$, $p(T = j|\mathcal{D}, y)$ , is a function of $y$ and it shows how

the probability of an individual to belong to group $j$ evolves as a function of the individual's value for the variable $Y$. Hence, it can be used to determine to which group an individual is more likely to belong to given $y$, the observed value of $Y$. Also, drawing these curves should permit to identify intervals over which each of the groups is the more likely to contain the individual, based on the value of $Y$. The intersection points can be taken as cutoff points, determining such regions in terms of values of $Y$.

### 2.1 The classification rule

It is expected that the data permit to find values $a$ and $b$, $0 < a < b$, such that:

- for $y \in [0, a]$

$$P[T = 1|\mathcal{D}, y] > P[T = 2|\mathcal{D}, y] > P[T = 3|\mathcal{D}, y]$$

- for $y \in (a, b)$

$$P[T = 2|\mathcal{D}, y] > P[T = 1|\mathcal{D}, y] \text{ and } P[T = 2|\mathcal{D}, y] > P[T = 3|\mathcal{D}, y]$$

- for $y \in (b, +\infty)$

$$P[T = 3|\mathcal{D}, y] > P[T = 2|\mathcal{D}, y] > P[T = 1|\mathcal{D}, y]$$

And hence, given an individual with an observed value $y$ for the variable $Y$, the classification rule would be

$$\begin{array}{ll} \text{classify in group 1,} & \text{if } y \in [0, a]; \\ \text{classify in group 2,} & \text{if } y \in (a, b); \\ \text{classify in group 3,} & \text{if } y \in [b, +\infty). \end{array} \tag{17}$$

In the case the groups represent stages of a disease, the classification rule allows also to say something about the way the disease is evolving in the patient and hence to say something about its prognosis. If a patient is classified in his or her own group, it means that the disease is taking its expected course. If the patient is classified in a group corresponding to a less severe stage, then we conclude that the disease is taking a milder course than expected and hence the patient has a good prognosis. Conversely, if the patient is classified in a group corresponding to a more severe stage, then we conclude that the disease is taking a course worse than expected and hence the patient has a bad prognosis.

## 3 Application to Rheumatoid Arthritis data

We consider data referring to Rheumatoid Arthritis (RA) patients collected in Portugal. The 554 patients included in the study fulfilled the American College of Rheumatology 1987 revised criteria [2]. Patients were randomly selected from different hospitals covering the whole country and the insular regions of Madeira and Azores.

A useful tool for the doctors would be to have a quick and easy procedure to evaluate the patient state concerning the evolution of the disease. Because of the disease aspects, the X-Ray exam, besides being cheap and non invasive, gives a quite clear picture of the state of the joints and the cartilages, providing an accurate measure of the disease state. An interesting thing to do then is to be able to say something about the severity of the disease in terms of its course. This is, doctors would like to be able to say if a certain patient has a good prognosis, a bad prognosis or if the disease is taking its usual course.

To achieve this, we considered the SvdH score to be $Y$, the variable to be measured, and the duration of the disease (DD) to be the categorical variable T. The duration of the disease is expressed in years and the patients were separated into three groups:

- Group 1: DD $\leq$ 2 years
- Group 2: 2<DD$\leq$ 10 years
- Group 3: DD>10 years

The first group corresponds to the patients in the initial stage of the disease, whereas the other two groups correspond to moderate duration and long term illness. These groups correspond to $T = 1$, $T = 2$ and $T = 3$ respectively.

Out of the 554 patients, 359 had information both on the SvdH score and DD and distributed the way described in Table 1. Since we considered

| $j$ | DD ($T = j$) | number of patients ($n_j$) |
|---|---|---|
| 1 | less than 2 years | 47 |
| 2 | from 2 to 10 years | 132 |
| 3 | more than 10 years | 180 |

Table 1: Patients distribution along the DD groups.

