FROM CROWDSOURCED RANKINGS TO AFFECTIVE RATINGS

Yoann Baveye[⋆†]*, Emmanuel Dellandréa*[†]*, Christel Chamaret*[⋆]*, Liming Chen*[†]

[⋆]Technicolor
975, avenue des Champs Blancs, 35576 Cesson Sévigné, France
{yoann.baveye, christel.chamaret}@technicolor.com
[†]Université de Lyon, CNRS
Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, Lyon, France
{emmanuel.dellandrea, liming.chen}@ec-lyon.fr

ABSTRACT

Automatic prediction of emotions requires reliably annotated data which can be achieved using scoring or pairwise ranking. But can we predict an emotional score using a ranking-based annotation approach? In this paper, we propose to answer this question by describing a regression analysis to map crowdsourced rankings into affective scores in the induced valence-arousal emotional space. This process takes advantages of the Gaussian Processes for regression that can take into account the variance of the ratings and thus the subjectivity of emotions. Regression models successfully learn to fit input data and provide valid predictions. Two distinct experiments were realized using a small subset of the publicly available LIRIS-ACCEDE affective video database for which crowdsourced ranks, as well as affective ratings, are available for arousal and valence. It allows to enrich LIRIS-ACCEDE by providing absolute video ratings for the whole database in addition to video rankings that are already available.

Index Terms— Gaussian Processes for Regression, Outlier detection, Affective video database, Affective computing

1. INTRODUCTION

Large multimedia databases are essential in various fields. They can be used to measure the performance of different works, enable benchmarks, and they can also be used to learn and test computational models using machine learning. The Affective Computing field is no exception to the rule and some well-known picture databases are already intensively used by researchers, such as the IAPS database [1]. But until very recently, there did not exist affective video databases publicly available large enough to be used in machine learning. Existing databases were too small, not sufficiently realistic or suffered from copyright issues. That is why Soleymani *et al.* wrote in [2] that

> "High-quality corpora will help to push forward the state of the art in affective video indexing."

Recently, a large affective video database called LIRIS-ACCEDE [3] has been made publicly available [1] in an attempt to solve most of these issues. In this database, 9,800 video excerpts have been annotated with pairwise comparisons using crowdsourcing along the induced dimensions of the 2D valence-arousal emotional space. Valence can range from negative (*e.g.*, sad, disappointed) to positive (*e.g.*, happy, elated), whereas arousal varies from inactive (*e.g.*, calm, bored) to active (*e.g.*, excited, alarmed). All the video clips being ranked along both dimensions, the rankings provide no information about the distances between them. Furthermore, these ranks are relative to this particular database which prevents the comparison with other video clips annotated with absolute valence and arousal values.

This is why the goal of this paper is to enrich the LIRIS-ACCEDE database by providing absolute video ratings in addition to video rankings that are already available. The new absolute video ratings are generated thanks to a regression analysis, allowing to map the ranked database into the 2D valence-arousal affective space. Gaussian Processes for Regression were preferred over other existing regression techniques since they can model the noisiness from measurements and thus take into account the subjectivity of emotions. The proposed regression analysis is performed using rating values collected on a subset of the database from a previous experiment. Results show that the predictions of our models are in line with these affective rating values and thus are able to estimate affective ratings from crowdsourced ranks.

The paper is organized as follows. Section 2 provides background material on affective multimedia self-reporting methods. Section 3 presents the database, the crowdsourced experiment that led to the ranking of the whole database and the controlled experiment in which arousal and valence ratings are collected for a subset of the database. For the purpose of estimating scores in the affective space from crowdsourced ranks, a regression analysis is performed in Section 4. Next, in Section 5, the outliers are detected and removed from the

[1] http://liris-accede.ec-lyon.fr/

process. Results are discussed in Section 6 and finally, conclusion and future work end the paper in Section 7.

2. RELATED WORK

In previous works, three different self-reporting methods with different advantages and drawbacks have been used to annotate affective multimedia databases: ratings, continuous annotations and ranking approaches.

Rating-based are the most popular in experiments to annotate a collection of movies. Annotators are asked to select on a rating scale the score that best represents the emotion induced by the video segment. For example, this process has been used to annotate, among others, the HUMAINE [4], FilmStim [5] and DEAP [6] affective video databases. The advantage of the scoring method is that annotators rate a video clip independently of any other video clip and thus, the scores can be easily compared even to those of video clips annotated in other experiments. But indeed, requesting an affective score requires annotators to understand the range of the emotional scale which is a sizable cognitive load. Furthermore, it may be quite difficult to ensure that the scale is used consistently [7], especially in experiments using crowdsourcing [3].

