An ensemble of ordered logistic regression and random forest for child garment size matching

A. Pierola (1), I. Epifanio(2)(*), S. Alemany (1)

(1) Biomechanics Institute of Valencia, Universidad Politécnica de Valencia, Valencia 46022, Spain. Email addresses: ana.pierola@ibv.upv.es (Ana Pierola), sandra.alemany@ibv.upv.es (Sandra Alemany)

(2) Ph. 34 964728390, fax 34 964728429, epifanio@uji.es. Dept. Matemàtiques and Institut de Matemàtiques i Aplicacions de Castelló. Universitat Jaume I. Castelló 12071. Spain (*) Corresponding author.

Abstract

Size fitting is a significant problem for online garment shops. The return rates due to size misfit are very high. We propose an ensemble (with an original and novel definition of the weights) of ordered logistic regression and random forest (RF) for solving the size matching problem, where ordinal data should be classified. These two classifiers are good candidates for combined use due to their complementary characteristics. A multivariate response (an ordered factor and a numeric value assessing the fit) was considered with a conditional random forest. A fit assessment study was carried out with 113 children. They were measured using a 3D body scanner to obtain their anthropometric measurements. Children tested different garments of different sizes, and their fit was assessed by an expert. Promising results have been achieved with our methodology. Two new measures have been introduced based on RF with multivariate responses to gain a better understanding of the data. One of them is an intervention in prediction measure defined locally and globally. It is shown that it is a good alternative to variable importance measures and it can be used for new observations and with multivariate responses. The other proposed tool informs us about the typicality of a case and allows us to determine archetypical observations in each class.

Keywords: Multivariate conditional random forest; Proportional odds logistic regression; Supervised learning; Ordinal classification; Childrenswear garment fitting; Variable importance

ACKNOWLEDGEMENTS

This work has been partially supported by Grants DPI2013-47279-C2-1-R and DPI2013-47279-C2-2-R.

An ensemble of ordered logistic regression and random forest for child garment size matching ☆

Abstract

Size fitting is a significant problem for online garment shops. The return rates due to size misfit are very high. We propose an ensemble (with an original and novel definition of the weights) of ordered logistic regression and random forest (RF) for solving the size matching problem, where ordinal data should be classified. These two classifiers are good candidates for combined use due to their complementary characteristics. A multivariate response (an ordered factor and a numeric value assessing the fit) was considered with a conditional random forest. A fit assessment study was carried out with 113 children. They were measured using a 3D body scanner to obtain their anthropometric measurements. Children tested different garments of different sizes, and their fit was assessed by an expert. Promising results have been achieved with our methodology. Two new measures have been introduced based on RF with multivariate responses to gain a better understanding of the data. One of them is an intervention in prediction measure defined locally and globally. It is shown that it is a good alternative to variable importance measures and it can be used for new observations and with multivariate responses. The other proposed tool informs us about the typicality of a case and allows us to determine archetypical observations in each class.

Keywords: Multivariate conditional random forest, Proportional odds logistic regression, Supervised learning, Ordinal classification, Childrenswear garment fitting, Variable importance

1. Introduction

Selecting the right size of any garment without try on it is a difficult problem when buying these items both in store and, especially, in online garment shops (Ding et al., 2011). Users can base their decision on their previous experience. But each company has its own sizing, and what is more, this can change over time

☆The code and data for reproducing the results are available at *[illegible]ata are also available at *** (2016).

(Schofield and LaBat, 2005). This is therefore not a reliable strategy (Sindicich and Black, 2011).

There is usually a sizing chart corresponding to several anthropometric measurements, together with their ranges to know the size assignation. However, when the user's measurements belong to different size assignations depending on which measurements are considered, this strategy is also not a very good idea, since users cannot know which size will fit them best (Labat, 2007). As a consequence, there is a high percentage of returns due to size fitting problems, which represents one of the main costs of this sales channel for distributors and manufacturers. This also affects customer satisfaction (Otieno et al., 2005). The return rates of some e-commerce businesses are between 20 and 50% (Eneh, 2015). Furthermore, the fear of a poor fit is the main barrier to buying clothes online.

In the case of adults, the choice of size could be based on their sales and fit history, i.e., their virtual closet, but this methodology is not appropriate for children because they are growing (their anthropometric measurements are constantly changing). As a result, size matching in children should be based on their anthropometric measurements and their relationship with the garment measurements. Note that nowadays customers can obtain their detailed body sizes using their own digital cameras or other measuring technologies (Cordier et al., 2003; Ballester et al., 2015).

To address the garment size matching problem, a fit assessment study was carried out. The anthropometric measurements of 113 children were obtained using a 3D body scanner. Children tested different garments of different sizes, and their fit was assessed by an expert. This expert labeled the fit as 0 (correct), -1 (if the garment was small for that child), or 1 (if the garment was large for that child) in an ordered factor called Size-fit. Moreover, the fit was numerically assessed from 1 (very poor fit) to 10 (perfect fit) in a variable called Expert evaluation.

Finding the garment size that best fits the customer can be seen as a statistical classification problem (Meunier, 2000), i.e., given a new customer and his or her features, this new observation should be assigned to one of the predefined categories, i.e., sizes, based on a training set of observations whose features and size are known.

The classification problem has been widely studied and the number of methods developed for classification is enormous, ranging from simple linear discriminant analysis to more sophisticated methods such as neural networks. Hastie et al. (2009) provide an excellent explanation of supervised learning methods, including their strengths and weaknesses. For example, black box systems such as neural networks or support vector machines have a high predictive power, but their interpretability is poor, in contrast to regression models, such as ordered logistic regression, that offer human interpretable models. Decision trees can be consid-

ered as an off-the-shelf procedure for data mining (Hastie et al., 2009, Ch. 10). They are good at handling missing values and outliers, dealing with irrelevant inputs, etc., aspects in which recent methods can fail, but that are present in our problem. Predictive accuracy is the Achilles heel of trees. A bagging strategy can be considered to improve their individual performance. Random forest ensembles a large collection of trees (Breiman, 2001). Recently, trees and random forest for multivariate response variables have been defined in a conditional inference framework (Hothorn et al., 2006b).

However, garment size matching is not an easy problem due to uncertainties. For example, in the classical supervised classification paradigm it is typically assumed that the classes are objectively defined, with no arbitrariness or uncertainty about class labels, but this is not our case. See Hand (2006) for a discussion about other sources of uncertainty that are not generally considered in the classical supervised classification paradigm. Note that our class definition is more quantitative than qualitative. Furthermore, it could happen that none of the sizes fits the child well, or two sizes could be right sizes. In those cases, the gain made by recent, sophisticated statistical methods can be marginal (Hand, 2006).

In order to solve our garment size matching problem, we propose to investigate two classification methods with complementary strengths. On the one hand, the ordered logistic regression, with a stepwise model selection by Akaike's information criterion (AIC), fits a proportional odds logistic regression (POLR) to the ordered factor Size-fit. It is a parametric method, a generalized linear model that seeks global relationships between inputs and predicted classes. On the other hand, random forest (RF) is a nonparametric and highly nonlinear method. In our case, instead of a univariate response as usual, the inputs of the RF estimate a multivariate response: the ordered factor Size-fit and the numeric variable Expert evaluation.

We also propose to use RF to understand our data, and we define two new measures based on the RF results to achieve interpretability. One measure assesses the variable intervention in the prediction, locally (for each observation) or globally, so it can be used for new observations in contrast to the local (or casewise) variable importance (Breiman, 2003). The other proposed measure informs us about the location of the observation with respect to all other observations belonging to its class in the training set, as a measure of the class purity of the observation, going from observations in pure regions to observations in the boundary with other classes or even inside the decision region of other classes if we consider a class-wise map. Those cases are very important for discovering interesting patterns, detecting novelties or outliers in that class. A modification of this measure can also be used to define an archetypical observation of each class.

In our problem, POLR and RF are accurate and diverse, so they are good

complements for an ensemble (Dietterich, 2000). Both methods return a posteriori probabilities for each class. We propose to use a weighted majority rule based on the ranked probability scores to combine these probabilities and build the ensemble classifier (Kittler et al., 1998). Other methods are also studied, but POLR and RF offer very high performances, and also allow a good interpretation of the data. In other kinds of problems, such as estimating controlled direct effects of restrictive feeding practices, the combined estimator of logistic regression and random forests gave the smallest standard errors of the estimated controlled effects (Zhu et al., 2016).

The outline of the paper is as follows. Section 2 describes our data. The methodology is explained in Section 3, and the results obtained by applying this methodology to our database of children are shown in Section 4, together with a comparison with other methods, for a well-known benchmark data set too in Section 5. In Section 6 conclusions are given. To the best of our knowledge, this is the first manuscript about the garment matching problem in children.