$Y|T = j \sim Gamma(\alpha, \beta_j)$, with the same $\alpha$ for all the groups, we started by estimating $\alpha$ within each of the groups using the moments estimator,

$\alpha^* = \frac{m_1^2}{m_2 - m_1^2}$, where $m_1$ and $m_2$ are the first and the second empirical moments respectively. The results are in Table 2. The value found for $\alpha$ was very similar for all the three groups and hence we considered $\alpha = 3$. For the purpose of calculating the posterior distibution of the parameters, the hyperparameters were all taken to be equal to zero. The parameters of the posterior distribution of the $\beta_j$, $j = 1, 2, 3$ , can be found in the same table. The posterior distribution of $(\pi_1, \pi_2)$ is $Dirichlet(47, 132, 180)$.

| Group ($j$) | estimated $\alpha$ | $G_j$ | $H_j$ |
|---|---|---|---|
| 1 | 3.191847 | 141 | 3458 |
| 2 | 3.080123 | 396 | 11749 |
| 3 | 3.000716 | 540 | 27000 |

Table 2: Estimated values of $\alpha$ and parameters of the posterior distributions.

The conditional predictive distributions $p(y|\mathcal{D}, T = j)$ and $p(T = j|\mathcal{D}, y)$ depend on these parameters and once these are calculated, the functions can be plotted.

The conditional predictive distributions of the SvdH Score ($Y$) given the DD group ($T$) are plotted in Figure 1. We recall that these curves are useful essentially for descriptive purposes.

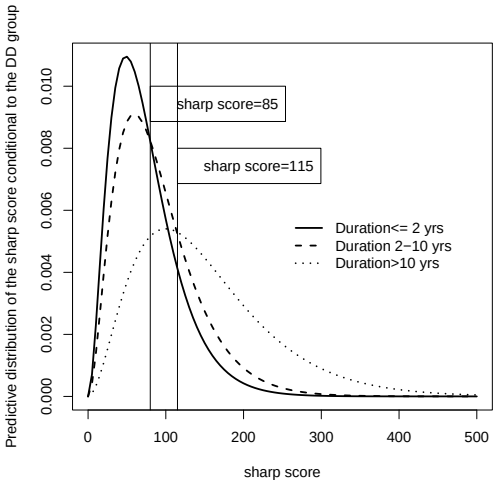


Figure 1: Estimated distribution of the Sharp/van der Heijde score (SvdH) according to the duration of the disease (DD)

The points where the curves intersect can not be found analitically but

a numerical solution is easy to find. The intersection points were approximatelly 85 and 115. This means that SvdH score values bellow 85 are more likely to occur in the first group, this is, among individuals with less than 2 years of disease duration. Similarly, SvdH score values between 85 and 115 are more likely to occur among individuals with from 2 to 10 years of disease duration and SvdH score values above 115 are more likely to occur among individuals with more than 10 years of disease duration.

Conversely, the conditional predictive probability of an individual to belong to group $j$ given a value of the SvdH score ($Y$), is a function of $y$ and can be used for classification purposes. These calculations take into account the way the patients distribute along the groups, namely the weight of each group in the sample. The functions are plotted in Figure 2.

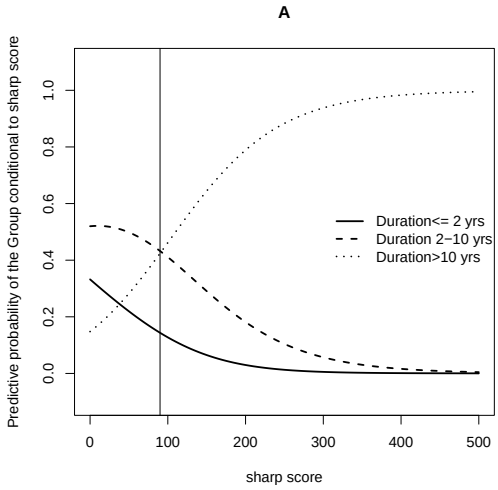


Figure 2: Estimated distribution of the Sharp/van der Heijde score (SvdH) according to the duration of the disease (DD)

From this figure we can conclude that patients with DD less than 2 years and patients with DD between 2 and 10 years can not be separated by any cutoff point. The method classifies them all as belonging to the second group. This is due to the relatively higher weight of the second group and to the similar way the SvdH score distributes in the two groups (check the similarity of the corresponding curves in Figure 1). Hence, for this reason and since there are almost three times more patients in the second group than in the first, it is always more likely that an individual belongs to the second group than to the first. However, a cutoff point separating the second and the third groups was found. A SvdH score close to 90 points allows to

distinguish between two groups: patients with less than 10 years of DD from patients with more than 10 years of DD (see Figure 2).