Continuous ratings are even more difficult to implement. Recently, Metallinou and Narayan investigated the challenges of continuous assessments of emotional states [8]. This technique is promising since the precision of the annotations is better than any other technique where one annotation is made for a large video segment. However, there are several issues that have to be addressed such as user-specific delays or the aggregation of multiple continuous subjective self-assessments.

Ranking approaches and more specifically pairwise comparisons are easier as they require less cognitive load. Annotators agree more when describing emotions in relative terms. Thus, pairwise comparisons enhance the reliability of the ground truth. This method has been used by Yang and Chen [9] to annotate the emotions induced by a collection of songs, and more recently in [3] to rank the LIRIS-ACCEDE affective video database described in the next section. Some approaches, such as [10], are able to derive ratio scale measures of the stimuli from only binary judgments from pairwise comparisons but they are not applicable in experiments dealing with a large number of stimuli. To deal with a huge number of stimuli, the quicksort algorithm can be used in order to reduce the amount of pairwise comparisons needed to sort the stimuli. But the main disadvantage of rating-by-comparison experiments, for which comparisons are generated using the quick sort algorithm as in [3], is that relative ranks do not provide indication about the distance between the excerpts, *i.e.* we do not know how much lower or higher a valence or arousal rank is than another one. It is also not sure that the median value represents a neutral data. In order to solve these drawbacks, Yang and Chen explored the possibility of representing songs in the emotion space according to relative emotion rankings but the ranks were converted to emotion scores in a linear way.

Fig. 1. Some examples of key frames extracted from several video clips included in the LIRIS-ACCEDE database. Credits can be found at http://liris-accede.ec-lyon.fr/database.php.

In this work, we show that it is possible to map the relative ranks of the 9,800 excerpts of the LIRIS-ACCEDE database into the valence-arousal emotional space in a finer way, thus combining the advantages of both the rating-based and ranking self-reporting methods, while considering the variability of annotations in emotion.

3. LIRIS-ACCEDE DATABASE

First, we describe the LIRIS-ACCEDE database used in this work and the previous experiments that led to the annotation of the whole database.

3.1. Description

LIRIS-ACCEDE is composed of 9,800 video clips extracted from 160 movies shared under Creative Commons Licenses [3]. Thus, the database can be shared publicly without copyright issues. The excerpts included in the database have been extracted from movies labeled under several movie genres such as comedy, drama, horror, documentary or even animation movies. Thus, the database is representative of current

most popular movie genres. The 9,800 video segments last between 8 and 12 seconds. It is large enough to get consistent video clips allowing the viewer to feel emotions and it is also small enough to reduce the probability for a viewer to feel more than one emotion per excerpt. In order to extract consistent segments from movies, a fade and scene cut detection based on [11] has been implemented to make sure that each extracted segment starts and ends with a fade or a scene cut. There are numerous different video scenes reflecting the variety of selected movies. This variety can be seen on Figure 1 showing the key frame of several excerpts included in the database.

Annotating such a large amount of data in laboratory would be very time-consuming. That is why crowdsourcing has been used to rank the whole database along the induced arousal and valence axes.

3.2. Crowdsourced ranking experiment

The annotations were gathered using the CrowdFlower platform, since it reaches several crowdsourcing services. Thus, collected annotations were made from workers with more different cultural backgrounds than experiments made only on one crowdsourcing service particularly popular for a few countries. Figure 2 shows that for the experiment made on CrowdFlower to rank the database along the induced arousal axis, the distribution of the country of annotators was very different depending on the crowdsourcing service. For example, most workers on Amazon Mechanical Turk were American and Indian whereas for instaGC workers are mostly American, Canadian and British.

To make reliable annotations as simple as possible, pairwise comparisons were proposed on CrowdFlower in order to rank the database along the induced arousal and valence axes. The ranking of the whole database was a two-stage process. First, annotations to rank the database along the induced valence axis were gathered. Workers were simply asked to select the video clip inducing the most positive emotion. Then, in the second step, workers were asked to select the excerpt inducing the calmest emotion, allowing to rank the database along the arousal axis. For both ranking experiments, the pairwise comparisons were generated and sorted using the quick sort algorithm. Workers were paid 0.01$ per comparison. More than 4,000 annotators participated in both stages, allowing to rank the LIRIS-ACCEDE database along the induced valence and arousal axes. Note that a more detailed description of the ranking process is given in [3].

3.3. Controlled rating experiment

To cross-validate the annotations gathered from various uncontrolled environments using crowdsourcing, another experiment has been created to collect ratings for a subset of the database in a controlled environment.