2. Data

Information about fit was gathered in a fit assessment study of a selection of garments from two childrenswear companies. A set of garments was selected as the most representative for the fit assessment. Specifically, this set consisted of a shirt, a t-shirt and two pair of trousers for boys, and a t-shirt, a skirt and a dress for girls. The fit assessment study consisted of an experimental study with 113 children aged between 3 and 12 years. They were scanned and measured. A fit test was conducted for the garment sample and the fit of each garment was assessed. The number of children participating in the experimental study was balanced according to age ranges (3-4, 4-6, 6-8, 8-10, 10-12 years), with an equivalent number of boys and girls. The sizes of these companies are denoted as year 2, 3, 4, 5, 6, 8, 10 and 12.

During the fitting test, the selected garments were tried on the children and a questionnaire about fit was answered by a pattern making expert, regarding perception of garment fit and children's comfort. This expert was an experienced anthropometry technician with a degree in pattern making. The fit of the different garments was evaluated in three sizes specifically for current use: their supposed right size, the immediately smaller size and the immediately larger size, if these are manufactured. The assessment process was repeated for all three sizes and then the expert chose the size which best suited the child. However, sometimes not all children tested the three sizes for all the garments, but only two sizes or even one, depending on their degree of cooperation.

The gender, birth date, weight and height of the children were recorded, as well as their usual size for upper and lower body garments. The children were

also scanned in a standing position with a 3D body scanner. The children wore appropriate clothing for scanning, tight underwear and a head cap, in order to obtain better quality measurements. A Vitus Smart 3D body scanner from Human Solutions was used, which is a non-intrusive laser system formed by four columns allocating the optic system, which moves from the head to the feet in ten seconds performing a sweep of the body. From the 3D mesh, 34 anthropometric measurements were calculated semi-automatically with a digital tape measurement software, combining automatic measurements based on geometric characteristic points with a manual review. Certain physical markers were placed during the scanning process and virtual landmarks were marked on the children's scans in order to highlight anatomical references that were useful for measurement extraction: knee, ankle, armpit, belly, breast, crotch, head, hip, neck, nipple, thigh, elbow and acromion.

Thus, for a specific garment such as the t-shirt, each child in the experimental study generates several observations in the database. Each of these observations corresponds to one of the sizes which has been assessed on the child. Observations consist of the child's anthropometric information, the evaluated size and the results of the assessment process. The latter include the size which best fits the child according to the pattern making expert's criteria, if any (it could happen that none of the sizes fitted the child well). This expert could select only one size as the correct one for the child for that garment. The correct size was labeled as 0. The rest of sizes assessed were labeled -1 or 1 depending on whether the garment was smaller or larger. This is the Size-fit variable. If an intermediate, but non-existent size between two consecutive sizes would have been the right size, this is recorded in the Int-size variable. Furthermore, the expert evaluated the fit with a number between 1 and 10, where 1 means a very poor fit and 10 a perfect fit, and 6 a normal fit. This is the Expert evaluation variable. The expert only used integer numbers. There is not analytic relationship between Size-fit and Expert evaluation.

In childrenswear garments, a base size is used in pattern design, which is usually established for each population group. For fit verification of the garment prototypes of the base sizes, and due to the difficulty of using live models, childrenswear companies usually use mannequins. Sizes in companies are not usually determined by an optimization method (Ibáñez et al., 2012; Vinué et al., 2014; Domingo et al., 2014); but the remaining sizes are obtained by scaling the base size pattern with linear grading factors (Ashdown, 2014). Some commercial families of child mannequins, obtained as a result of anthropometric child studies are Children Formax®, ASTM Standard Child mannequins, and MNQ 0-12 from ASEPRI (Spanish Association of Children's Products), used as a standard in Spain. Both of the childrenswear companies included in this study use the latter to verify fit.

These physical mannequins were scanned and measured according to the previous procedure in order to characterize their dimensions, scaling them to the rest of the sizes.

Mannequin and child body shapes and dimensions were compared. The anthropometric measurements for each child observation were transformed by calculating the differences between the reference mannequin of the evaluated size and the child's anthropometric measurements. Hence, it is possible to gather the observations with a different evaluated size in the same data set and increase sample sizes. Otherwise, observations would have to be divided according to evaluated sizes, resulting in several data sets with small sample sizes. The idea behind this transformation is that perfect garment fit is obtained with a child body shape similar to the mannequin's. Therefore, dissimilarities between real children and mannequin body shapes will produce regions of poor fit.

In the interests of brevity, only the results for the t-shirt for boys and girls are analyzed in this paper. The complete list of anthropometric measurements used for this garment are provided in Table 1 of Section 4. Note that only the 27 variables that could influence according to design experts in the fitting of this garment are analyzed, instead of the whole set of 34 variables. For example, variables such as ankle perimeter are previously discarded in the t-shirt analysis.

3. Methodology

Let $\mathbf{Y}$ be the response matrix, and $\mathbf{X}$ the $N \times M$ matrix with M observed explanatory variables in N observations. In Section 3.2 $\mathbf{Y}$ is a vector, an ordered factor with K levels. In Section 3.3, $\mathbf{Y}$ is a matrix recording two responses for the N observations: the previous ordered factor with K levels and a numeric variable. The implementation of our methodology is written in the free software R (R Development Core Team, 2015) and it is available together with the database at http://*** . Data are also available at *** (2016).

3.1. Preliminary definitions

The following statistics will be used. Let us briefly review their meaning.

Definition 1. *The ranked probability score (RPS) for probabilistic forecasts of ordered events (Wilks, 2006; NCAR - Research Applications Laboratory, 2015) is a squared measure that compares the cumulative density function (CDF) of a probabilistic forecast with the CDF of the corresponding observation over a given number of discrete probability categories; therefore, the order of the classes is taken into account. The Brier score is a special case of an RPS with two categories. It measures how well the probability forecast predicts the category that the observation falls into. RPS ranges from 0 (perfect forecast) to 1. Note that RPS is more*

informative than accuracy, which simply takes into account whether or not the forecast is correct.

Definition 2. *Kendall's coefficient W is an index of interrater reliability of ordinal data (Gamer et al., 2012). Kendall's W ranges from 0 (no agreement) to 1 (complete agreement).*

Definition 3. *Yule's Q coefficient (Kuncheva and Whitaker, 2003; Yule, 1900) can assess the similarity of two classifier outputs. For statistically independent classifiers, the expected value is 0. Q ranges from -1 to 1. The greater the coincidence in the classification of both classifiers, the nearer to 1 Q is.*

3.2. Ordered logistic regression

Agresti (2002, Ch. 7) explains the cumulative link model in detail. The model is $logit\ P(Y <= k|\mathbf{x}) = \zeta_k$ - η, where the logit link function is the inverse of the standard logistic cumulative distribution function, i.e. $logit(p) = log(p/(1-p))$, ζ_k parameters provide each cumulative logit, and η is the linear predictor $\beta_1 x_1$ + ... + $\beta_M x_M$. A forward stepwise model selection using Akaike's information criterion (AIC) is performed to choose the model. When the parameters have been estimated, this model allows us to predict the class probabilities for a new observation. The *polr* and *extractAIC* functions from the R package **MASS** (Venables and Ripley, 2002) have been used in the implementation. The class assignation is implemented by choosing the class with the highest probability.

3.3. Random forest

Trees are a nonlinear regression technique. A tree is grown by binary recursive partitioning. In simple words, the generic idea behind binary recursive partitioning is to iteratively select one of the explanatory variables and the binary split in this variable in order ultimately to fit a constant model in each cell of the resulting partition, which constitutes the prediction. Two known problems with such models are overfitting and a selection bias towards explanatory variables with many possible splits or missing values. To solve these problems, Hothorn et al. (2006b) considered a conditional inference framework for recursive partitioning, which is also applicable to multivariate response variables. Details about how this algorithm carries out variable selection, stopping criteria, splitting criteria and other elements of their algorithm are thoroughly described in Hothorn et al. (2006b).

Trees are a low-bias but high-variance procedure, which makes them especially suited for bagging (Hastie et al., 2009). Growing an ensemble of trees significantly improves the accuracy. The term random forest was coined by Breiman (2001) for procedures where random vectors that govern the growth of each tree in the ensemble are generated. This randomness comes from randomly choosing a group of m ($m << M$) input variables to split on at each node and bootstrapping a sample from the training set. The non-selected cases are called out-of-bag (OOB).