Given that it was not possible to distinguish between the first two groups, the patients were then separated in two groups only - less than and more than 10 years of DD - and the distributions were recalculated. In Table 3 the new values for the parameters of the posterior distributions can be found, as well as the estimated values for $\alpha$. Again $\alpha = 3$ is an acceptable value.

| (new) Group ($j$) | estimated $\alpha$ | $G_j$ | $H_j$ |
|---|---|---|---|
| 1 | 3.022168 | 537 | 15207 |
| 2 | 3.000716 | 540 | 27000 |

Table 3: Estimated values of $\alpha$ and parameters of the new posterior distributions.

A new cutoff point of 110 points for the SvdH score was found. See Figure 3. Then, the classification rule is:

- Classify the patient in group 1 (DD $\leq$ 10 years) if SvdH score $\leq$ 110
- Classify the patient in group 2 (DD $>$ 10 years) if SvdH score $>$ 110

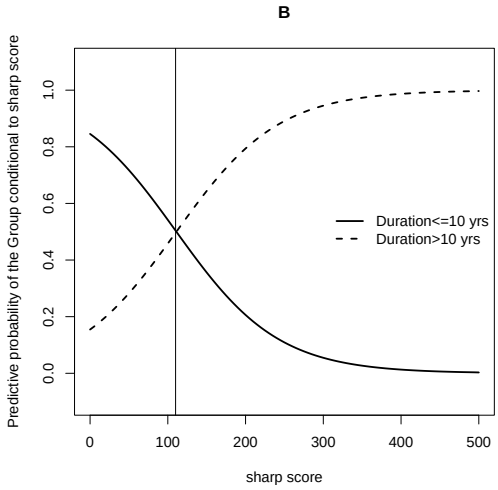


Figure 3: Estimated distribution of the Sharp/van der Heijde score (SvdH) according to the duration of the disease (DD)

This approach allow us to classify the patients in the study as having a "good prognosis", an "expected course of the disease" or a "bad prognosis"

by comparing the group the patient belongs to and the group the patient is classified in as shown in Table 4.

| | Disease Duration | |
|---|---|---|
| Sharp/van der Heijde score | $\leq 10$ years | $> 10$ years |
| $\leq 110$ (classify in group 1) | expected course | good prognosis |
| $> 110$ (classify in group 2) | bad prognosis | expected course |

Table 4: Classification rule and disease course evaluation according to SvdH score and DD.

The patients in the complete dataset were classified according to these rules. The results are presented in Table 5.

| | Disease Duration | |
|---|---|---|
| Sharp/van der Heijde score | $\leq 10$ years | $> 10$ years |
| $\leq 110$ | 133 (37.05%) | 78 (21.73%) |
| | expected course | good prognosis |
| $> 110$ | 46 (12.81%) | 102 (28.41%) |
| | bad prognosis | expected course |

Table 5: Classification and disease course evaluation for the complete dataset.

The results in Table 5 show that this cutoff point classified 65.46% of the patients as having had a disease course as theoretically expected (short DD and low SvdH score or long DD and high SvdH score), 21.73% as having had a disease course milder than theoretically expected, possibly revealing patients with prognosis (long DD and short SvdH score), and 12.81% as having had a disease course worse than theoretically expected, possibly depicting a subset of patients with bad prognosis (short DD and high SvdH score).

The cutoff point of 110 points was confirmed using cross-validation by separating randomly the available dataset into an experimental group (2/3 of the patients) and a test group (1/3 of the patients). All the distributions were recalculated using the data in the experimental group and the same cutoff point was found. Consequently, the patients in the test group were classified according to the same rule. As expected, the results (see Table 6) were similar to the ones achieved using the complete dataset, since the same classification rule was found.

| | Disease Duration | |
|---|---|---|
| Sharp/van der Heijde score | $\leq 10$ years | $> 10$ years |
| $\leq 110$ | 45 (37.50%) expected course | 25 (20.83%) good prognosis |
| $> 110$ | 16 (13.33%) bad prognosis | 34 (28.33%) expected course |

Table 6: Classification and disease course evaluation for the test set.

# 4 Conclusions

The method produces an extremely easy to use rule which may help doctors to evaluate the disease evolution, in particular in the task of identifying the subgroup of patients for whom the disease is causing more damage in the joints than it would be expected for the patients disease duration.

As far as the methodology is concerned, the conditions of application are easy to check and it can easily be extended to cases where $k \neq 3$.

# References