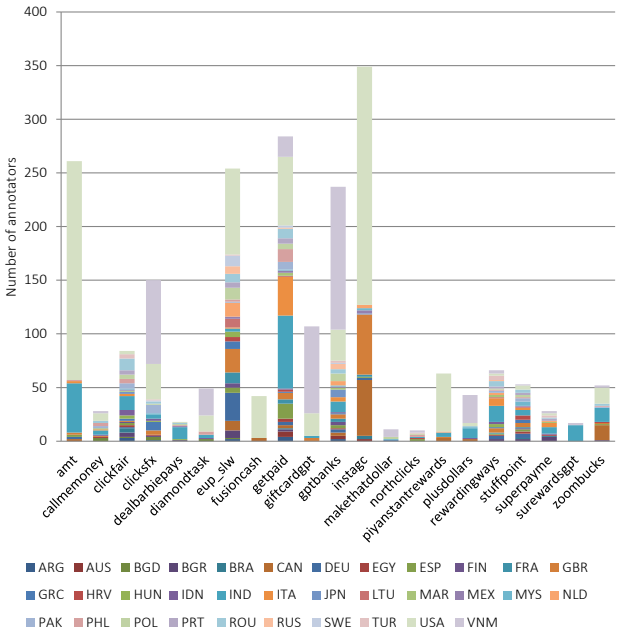


Fig. 2. Distribution of most represented countries for major crowdsourcing services used in October 2013.

In this controlled experiment, 28 volunteers were asked to rate a subset of the database carefully selected using the 5-point discrete Self-Assessment-Manikin scales for valence and arousal [12]. 20 excerpts per axis that are regularly distributed have been selected in order to get enough excerpts to represent the whole database while being relatively few to create an experiment of acceptable duration. Thus, their optimum rank positions are those with value $\frac{9800}{19} \times n$ with $n \in \{0, \ldots, 19\}$. For each axis and each optimum rank, we select the excerpt $i \in \{0, \ldots, 9799\}$ with $\alpha^i \geq 0.6$ as close as possible to the optimum rank, where α^i is the Krippendorff's alpha [13] measuring the inter-rater reliability of excerpt i during the crowdsourced ranking experiment. This process ensures that the selected film clips have different levels of valence and arousal and thus are representative of the full database while being highly reliable in eliciting such induced emotions. To rate selected excerpts, participants were instructed to focus on the emotion they felt while watching the 40 video clips. All the videos were presented in a random order and the participants had to perform immediately a self-assessment of their level of arousal or valence directly after viewing each film clip. Finally, all the annotations of a video clip were averaged to compute the affective rating of the excerpt. Due to the 5-point discrete scales used in this experiment, affective ratings range from 1 to 5.

The Spearman's rank correlation coefficient (SRCC) between the rankings of the film clips in the LIRIS-ACCEDE database and the ratings collected in this experiment exhibited a statistically highly significant correlation for both arousal

(SRCC = 0.751) and valence (SRCC = 0.795), thus validating the annotations made from various uncontrolled environments.

Hence, the goal of this paper is to use these rankings and ratings available for the 40 video clips selected for this second experiment to perform a regression analysis between the rankings and the ratings to convert the relative rankings into absolute scores.

4. REGRESSION ANALYSIS

Among all existing regression models, we used the Gaussian Processes for Regression as they can model the noisiness from measurements and thus take into account the subjectivity of emotions.

Two different Gaussian Process Regression Models are learned in this part, one for the valence axis and a second one for arousal. From the rank given as input (ranging from 0 to 9799), the goal of the models is to predict its affective rating for the dedicated axis (ranging from 1 to 5). To learn the models, we will use the crowdsourced ranks and the corresponding affective ratings gathered in section 3.3. The variance of the annotations gathered in this controlled rating experiment will be used to provide guidance to learn the models. Thus, these variances will be needed only during the learning step and will no longer be necessary to predict new affective ratings.

Knowing the rank x for valence or arousal of a video clip in the database, the goal of the Gaussian Processes regression models is to estimate the score $g(x)$ of the video clip in the affective space. Rasmussen and Williams [14] define a Gaussian Process (GP) as a collection of random variables, any GP finite number of which have a joint Gaussian distribution. The predictions from a GP model take the form of a full predictive distribution:

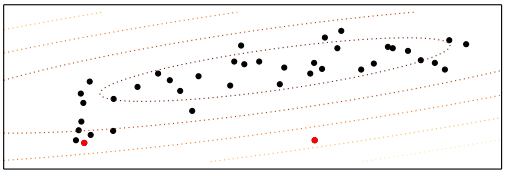
$$g(x) = f(x) + \mathbf{h}(x)^T\beta, \text{ with } f(x) \sim GP(0, k(x, x')) \quad (1)$$

where $f(x)$ is a zero mean GP, $\mathbf{h}(x)$ are a set of fixed basis functions, and β are additional parameters. For valence we used linear basis functions whereas quadratic basis functions were selected for arousal.