3.3.1. RF computation

For the computation of random forest we have used the function $cforest$ from the R package **party** (Hothorn et al., 2006a; Strobl et al., 2008). For regulating the construction of the random forest, we have considered two settings (in both $NT = 500$ trees are grown): the one suggested by Strobl et al. (2007) through function *cforest_unbiased* and the one that mimics the behavior of $randomForest$ (Liaw and Wiener, 2002), which implements Breiman (2003)'s random forest algorithm. We refer to this setting as *cforest_classical*. According to Breiman and Cutler (2004), the only tuning parameter to which random forest is somewhat sensitive is m (Breiman, 2003). The smaller m is, the lower the correlation between trees and consequently the lower the error rate. However, the larger m is, the higher the strength of each tree and consequently the lower the error rate. To find a trade-off value m, Breiman and Cutler (2004) suggest choosing the m value that gives the lowest OOB error rate, which is an unbiased estimate of the test set error.

3.3.2. RF tools for data mining

Besides prediction, a random forest offers several helpful tools for data interpretation. One of these is variable importance measures. Proximities are another very useful tool, which makes it possible to determine outliers and prototypes. Let us review them briefly.

The importance of variable k is measured by averaging over all trees the decrease in accuracy between the prediction error for the OBB data of each tree and the same after permuting that predictor variable. It can be computed for each class, and it can also be computed casewise. The local or casewise variable importance is the increase in percent of times a case i is OOB and misclassified when the variable k is permuted.

The node impurity is measured by the Gini index for classification and by residual sum of squares for regression. Another importance measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. Due to the definition, it is a global measure for each variable, it is not defined locally or by class. According to Breiman and Cutler (2004), it is often very consistent with the mean decrease in accuracy importance measure. These importance measures can be computed with the R library **randomForest**.

The proximity between two cases i and j is defined as the proportion of trees where cases i and j are in the same terminal node. The outlying measure of a case i is computed as the quotient between the number of samples and the sum of the squared proximities between case i and the rest of cases in its class, normalized by subtracting the median and divided by the median absolute deviation, within its class. According to Breiman (2003), cases with values larger than 10 are outlier suspects. This outlying measure is available in the R library **randomForest**, as well as the proximities, although they can also be calculated with the R library

party.

With the proximities, we can compute the k nearest neighbors for each case. To define the prototype of each class, Breiman and Cutler (2004) identified the case that has most neighbors in each class, and then computed the medians of those neighbors coordinate-wise.

Proximities can also be visualized. A dissimilarity matrix $\boldsymbol{D}$ with elements $\{d_{ij}\}$ $i, j = 1, ..., N$ is built as one minus the matrix of proximities. Let $\boldsymbol{M}$ be the matrix with elements m_{ij} = -0.5* d_{ij}^2, and $\boldsymbol{B} = (\boldsymbol{I} - N^{-1}\boldsymbol{e}\boldsymbol{e}')\boldsymbol{M}(\boldsymbol{I} - N^{-1}\boldsymbol{e}\boldsymbol{e}')$, where $\boldsymbol{I}$ is the $N \times N$ identity matrix and $\boldsymbol{e}$ is the $N \times 1$ vector with all its elements equal to unity. The matrix $\boldsymbol{D}$ is Euclidean, only if $\boldsymbol{B}$ is positive semidefinite (Mardia et al., 1979, Theorem 14.2.1). If these dissimilarities are Euclidean distances, classical multidimensional scaling (cMDS) allows us to represent them exactly in at most N - 1 dimensions (Mardia et al., 1979, Theorem 14.4.1). With cMDS we obtain a set of points such that the distances between the points are approximately equal to the dissimilarities, since the dimension of the space which the data are to be represented in is usually less than N - 1. If these dissimilarities are a distance but not an Euclidean distance, we can apply cMDS as an approximation, which is optimal for a kind of discrepancy measure (Mardia et al., 1979, Theorem 14.4.2), or we can apply the h-plot (Epifanio, 2013), a technique that also performs well when the dissimilarity is not a distance, or any other multidimensional scaling methodology.

We propose two new tools that can help us understand the data.

3.4. New measures for data mining

3.4.1. Intervention in prediction measure

Variable importance was defined for univariate responses, and it is computed with data whose true response is known. In our problem, we have considered a multivariate response, and it would also be interesting to know which variables are involved in the prediction. The user could know which variables the particular prediction is mainly based on if this would help make a final decision. For those reasons, we define a new measure that we call: *Intervention in Prediction Measure* (IPM).

It can be defined globally, for each class and locally, as well as for multivariate responses, and for new cases for which we do not know their true response. Given a new case, it is put down each of the NT trees. For each tree, the case goes from the root node to a leaf through a series of nodes. We record the variable split in these nodes. For each tree, we compute the percentage of times a variable is split along the case's way from the root to the terminal node. Note that we do not count the percentage of times a split occurred on variable k in tree t, but only those variables that have intervened in the prediction of the case. The IPM for this new case is obtained by averaging those percentages over the NT trees. Therefore, for

IPM computation, we only need to know the trees that form the forest. The IPM for a case in the training set is obtained by considering and averaging over only those trees in which the case belongs to the OOB set. Once the casewise IPMs are estimated, the IPM can be computed for each class and globally, averaging over the cases in each class or all the cases, respectively.

3.4.2. Typicality measure

In data mining applications, the objective of a classification problem is not simply a class assignment, but class membership probabilities are often more interesting (Hastie et al., 2009). RF returns a matrix of class membership probabilities. In a similar vein, for a positive integer k, we consider the k nearest neighbors to a case i ($\{n_j^i\}$, $j = 1, ..., k$) according to the proximities defined by the RF, i.e., cases with the highest proximities to that case. Of these cases, they could either be of the same class and a weight w_j^i of one is assigned to them, or of other classes, in which case a weight of zero is assigned to them. We compute the weighted sum: $W_i^k = \sum_{j=1}^{k} w_j^i P(i, n_j^i)$, where P denotes the proximity between two cases obtained by RF. The quotient $Q_i^k = \sum_{j=1}^{k} w_j^i P(i, n_j^i) \ / \ \sum_{j=1}^{k} P(i, n_j^i)$ is a location measure that gives us information about the class purity of the case. This quotient can go from 0 to 1.

If we think of a class-wise map, the weights w_j^i for cases in pure regions will be one, and therefore that quotient will be one. For cases on the boundary with other classes, some weights will be one and others zero, and there will be more zeros if the case is inside the decision region of other classes, which leads to low values of W_i^k, near zero, and Q_i^k. This quotient is also a membership measure in the class, like the class probabilities returned by RF. In fact, the quotient values Q_i^k and the probabilities of belonging to each true class are often very consistent.

However, Q_i^k only takes into account whether its nearest neighbors belong to its class or not, but not whether the neighbor proximities are high or low. Let us suppose that we have a case i with low proximities to the majority of the other cases in its class, and lower proximities to the cases in other classes, then Q_i^k will be one. According to the definition of prototype proposed by Breiman and Cutler (2004), this case could be the one that has most neighbors in that class, but it was really not a good representative of its class as it had low proximities to the majority of the other cases in its class, in fact it could be an archetype or extreme of that class (archetypes were introduced by Cutler and Breiman (1994); for an archetype that is real observation, the term archetypoid was introduced by Vinué et al. (2015)).

For that reason, we define what we call *Typicality measure* (TM) for case i and k neighbors as $TM_i^k = \sum_{j=1}^{k} w_j P(i, n_j^i) \ / \ M_{c_i}^k$, where $M_{c_i}^k$ is the maximum Q_i^k among observations in the class c_i, to which case i belongs. TM_i^k takes values

from 0 to 1. For the situation described above, the TM value for that case will be low. The observation where that maximum is attained will have a TM value of one, and it can be considered as a prototype or a typical representative of that class, as it is the case in that class with the largest number of neighbors in its class and high proximities to it. According to the definition, a prototype can be determined for each class.

If there is an overlap between the observations in different classes, the number of observations with Q_i^k equal to one, i.e., having all the neighbors in the same class will decrease when k is increased and there will be a certain k where no point takes the value one for Q_i^k. It is interesting to know which the highest k_b^Q is for which Q_i^k is still one, and for which observation the value one is attained, for each class. This observation would be an archetype of the class or the purest observation, and the value k_b^Q for each class can inform us about the overlap between classes. Note that if k_b^Q is small the overlap is large.

With TM we assess not only whether there are many neighbors in the same class, but also the proximity to them. If k is small, observations with neighbors near to them in the same class, i.e., with a concentration of observations in that class will have high TM values, and their Q values will probably be one. As k increases, observations with TM_i^k equal to one do not necessarily correspond with those observations that have Q_i^k values equal to one. The highest k_b^{TM} value in each class for which the observation's TM value is one and its corresponding Q is also one will be determined. This observation can also be seen as an archetype of its class, and again the k_b^{TM} value for each class can inform us about the overlap between classes since if k_b^{TM} is small, the overlap is large.

Trees are invariant under strictly monotone transformations of the individual predictors, so the proximities generated by the random forest are immune to affine transformations of the predictors, as are therefore the measures Q, TM and their derivative.

3.5. Ensemble

An ensemble of classifiers is a set of classifiers whose individual predictions are combined in some way to classify new cases. Ensemble of classifiers have been used in many different fields (Perikos and Hatzilygeroudis, 2016; Yang et al., 2015). The accuracy obtained by ensembles is usually higher than that of the individual classifiers that comprise them. A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are diverse and accurate, i.e., their error rates are lower than random guessing (Dietterich, 2000). The more different the errors they make on new cases, the more diverse the classifiers are.

To combine the decision of POLR and RF, we use soft voting. We consider a convex combination of the predicted class probabilities for each classifier, i.e., we

consider weights whose sum is one, and multiply the predicted class probabilities for each classifier by the classifier weight, and take the average. The final class label is assigned based on the maximum of these weighted average probabilities.

There are many weighting methods for combining the classifier. See Rokach (2010) for a review about combination methods. From all of them, we have chosen a weighting based on performance due to its simplicity and successful results. However, we propose two novelties for determining the weights.

Firstly, the classical accuracy performance measure is not used. RPS is used as performance measure. The second novelty is the weight definition. Let S_i be the score performance of classifier i, with $i = 1, 2$ in our case. The closer to zero S_i is, the more accurate the classifier. One should believe accurate classifiers more than inaccurate ones, so more accurate classifiers have more weight. An intuitive definition of the weights w_i of each classifier is to set them proportional to their performance: $w_i = 1 - S_i/(S_1 + S_2)$. Nevertheless, in practice the classifier scores are not very different, and the effective calculation degenerates to a simple average or unweighted majority vote (Procopio, 2007). For example, if the scores of two classifier are 0.1 and 0.15, the weights would be 0.6 and 0.4.According to Procopio (2007), the actual range of the scores may be less than, say, 5% or even 2% of the mean score of those values. To solve this issue, he proposed weighting by ranking, as even very small differences in relative scores of the models will result in a ranking. However, information is lost by ranking. If weights are determined by ranking the scores, the same weights would be assigned, for instance, in the following two different situations: the scores of the two classifiers are 0.1 and 0.15 versus the situation where the scores are 0.1 and 0.101. In both situations, the weights by ranking would be 2/3 and 1/3.

Therefore, we propose to raise the scores to a power p: $w_i = 1 - S_i^p/(S_1^p + S_2^p)$ to expand the effective range of weight values. This is not a new concept. For example, input measures are raised to high powers in Podani and Miklós (2002) in multidimensional scaling, or image intensity values are gamma-corrected by a power-law expression in image processing. However, in this context it appears to be a novel idea not found elsewhere in the literature to the best of our knowledge. An intuitive technique for choosing p, which could be refined in the future, would be the following. In our problem, RPS values for the different classifiers in different subsets (girls and boys) range more or less from 0.1 to 0.15, so we would like to put the majority of weight, for example 90%, on the classifier with score 0.1, and 10% on the classifier with score 0.15. We have considered the function $f(x) = 1 - 0.1^p/(0.1^p + x^p)$ for different natural p values, and $p = 6$ is the one that satisfies the criterion, although the weights and results are quite similar in our problem for $p = 4$, 5 or 7.

4. Results

4.1. Exploring data by random forest

We applied the methodology explained in Section 3.3 to our database for boys in the case of t-shirts: 52 boys, with a total of 109 observations. The multivariate response was composed of the ordered factor Size-fit and the numeric variable Expert evaluation. With the setting suggested by Strobl et al. (2007) through function *cforest_unbiased*, the error rate for the OOB samples is 33.21%, whereas it is 27.52% for $m = 15$ (m values from 5 to 15 were considered) with the one that mimics the Breiman and Cutler (2004)'s random forest algorithm. We chose this last setting to estimate the proximities with the OOB samples.

4.2. TM analysis

For the dissimilarity matrix $\boldsymbol{D}$, the last 44 eigenvalues of $\boldsymbol{B}$ are negative, but they are small (the smallest is -0.4110788) in relation to the first two eigenvalues (9.541488, 6.469156). For a two-dimensional representation, the goodness-of-fit (GOF) measure α_2 proposed by Mardia et al. (1979, eq. 14.4.8) for the cMDS representation is 66.9%, whereas the goodness-of-fit measure for h-plotting (Epifanio, 2013) is 88.4%. The proximities are displayed in Figure 1 using the h-plot methodology. We can observe a horseshoe effect, which is usually observed when there is a latent ordering of the data (Diaconis et al., 2008). This is very reasonable in our case, since Size-fit is an ordered factor, ranging from cases for which the garment is very large to those where it is very small and passing through those with a good fit in between, as a continuum, without separation between groups.

The outlying measure is computed and the maximum value obtained by a case is 6.5, less than 10. According to the rule suggested by Breiman and Cutler (2004) this would imply that there are no outliers. However, looking at Figure 1 some points are clearly very far from their class, and quite far inside other classes. However, our proposed measure can detect this situation.

Firstly, we compute the measures Q_i^k and TM_i^k for $k = 21$, the default number in the R package **randomforest** (Liaw and Wiener, 2002) for computing the prototypes as Breiman and Cutler (2004) suggested. This number is generated by the fact that the number of cases in the group with smallest sample size is 22, and therefore the maximum number of potential neighbors in its class is 21. The Q_i^{21} values and the predicted probabilities in their true class are quite similar; in fact, the median absolute difference between these values is 0.06, and the maximum is 0.22. The values of Q_i^{21} and TM_i^{21} for the three points of the group Size-fit -1 situated at the bottom of Figure 1 are around 0.07 and 0.09 (the predicted

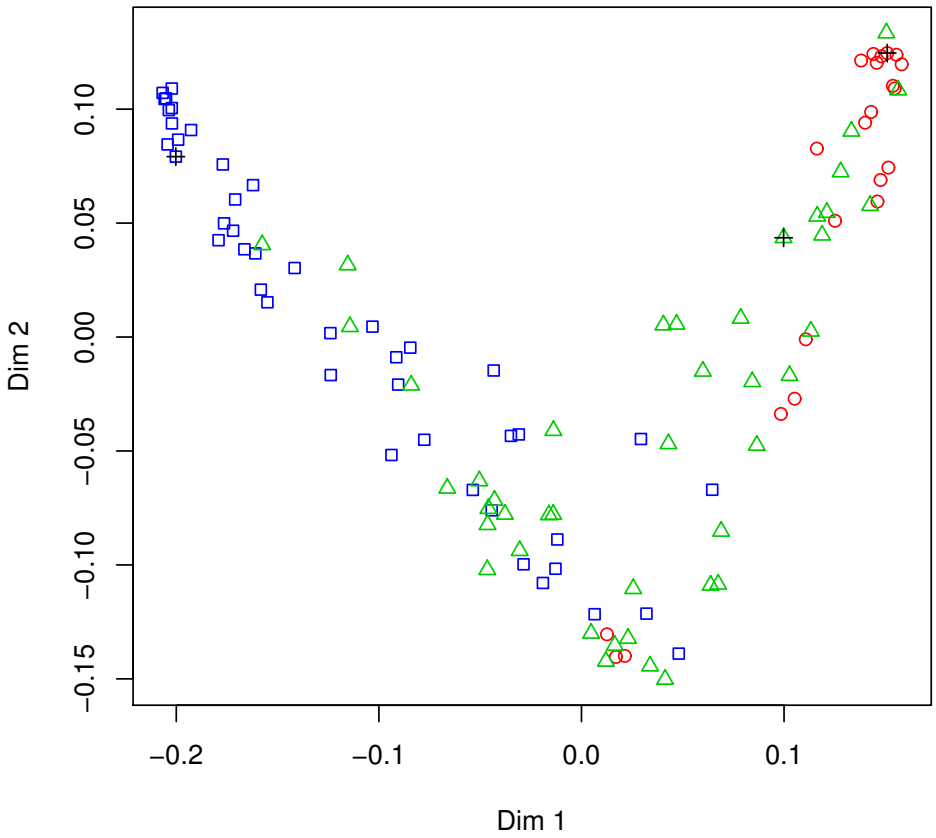

Figure 1: 2D h-plot representation of proximities obtained with RF. Cases with class labels -1, 0, 1 are represented by blue squares, green triangles and red circles, respectively. The representatives are marked with black crosses.

probabilities in their true class range from 0.1 to 0.15), indicating that they are quite far inside the decision region of other classes.