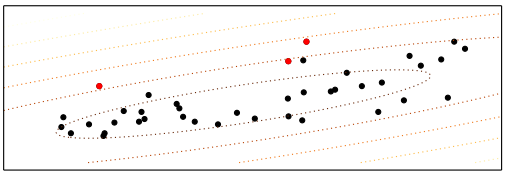
We used the squared exponential kernel for which, during interpolation at new values, distant observations will have negligible effect:

$$k(x, x') = \sigma_f^2 \times exp\left(\frac{-(x - x')^2}{2l^2}\right) + \sigma_n^2\delta(x, x') \quad (2)$$

where the length-scale l and the signal variance σ_f are hyperparameters, σ_n is the noise variance and $\delta(x, x')$ is the Kronecker delta. All the parameters are estimated using the maximum likelihood principle. In this work, σ_n values are not hyperparameters since they represent the known variance of annotations gathered in the controlled rating experiment described in section 3.3. They are added to the diagonal of the assumed training covariance. As a consequence, the GP is also able to model the subjectivity of emotions from this experiment.



(a) Valence



(b) Arousal

Fig. 3. Mahalanobis distances between the 40 video clips and the estimated center of mass, with respect to the estimated covariance in the ranking/rating space. Red points are the video clips considered as outliers.

5. OUTLIER DETECTION

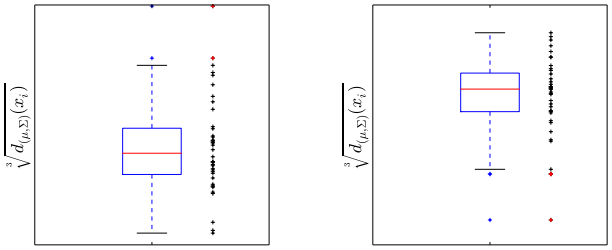
To perform a regression analysis on clean data, the first step is to detect outliers.

The Minimum Covariance Determinant (MCD) estimator introduced by Rousseeuw in [15] is a highly robust estimator for estimating the center and scatter of a high dimensional data set without being influenced by outliers. Assuming that the inlier data are Gaussian distributed, it consists in finding a subset of observations whose empirical covariance has the smallest determinant. The MCD estimate of location μ is then the average of the "pure" observations in the selected subset and the MCD estimate of scatter is their covariance matrix Σ. In this work, we used the fast MCD algorithm implemented in [16] to estimate the covariance of the 40 video clips defined in section 3.3, described in the 2D ranking/rating space by their rank and rating score.

Once the center and covariance matrix have been estimated, the Mahalanobis distance of centered observations can be computed. It provides a relative measure of an observation from the center of mass taking into account the correlation between those points. The Mahalanobis distance of a video clip x_i is defined as:

$$d_{(\mu,\Sigma)}(x_i)^2 = (x_i - \mu)\Sigma^{-1}(x_i - \mu) \quad (3)$$

where μ and Σ are the estimated center of mass and covari-



(a) Valence (b) Arousal

Fig. 4. Box plot of the Mahalanobis distances for valence and arousal. The whiskers show the lowest and highest values still within the 1,5 IQR. Red points are the video clips considered as outliers.

ance of the underlying Gaussian distribution. Figure 3 shows the shape of the Mahalanobis distances for the valence and arousal data sets.

By considering the covariance of the data and the scales of the different variables, the Mahalanobis distance is useful for detecting outliers in such cases. As a rule of thumb, a video clip x_i is considered as an outlier if $d_{(\mu,\Sigma)}(x_i) < Q1 - 1.5 \times IQR$ or if $d_{(\mu,\Sigma)}(x_i) > Q3 + 1.5 \times IQR$ with $Q1$ and $Q3$ the first and third quartiles and IQR the Inter-quartile Range. The boxplots showing the outliers detected for valence and arousal during this process are illustrated in Figure 4.

In our experiments, two video clips are categorized as outliers for valence and three video clips for arousal. Thus, these video clips are removed from the data sets in order to perform a regression analysis only on "clean" data sets. As a consequence, 38 video clips are used to perform the regression analysis for valence while for arousal the data set is composed of 37 video clips.