We follow the strategy explained in Section 3.3 to determine the representatives of each class. The k_b^Q are 7, 8 and 20 for class -1, 0 and 1, respectively. While the k_b^{TM} are 7, 7 and 19. This is consistent with that shown in Figure 1, where there is a larger overlap between classes 0 and -1,whose k_b^Q are small, than with class 0 and 1. The overlap between classes can be easily observed by looking at the location of the representative points. For classes -1 and 0, their representative points are nearer than the representative of class 1. The representatives obtained for k_b^{TM} are displayed with a black cross in Figure 1. They are near those obtained for k_b^Q which we omit in order to keep the figure clear.

4.3. IPM analysis

Let us study the importance of the variables. Three observations have missing values in the predictors. This was not a problem using the R library **party**, but the *randomForest* function from the R library **randomForest** does not handle missing values in the predictors directly. Although values can be imputed, we preferred simply not to consider these three observations, because we would have the same problem with POLR. Therefore, we now have 50 boys, with a total of 106 observations.

We computed the forest again with two responses with the setting *cforest_classical*, for the 106 cases. Now $m = 7$ is the lowest number that gives the highest accuracy (the error rate with OBB samples is 28.30%), and the five variables with the highest global IPM, in descending order, are: height, 7 cervical vertebrae (CV) height, hip girth, buttock girth and arm length coupled with 7 CV. The same variables are obtained if we compute the IPM for a new artificial case with a value of zero in all the variables, i.e. an artificial boy with all his measures equal to the mannequin. The importance measures cannot be computed for new cases.

Remember that variable importance measures cannot be computed for multivariate response, unlike for IPM. This is the reason why we compute a random forest with only one response, the ordered factor Size-fit, to compare IPM with the importance measures, with the R library **randomForest**. The lowest error rate for OBB samples is 27.36% for $m = 8$. Table 1 shows the importance measures and IPM by group and globally.

In global terms, the ranking of importance measures and IPM seems quite consistent at a glance; in fact, the ranking for the mean decrease in Gini index and IPM are very similar. We computed Kendall's coefficient W (Kendall, 1948) to assess the concordance between them. The agreement between the rankings of the mean decrease in Gini index and IPM is very high (0.995), higher than the rankings of the mean decrease in accuracy and IPM (0.956), which in turn is higher

Table 1: The first column is the name of the variables. The three following columns correspond to the ranking in decreasing order of the mean decrease in accuracy for each class, whereas the fifth column is the same but calculated globally (labeled as G). The sixth column is the ranking for the mean decrease in Gini index (labeled as Gini). The last four columns contain the ranking of the IPM values firstly by group and the last column computed globally (labeled as G).

Measures	Mean Decrease Accuracy				Gini	IPM			
	-1	0	1	G		-1	0	1	G
Body height	1	1	1	1	1	1	1	1	1
7 cervical vertebrae (CV) height	2	2	3	2	2	2	2	2	2
Mid neck girth	14	23	5	10	11	13	11	8	11
Neck at base girth	6	21	14	7	8	6	8	12	8
Head circumference	18	12	25	24	26	23	26	26	25
Horizontal shoulder width between acromia	20	11	21	23	17	24	20	17	19
Left shoulder length	25	18	26	26	22	21	22	22	23
Width armpits	24	14	16	22	25	25	25	27	26
Bust points width	21	4	18	18	19	15	19	18	17
Length neck-waist over chest	16	15	4	8	7	9	5	4	5
Bust point to neck	23	25	9	19	23	26	23	20	22
Bust/chest girth (horizontal)	17	16	10	12	14	14	15	14	14
Across back width (armpit level)	27	8	23	25	21	20	21	21	20
Length neck-armpits line	26	19	27	27	27	27	27	24	27
Vertical length neck-waist	22	5	24	20	24	22	24	25	24
Crotch length	4	26	15	11	9	7	9	13	9
Front crotch length	10	27	13	17	13	12	13	11	13
Rear crotch length	12	24	12	14	12	11	12	10	12
Waist girth	13	7	22	15	15	16	16	16	16
Buttock girth	9	20	7	6	6	8	6	6	6
Hip girth	3	9	2	3	3	4	3	3	3
Belly circumference	8	17	19	13	10	10	10	9	10
Arm length left to 7 CV	7	3	11	5	5	5	7	7	7
Arm length left	5	6	6	4	4	3	4	5	4
Upper arm length left	15	13	8	9	20	18	17	19	18
Upper arm girth	19	10	17	21	16	19	14	15	15
Wrist girth left	11	22	20	16	18	17	18	23	21

than the rankings of the mean decrease in accuracy and Gini index (0.954).

In class-specific terms, the ranking of the mean decrease in accuracy for class 0 is not consistent with that for class -1 and 1. Their Kendall's W are 0.545 and 0.539, respectively. Note that the ranking of mean decrease in accuracy for class 0 gives as the fourth and fifth most important the variables: Bust points width and Vertical length neck-waist, while the importance of both of them in other classes is very low. This is not consistent with what we might expect, as these variables are not in the pole position (or close to it) in design. On the contrary, it underestimates the importance of variables that are key to good fit such as Buttock and Hip girth, which are considered to be relevant in class -1 and 1. This is not the case for the rankings of IPM, which give information that is more consistent with that expected according to the variables used in apparel design. With IPM, the agreement between classes is high (greater than 0.95), and is consistent with the ranking of mean decrease in accuracy for classes -1 and 1. The Kendall's W for the ranking of mean decrease in accuracy and IPM between classes are 0.955, 0.586 and 0.905, for class -1, 0, and 1 respectively.

4.4. Classification

As mentioned above, individuals with missing values are not considered, so we therefore have 50 boys, with a total of 106 observations (21, 42 and 43 labeled -1, 0 and 1, respectively), and 58 girls, with a total of 119 observations (26, 42 and 51 labeled -1, 0 and 1, respectively).

4.4.1. POLR classification

In order to obtain an estimation of the error rate for POLR we use a leave-one-out strategy, but instead of leaving out an observation, each time we leave out a complete individual with all his/her observations. The methodology, which is explained in Section 3.2, is carried out without this individual, and then the predictions for all the observation for this individual are preserved to produce the leave-one-out estimate of the prediction error, because this process is repeated in turn for each of the individuals. The error rates are 28.30 and 26.05% for boys and girls, respectively.

4.4.2. RF classification

The same strategy is followed with RF. We delete each individual in turn, and the random forest with the $cforest\_classical$ setting is built for m = 5, ...,15, without that individual. The random forest with m giving the lowest error rate with OBB samples is selected, and the predictions for all the observations for that individual are preserved to produce the leave-one-out estimate of the prediction error. The predictions of the Expert evaluation variable are also preserved. The error rates are 27.36% and 27.73% for boys and girls, respectively. The summary of the absolute difference between the predicted and true values of Expert evaluation,

Table 2: Summary of the absolute difference between the predicted and true values of Expert evaluation, for boys and girls.

	Median	Third quartile	Maximum
Boys	1.32	1.99	6.76
Girls	1.28	1.96	5.36

which were integer numbers, can be seen in Table 2 for both genders. There is a maximum of only 1 point of difference for 50% of the observations for both boys and girls. For 75% of the observations the difference is smaller than 2 points. The difference between the true and predicted values is only greater than 4 points in absolute value for four and six observations for boys and girls, respectively.

4.4.3. Ensemble classification

The error rates for the simple parametric ordered logistic regression and the more recent nonparametric random forest are similar. It may sound surprising, but Hand (2006) already argued that the performance yielded by simple methods is usually as good as more modern sophisticated classification methods. When other sources of uncertainty are present, such as mislabeling or arbitrariness in the class definition, the gain made by recent, sophisticated statistical methods can be marginal. Note that class definition in our problem is based on the subjective assessment of the expert, with the constraint that one and only one size could be labeled as 0. Furthermore, the class definition is more quantitative than qualitative, in fact, the label for each observation comes with the numeric Expert evaluation variable.

As an example of this uncertainty, we have six observations with a label of -1 (different from 0), but with a high Expert evaluation value (four with a value of 7 and two with a value of 9), and nine observations labeled as 1 with a high Expert evaluation value (five with 7, three with 8 and one with 9). Furthermore, there is one observation labeled with 0, and an Expert evaluation value of 6 (for the remaining observations in class 0, the values are higher than 6). Not only that, but for that child an observation for a larger size, labeled 1, has an Expert evaluation of 7, which is greater than that of the right size.

A similar situation arose in Ruiz-Gazen and Villa (2007), where comparable results were obtained by the logistic regression and random forest method in a meteorological problem.

We have accurate classifiers, since their error rates are smaller than the random guessing error rate, which is $(K-1)/K$, i.e. 2/3, in our problem. Let us study the diversity of POLR and RF. To assess the diversity of both classifier, Yule's Q coefficient is used. The value of Q is 0.56 for boys and 0.91 for girls. Table 3 shows the relationship between POLR and RF classifiers for both boys and

Table 3: Relationship between the classifiers, for boys and girls.

	Boys		Girls	
	RF		RF	
POLR	wrong	correct	wrong	correct
wrong	14	16	23	8
correct	15	61	10	78

girls. It is confirmed that the results of both classifiers coincide more for girls than for boys, i.e., results are more diverse in the group of boys. There are 16 of 106 observations in boys correctly classified by RF, but not by POLR, and 15 in the opposite situation. Nevertheless, there are only 8 of 119 observations in girls correctly classified by RF, but not by POLR, and 10 in the opposite situation.

We combine the POLR and RF classifiers as explained in Section 3.5. Figure 2 shows the error rates estimated for weights *we* from 0 to 1 in a convex combination (*we* for RF and 1 - *we* for POLR) with an increment of 0.01, in the case of both genders. When the weight is zero, the error rate for POLR is obtained, whereas weight one gives the error rate for RF. Note that none of the error rates in the combination is higher than the highest of those of the source classifiers. RPS for POLR is 0.134 and 0.118, while RPS for RF is 0.102 and 0.116 for boys and girls, respectively. In the case of boys with $p = 6$, the weight for RF probabilities is $w_1 = 1 - 0.102^p/(0.102^p + 0.134^p)$, which is equal to 0.84, so the weight for POLR classifier probabilities is 0.16. Therefore, the error rate with the POLR-RF ensemble for boys is 25.47%. The error rate with the POLR-RF ensemble for girls is computed analogously and is 24.37%. For other values of p, other weights are selected. The models chosen for boys for $p = 4$, 5 and 7 give the following error rates, respectively: 26.42%, 24.53% and 27.36%; and for girls: 24.37% in all the cases.

4.4.4. Comparison with other methods

Table 4 shows performance statistics for several classification methods (besides POLR, RF and POLR-RF ensemble) for boys and girls by leaving one complete individual out: LDA refers to linear discriminant analysis (Venables and Ripley, 2002); LDAVS refers to LDA with variable selection as implemented in Weihs et al. (2005) (a stepwise forward variable/model selection using the Wilk's Lambda criterion); LDA-RF and LDAVS-RF are the ensemble of those methods; k-NN denotes the k-nearest neighbor classifier, where the number of neighbors is determined by leave-one-out cross validation in the internal cross-validation, which is a double or nested cross-validation (Venables and Ripley, 2002); MARS1 and MARS2 denote multivariate adaptative regression splines (MARS) with Size-fit as response and Size-fit together with Expert evaluation as multivariate response, respectively

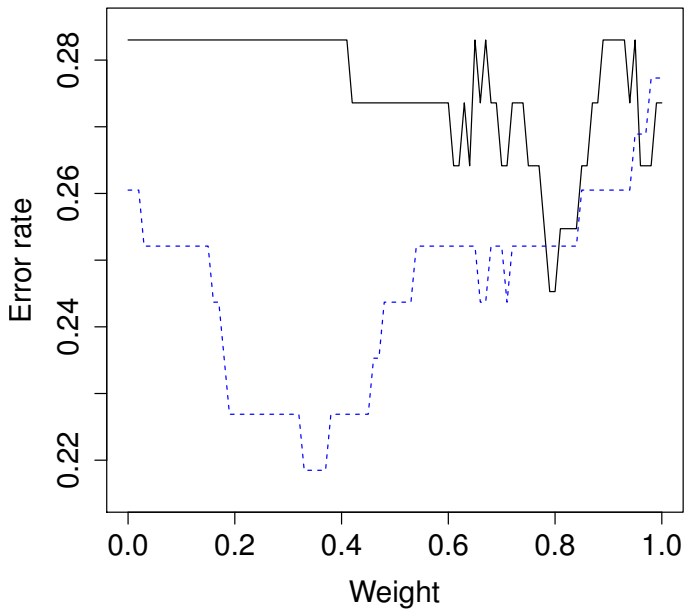

Figure 2: Error rates for weights in convex combination of POLR and RF for boys (solid line) and girls (dashed line). The error rate of the ensemble is determined by that corresponding to the weights w_i.

Table 4: Performance statistics for several classifiers, for boys and girls: accuracy, recall or sensitivity for each class, and precision or positive predictive value. The maximum value in each column appears in bold.

Method	Boys							Girls						
	Accuracy	Recall			Precision			Accuracy	Recall			Precision		
		-1	0	1	-1	0	1		-1	0	1	-1	0	1
POLR	0.72	0.71	0.6	0.84	0.65	0.66	0.8	0.74	**0.65**	0.67	0.84	0.68	0.64	**0.86**
RF	0.73	0.67	0.67	0.81	**0.88**	0.65	0.74	0.72	0.62	0.64	0.84	0.7	0.64	0.8
POLR-RF	0.75	0.62	**0.71**	0.84	0.81	0.67	0.80	**0.76**	**0.65**	0.71	0.84	**0.77**	**0.67**	0.83
LDA	0.75	0.81	0.62	0.84	0.74	0.71	0.78	0.58	0.46	0.52	0.69	0.57	0.48	0.67
LDAVS	0.7	0.57	0.67	0.79	0.67	0.61	**0.81**	0.65	0.46	0.55	0.82	0.67	0.53	0.72
LDA-RF	0.76	0.86	0.64	0.84	0.78	0.73	0.78	0.74	0.62	0.69	0.84	0.73	0.66	0.81
LDAVS-RF	0.72	0.62	0.69	0.79	0.72	0.63	**0.81**	0.74	0.62	0.69	0.84	0.73	0.66	0.81
k-NN	0.7	0.62	0.64	0.79	0.76	0.63	0.74	0.7	0.54	0.62	0.84	0.64	0.6	0.8
MARS1	0.67	0.62	0.57	0.79	0.65	0.62	0.72	0.65	0.62	0.6	0.71	0.64	0.54	0.75
MARS2	0.65	0.52	**0.71**	0.65	0.73	0.56	0.76	0.55	0.23	0.69	0.61	0.6	0.44	0.72
SVM	0.71	0.67	0.67	0.77	0.7	0.62	0.80	**0.76**	0.58	**0.74**	**0.86**	0.71	**0.67**	0.85
FDA	0.68	0.67	0.64	0.72	0.67	0.61	0.76	0.7	0.58	0.69	0.76	0.68	0.59	0.81
PPR	**0.78**	**0.95**	0.57	**0.91**	0.71	**0.83**	0.8	0.55	**0.65**	0.29	0.71	0.52	0.4	0.64
ClusterSVM	0.7	0.67	0.6	0.81	0.67	0.64	0.76	0.71	**0.65**	0.6	0.82	0.61	0.61	0.84

(Hastie and Tibshirani, 2015); SVM corresponds to a support vector machine as implemented in Meyer et al. (2012); FDA, the flexible discriminant analysis with MARS as implemented in Hastie and Tibshirani (2015); PPR refers to a projection pursuit regression model as implemented in R Development Core Team (2015), where the number of terms is chosen by an internal leave-one-out cross-validation. PPR consists of sums of nonlinearly transformed linear models, as neural network models; ClusterSVM refers to clustered support vector machines as implemented in He and Demircioglu (2015), proposed by Gu and Han (2013) and used by Harris (2015) for credit scoring. The number of clusters is set to 8, as recommended by Gu and Han (2013).

This comparison illustrates perfectly what Hand (2006) explained and we discussed previously: the apparent superiority of highly sophisticated methods in classification accuracy, obtained in "laboratory conditions", may be illusory and may not translate to a superiority in real-world conditions. POLR and RF give very good accuracies for both boys and girls, comparable to the results with SVM and k-NN, the strength of which is precisely predictive power, whereas their flaw is interpretability. In fact, the results with POLR-RF are better than with SVM for boys (equal for girls), and better for k-NN. The results with MARS are not good, especially for girls. The results with FDA are not so bad, but they do not improve on the results with POLR and RF. LDA and PPR give excellent results for boys and very poor results for girls. The results with LDAVS are good for boys and girls, but not as good as POLR or RF. The ensembles LDA-RF and LDAVS-RF are very good, similar to the results achieved with of POLR-RF. Although LDA could also be a good candidate for combining with RF, LDA does not take into account the order of the classes, as POLR does.

As regards the other measures, recall and precision, POLR-RF obtains the highest or second highest value in the majority of columns. In fact, the maximum is reached in 5 columns for POLR-RF, and the second highest value in another 5

columns in Table 4. None of the other methods give such results.

4.5. Decision for the user

Given a new child with their anthropometric measurements and several possible size choices, the first internal step is to calculate the differences between the reference mannequin of the possible sizes and the child's anthropometric measurements. These transformed measurements are used in our procedure, and a predicted label and predicted probabilities in each class for Size-fit is computed for each of the possible size choices, together with a predicted Expert evaluation value. The system does not make a decision for the user, but the information is supplied to the user and we leave the user to decide which size they wish to choose according to that information and their preferences.