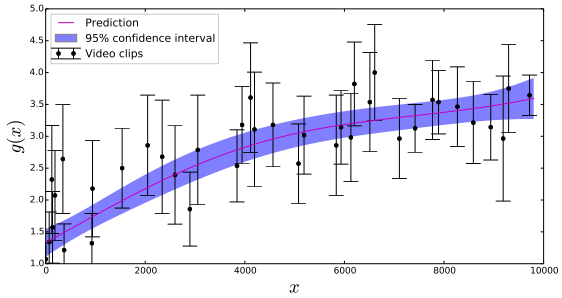
6. RESULTS

Figure 5 shows the regression models trained on all "clean" video clips for both valence and arousal axes. The 95% confidence interval shows that the models successfully used the variance of the annotations to learn the models.

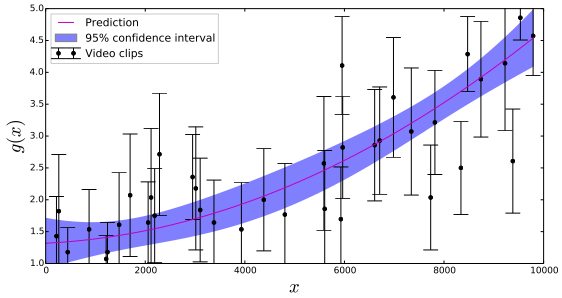
To measure the prediction power of the learned regression models, we calculated in addition to the well-known conventional squared correlation coefficient R^2, the predictive leave-one-out squared correlation coefficient Q^2_{loo} defined as:

$$Q^2_{loo} = 1 - \frac{\sum_{i=1}^{N} \left(y_i^{pred(N-1)} - y_i\right)^2}{\sum_{i=1}^{N} \left(y_i - y_{mean}^{N-1,i}\right)^2} \quad (4)$$

with y_i the true rating value of the video clip i and $y_i^{pred(N-1)}$ the prediction of the model learned with the initial training set from which the video clip i was removed. Note that the arithmetic mean used in equation (4), $y_{mean}^{N-1,i}$, is different for each test set and calculated for the observed values comprised in the training set.



(a) Valence



(b) Arousal

Fig. 5. Gaussian Process Models learned for valence and arousal converting ranks (horizontal axis) into ratings (vertical axis). Black bars show the variance of the annotations gathered in section 3.3.

Table 1. Performance of the Gaussian Process Models learned predicting valence and arousal.

Measure	Valence	Arousal
R^2	0.657	0.632
Q^2_{loo}	0.621	0.586

R^2 measures the goodness of fit of a model while Q^2_{loo} computed using the leave-one-out cross-validation technique measures the model prediction power. Both values for valence and arousal are shown in Table 1.

These results are remarkably high considering that we are modeling crowdsourced ranks and affective ratings that are both subject to the subjectivity of human emotions. Thus, our proposed regression models successfully learned to fit input observations. Furthermore, Q^2_{loo} values show that the models are also able to provide valid predictions for new observations.

7. CONCLUSION

In this work, we show that it is possible to estimate absolute values in the emotional space using affective ranks while taking into account the subjectivity of emotions.

First, we used the relative ranks available for the 9,800 video clips of the LIRIS-ACCEDE database as well as absolute scores available for a subset of the database. Outliers were detected using the minimum covariance determinant estimator and removed from the dataset in order to create a subset of "clean" observations. Finally, a regression analysis was performed for valence and arousal. The Gaussian process regression models, taking into account the variance of the annotation of the absolute scores, achieved a good performance, confirming our intuition that absolute scores in the affective space can be estimated using relative ranks.

In a near future, we plan to consider the continuous annotation process as a way to get finer annotations for longer video clips. The comparisons between these continuous annotations and absolute scores will allow to study the effect of the memorability on the induced emotions.

8. REFERENCES

[1] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, *International affective picture system (IAPS): Technical manual and affective ratings*. The Center for Research in Psychophysiology, University of Florida, 1999.

[2] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Transactions on Multimedia*, pp. 1–1, 2014.

[3] Y. Baveye, J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamaret, "A large video data base for computational models of induced emotion," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 2013, pp. 13–18.

[4] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4738, pp. 488–500.

[5] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, Nov. 2010.

[6] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: a database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[7] S. Ovadia, "Ratings and rankings: reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.

[8] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, Apr. 2013, pp. 1–8.

[9] Y.-H. Yang and H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 762–774, 2011.

[10] F. Wickelmaier and C. Schmid, "A matlab function to estimate choice model parameters from paired-comparison data," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 1, pp. 29–40, Feb. 2004.

[11] R. W. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Electronic Imaging'99*, 1998, pp. 290–301.

[12] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[13] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, Apr. 1970.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2006.

[15] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, Dec. 1984.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.