We believe this is the best strategy because there are many subjective factors that can affect the user's decision. Note that the procedure is based on data about current garment use, but users may be thinking of future use, i.e. they may prefer the garment to be a little larger so that it can be worn now and in the future, maybe next year.

Observe that we are considering childrenswear and children grow over time. On the other hand, user may prefer to wear a loose garment or a tight garment. It could also happen that none of the tested sizes obtain a predicted zero label, indicating that none of them provide a good fit for that child, maybe because a non-existent intermediate size would be the right size (there were 7 boys and 9 girls recorded in Int-size). In this way, the user can decide on his/her risk and whether or not to buy the garment. In this case, IPM can be used together with the differences between the child and the mannequin to help the user take a decision. Note that garments that are slightly too large could be altered with a needle and thread. It could also happen that two of the tested sizes obtain a predicted zero label, indicating that both sizes would be appropriate for that child. In that case, the predicted Evaluation value could help users in their decision. Remember that in the experiment, the expert had the constraint that only one zero label could be assigned to a child.

If the user wanted to know which size fits him or her best, even if the fit is not perfect; i.e., if the user wanted to be given a specific recommendation, we could recommend the size with label equals zero. If two sizes had a predicted label of zero, then the one with the highest predicted Expert evaluation value could be chosen. If neither of the sizes had a predicted label of zero, the one with the highest predicted Expert evaluation value could be chosen. According to this strategy for selecting the right size, the percentage of errors for the 50 boys is 33.33% and 24.14% for the 58 girls, 28.70% overall, with the ensemble POLR-RF (for the cases where an intermediate non-existent size had been recommended by

Table 5: The first column is the name of the variables for the diabetes data set. The two following columns correspond to the ranking in decreasing order of the mean decrease in accuracy for each class, whereas the fourth column is the same but calculated globally (labeled as G). The fifth column is the ranking for the mean decrease in Gini index (labeled as Gini). The last three columns contain the ranking of the IPM values firstly by group and the last column computed globally (labeled as G).

Measures	Mean Decrease Accuracy			Gini	IPM		
	neg	pos	G		neg	pos	G
pregnant	3	8	5	7	7	7	7
glucose	1	1	1	1	1	1	1
pressure	8	6	8	8	8	8	8
triceps	7	5	7	6	6	6	6
insulin	4	3	7	2	2	2	2
mass	6	4	4	4	4	4	4
pedigree	5	7	6	5	5	5	5
age	2	2	2	3	3	3	3

the expert, the size with a higher Expert evaluation value given by the expert was considered as the right size in this computation).

These results are extremely good if compared with the error rates obtained if a child size chart was used to choose the correct size. If the size was selected according to the recommendation by Guerrero and ASEPRI (2000) based on the child's stature, the error rates would be 74% and 68.97% for boys and girls, respectively (71.30% overall). Note that the current garment sizing system for childrens wear does not accurately reflect the body sizes or current growth rate of todays children (Kilinc, 2011; Kang et al., 2001).

5. Analysis of a well-known benchmark data set

The well-known Pima Indians diabetes database (Lichman, 2013) is considered for illustration purposes. The version of the original data set corrected by Leisch and Dimitriadou (2010) is used, where details about the data set can be found. Eight continuous variables are used for classifying the data into two groups: negative (neg) or positive (pos) for diabetes. Table 5 shows the importance measures and IPM by group and globally. The ranking for the mean decrease in Gini index and IPM are identical, both for each class and globally. They are quite consistent with the ranking of the mean decrease in accuracy in global terms. However, in class-specific terms, the ranking of the mean decrease in accuracy for the variable pregnant is surprisingly very different for each class: it is the third most important variable in class neg, but the least important in class pos.

Table 6: Performance measures for several classifiers, for the diabetes data set: accuracy, recall or sensitivity, specificity, precision or positive predictive value, and negative predictive value (NPV), assuming class pos as positive class. The maximum value in each column appears in bold.

Method	Accuracy	Recall	Specificity	Precision	NPV
LR	0.7755	0.5538	0.8855	0.7059	0.8000
RF	0.7883	0.5923	0.8855	0.7196	0.8140
LR-RF	**0.7934**	0.5923	**0.8931**	**0.7333**	0.8153
LDA	0.7832	0.5769	0.8855	0.7143	0.8084
LDAVS	0.7883	0.5923	0.8855	0.7196	0.8140
LDA-RF	**0.7934**	**0.6000**	0.8893	0.7290	**0.8175**
LDAVS-RF	0.7908	0.5923	0.8893	0.7264	0.8147
k-NN	0.7066	0.4538	0.8321	0.5728	0.7543
MARS1	0.773	0.5615	0.8779	0.6952	0.8014
SVM	0.7602	0.5385	0.8702	0.6731	0.7917
FDA	0.7755	0.5692	0.8779	0.6981	0.8042
PPR	0.7653	**0.6000**	0.8473	0.6610	0.8102
ClusterSVM	0.6735	0.5000	0.7595	0.5078	0.7538

The different classification methods are compared below. Note that POLR cannot be used since classes are not ordered. Instead of POLR, logistic regression (LR) is considered. Note that RF is used with only one response instead of two, as in our application. As only one response appears in the diabetes data set, MARS2 cannot be considered. RPS values range from 0.147 to 0.151 in this data set, so $p = 85$ is considered for the ensemble. Performance measures can be seen in Table 6. The ensembles return the best results. The highest accuracies are achieved by the ensembles LR-RF and LDA-RF (the second highest is the ensemble LDAVS-RF). The highest values in recall are obtained by LDA-RF and PPR, while RF, LR-RF, LDAVS and LDAVS-RF obtain the second highest values. The maximum in specificity and precision is achieved by LR-RF, while LDA-RF is in second position for specificity and precision and LDAVS-RF for specificity. LDA-RF gives the highest NPV, and LR-RF the second highest.

6. Conclusions

Unlike the usual classification problem with nominal classes, the responses in our problem are ordered classes and a numeric variable. A random forest for multivariate response variables, which can handle this kind of data, has been used to explore and classify the data. We have proposed the use of an ensemble of POLR and RF to solve the size matching problem in children. These two classifiers are good candidates for combined use due to their complementary characteristics,

as the aggregation process is likely to lead to stronger and less biased results (Levin et al., 2016). The ensemble weights based on the classifier performances are determined in a novel way. RPS, which is specifically for ordered factors, is used instead of accuracy as a performance measure, since RPS is more informative. Furthermore, RPS is raised to a power to expand the effective range of weight values, rather than simply averaging the classifier results.

We have obtained promising results with our data. With ordinal data, the ensemble POLR-RF obtains the best results. Note that both POLR and RF take into account the class order. In fact, POLR is specifically designed for ordered classes, unlike LDA or LDAVS, which are also used in an ensemble with RF. For the diabetes data set, which is a binary classification problem with nominal classes, the ensembles LR-RF, LDA-RF and LDAVS-RF give the best results. Therefore, the proposed ensemble is not restricted to classification of ordered classes.

We have introduced two new measures based on RF with multivariate responses to achieve data interpretation. We have shown that local, by class and global IPM is a good alternative to variable importance measures and it can be used with multivariate responses, or any kind of response, and for new observations, unlike the classical variable importance measures. IPM results by class have given information that is more consistent with that expected than the measure based on the mean decrease in accuracy by class. We have also seen the advantage of IPM in a binary classification problem such as diabetes, over the classical variable importance measures. IPM results are consistent with the results of the measure based on the mean decrease in Gini index, but this measure is only defined globally. The other proposed tool informs us about the typicality of a case and allows us to determine archetypical observations in each class. These two new measures can be used in other classification problems to achieve interpretability.

Instead of calculating the differences between the reference mannequin of the evaluated size and the child's anthropometric measurements, the proportions between both group of measurements were also considered, but differences give the best results. We have treated all the misclassifications equally here. If misclassifying an observation in a smaller size, for example, were considered more serious, observation weighting could be used (Hastie et al., 2009, Ch. 10).

As future work, this methodology will be tested with measurements taken by users with digital domestic technology. It is expected that if more data were provided, a better learning would be achieved and also results could be improved. The sample size could be increased over time by knowing garment sales and customers' anthropometric measurements, and adding some feedback from users. As discussed by Robinette and Veitch (2016), sustainable sizing methods improve efficiency.

References

 2016. Child t-shirt size data set. Data in Brief (submitted).

Agresti, A., 2002. Categorical Data Analysis. Wiley.

Ashdown, S., 2014. 2 - creation of ready-made clothing: the development and future of sizing systems. In: Faust, M.-E., Carrier, S. (Eds.), Designing Apparel for Consumers. Woodhead Publishing Series in Textiles. Woodhead Publishing, pp. 17 – 34.

Ballester, A., Parrilla, E., Vivas, J. A., Pierola, A., Uriel, J., Puigcerver, S. A., Piqueras, P., Solve, C., Rodriguez, M., Gonzalez, J. C., Alemany, S., 2015. Low-cost data-driven 3D reconstruction and its applications. In: Proc. of 6th Int. Conf. on 3D Body Scanning Technologies, doi:10.15221/15.184. Lugano, Switzerland.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., 2003. Manual On Setting Up, Using, and Understanding Random Forests V4.0. Statistics Department, University of California, Berkeley.

Breiman, L., Cutler, A., 2004. Random forests.
URL `http://www.stat.berkeley.edu/~breiman/RandomForests/`.

Cordier, F., Seo, H., Magnenat-Thalmann, N., 2003. Made-to-measure technologies for an online clothing store. IEEE Computer Graphics and Applications 23 (1), 38–48.

Cutler, A., Breiman, L., 1994. Archetypal Analysis. Technometrics 36 (4), 338–347.

Diaconis, P., Goel, S., Holmes, S., 2008. Horseshoes in multidimensional scaling and local kernel methods. The Annals of Applied Statistics 2 (3), 777–807.

Dietterich, T. G., 2000. Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems. MCS '00. Springer-Verlag, London, UK, pp. 1–15.

Ding, Y.-S., Hu, Z.-H., Zhang, W.-B., 2011. Multi-criteria decision making approach based on immune co-evolutionary algorithm with application to garment matching problem. Expert Systems with Applications 38 (8), 10377 – 10383.

Domingo, J., Ibáñez, M. V., Simó, A., Dura, E., Ayala, G., Alemany, S., 2014. Modeling of female human body shapes for apparel design based on cross mean sets. Expert Systems with Applications 41 (14), 6224 – 6234.

Eneh, S., 2015. Showroom the future of online fashion retailing 2.0: Enhancing the online shopping experience. Master's thesis, University of Borås, Faculty of Textiles, Engineering and Business, Borås, Sweden.

Epifanio, I., 2013. h-plots for displaying nonmetric dissimilarity matrices. Statistical Analysis and Data Mining 6 (2), 136–143.

Gamer, M., Lemon, J., Singh, I. F. P., 2012. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.
URL `http://CRAN.R-project.org/package=irr`

Gu, Q., Han, J., 2013. Clustered support vector machines. In: AISTATS. Vol. 31 of JMLR Workshop and Conference Proceedings. JMLR.org, pp. 307–315.

Guerrero, J., ASEPRI, 2000. Estudio de tallas y medidas de la población infantil internacional. Asociación Española de Fabricantes de Productos para la Infancia (ASEPRI).

Hand, D. J., 2006. Classifier technology and the illusion of progress. Statistical Science 21 (1), 1–14.

Harris, T., 2015. Credit scoring using the clustered support vector machine. Expert Systems and Applications 42 (2), 741–750.

Hastie, T., Tibshirani, R., 2015. mda: Mixture and flexible discriminant analysis. R port by Leisch, F., Hornik, K. and Ripley, B. D. R package version 0.4-8.
URL `http://CRAN.R-project.org/package=mda`

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data mining, inference and prediction. 2nd ed., Springer-Verlag.

He, T., Demircioglu, A., 2015. SwarmSVM: Ensemble Learning Algorithms Based on Support Vector Machines. R package version 0.1.
URL `https://CRAN.R-project.org/package=SwarmSVM`

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., Laan, M. V. D., 2006a. Survival ensembles. Biostatistics 7 (3), 355–373.

Hothorn, T., Hornik, K., Zeileis, A., 2006b. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15 (3), 651–674.

Ibáñez, M. V., Vinué, G., Alemany, S., Simó, A., Epifanio, I., Domingo, J., Ayala, G., 2012. Apparel sizing using trimmed PAM and OWA operators. Expert Systems with Applications 39 (12), 10512 – 10520.

Kang, Y., Choi, H.-S., Do, W. H., 2001. A study of the apparel sizing of children's wear - an analysis of the size increments utilized in childrens wear based on an anthropometric survey. International Journal of Human Ecology 2 (1), 95–110.

Kendall, M., 1948. Rank correlation methods. Oxford, England: Griffin.

Kilinc, N., 2011. Assessment of the size charts of apparel business in online clothing sale for children between the ages of 3 and 6. INDUSTRIA TEXTILA 62 (5), 248–253.

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3), 226–239.

Kuncheva, L., Whitaker, C., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51 (2), 181–207.

Labat, K., 2007. Sizing standardization. In: Sizing in Clothing: Developing effective sizing systems for ready-to-wear clothing. Elsevier Ltd, pp. 88–107.

Leisch, F., Dimitriadou, E., 2010. mlbench: Machine Learning Benchmark Problems. R package version 2.1-1.

Levin, I., Pomares, J., Alvarez, R. M., 2016. Using machine learning algorithms to detect election fraud. In: Alvarez, R. M. (Ed.), Computational Social Science. Cambridge University Press, pp. 266–294.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.
URL `http://CRAN.R-project.org/doc/Rnews/`

Lichman, M., 2013. UCI machine learning repository.
URL `http://archive.ics.uci.edu/ml`

Mardia, K., Kent, J., Bibby, J., 1979. Multivariate Analysis. Academic Press.

Meunier, P., 2000. Use of body shape information in clothing size selection. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 44 (38), 715–718.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2012. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1.
URL `http://CRAN.R-project.org/package=e1071`

NCAR - Research Applications Laboratory, 2015. verification: Weather Forecast Verification Utilities. R package version 1.42.
URL http://CRAN.R-project.org/package=verification

Otieno, R., Harrow, C., Lea-Greenwood, G., 2005. The unhappy shopper, a retail experience: exploring fashion, fit and affordability. International Journal of Retail & Distribution Management 33 (4), 298–309.

Perikos, I., Hatzilygeroudis, I., 2016. Recognizing emotions in text using ensemble of classifiers. Engineering Applications of Artificial Intelligence 51, 191201.

Podani, J., Miklós, I., 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology 83 (12), 3331–3343.

Procopio, M. J., 2007. An experimental analysis of classifier ensembles for learning drifting concepts over time in autonomous outdoor robot navigation. Ph.D. thesis, University of Colorado, Boulder, Colorado, United States.

R Development Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org/

Robinette, K. M., Veitch, D., 2016. Sustainable sizing. Human Factors: The Journal of the Human Factors and Ergonomics Society 58, 657–664.

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intelligence Review 33 (1), 1–39.

Ruiz-Gazen, A., Villa, N., 2007. Storms prediction: Logistic regression vs random forest for unbalanced data. Cases Studies in Business, Industry and Government Statistics (CSBIGS) 1 (2), 91–101.

Schofield, N. A., LaBat, K. L., 2005. Exploring the relationships of grading, sizing, and anthropometric data. Clothing and Textiles Research Journal 23 (1), 13–27.

Sindicich, D., Black, C., 2011. An assessment of fit and sizing of men's business clothing. Journal of Fashion Marketing and Management: An International Journal 15 (4), 446–463.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9 (307).
URL http://www.biomedcentral.com/1471-2105/9/307

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8 (25).
URL http://www.biomedcentral.com/1471-2105/8/25

Venables, W. N., Ripley, B. D., 2002. Modern Applied Statistics with S, 4th Edition. Springer, New York, ISBN 0-387-95457-0.
URL http://www.stats.ox.ac.uk/pub/MASS4

Vinué, G., Epifanio, I., Alemany, S., 2015. Archetypoids: A new approach to define representative archetypal data. Computational Statistics & Data Analysis 87, 102 – 115.

Vinué, G., León, T., Alemany, S., Ayala, G., 2014. Looking for representative fit models for apparel sizing. Decision Support Systems 57 (0), 22–33.

Weihs, C., Ligges, U., Luebke, K., Raabe, N., 2005. klar analyzing German business cycles. In: Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.), Data Analysis and Decision Support. Springer-Verlag, Berlin, pp. 335–343.

Wilks, D., 2006. Statistical Methods in the Atmospheric Sciences. Academic Press.

Yang, L., Liu, S., Tsoka, S., Papageorgiou, L. G., 2015. Sample re-weighting hyper box classifier for multi-class data classification. Computers & Industrial Engineering 85, 44 – 56.

Yule, G. U., 1900. On the association of attributes in statistics. Philosophical Transactions A (194), 257–319.

Zhu, Y., Ghosh, D., Coffman, D. L., Savage, J. S., 2016. Estimating controlled direct effects of restrictive feeding practices in the early dieting in girls study. Journal of the Royal Statistical Society: Series C (Applied Statistics) 65 (1), 115–130.