Discrete Choice Models for Non-Intrusive Quality Assessment

Petko N. Petkov[1], *W. Bastiaan Kleijn*[2], *Bert de Vries*[3]

[1] Sound and Image Processing laboratory, School of Electrical Engineering,
KTH-Royal Institute of Technology, Stockholm, Sweden
[2] School of Engineering and Computer Science, Victoria University of Wellington,
Wellington, New Zealand
[3] DSP Research, GN ReSound A/S, Eindhoven, Netherlands
petkov@kth.se, bastiaan.kleijn@ecs.vuw.ac.nz, bdevries@gnresound.com

Abstract

Non-intrusive signal quality assessment in general, and its application to speech signal processing, in particular, builds extensively upon statistical regression models. Commonly, the raw preference scores used for fitting these models belong to a categorical scale. Averaging the scores over a number of test subjects results in smooth, close-to-continuous ratings, thus justifying the use of regression as opposed to classification models. A form of marginalization, averaging subjective ratings takes away useful information about the reliability of individual test points. Using a model tailored to the raw data achieves highly competitive performance in terms of conventional performance measures while providing the additional advantage of identifying the usability of individual test points. In this paper, we consider the application of discrete choice models to non-intrusive quality assessment of speech.

Index Terms: non-intrusive quality assessment, discrete choice model, distribution fitting

1. Introduction

Model-based quality assessment (QA), also referred to as objective QA, has as goal the replacement of costly, time-consuming and, at times, infeasible subjective testing. Ideally, a model for QA would replicate the processing stages taking place in the human auditory periphery and the central nervous system. As little is known, however, about how preference ratings are formed in the brain, an algorithmic mapping from features to preference scores is used instead. The parameters of the mapping are inferred using large databases with speech utterances and corresponding subjective ratings.

With respect to the origin of the features a QA model is classified as intrusive if both the clean and the noisy versions of the signals are used [1], and as non-intrusive if only the noisy signals are used [2], [3]. In terms of established performance measures, intrusive models achieve, on average, higher performance. The dependence on the clean signals, however, precludes their use in on-line applications.

A number of different protocols for absolute preference elicitation are used in practice. While the scale is typically discrete, the resolution varies. Averaging the original ratings over the test subject dimension effectively produces a continuous scale where applying a regression model is the natural choice. Averaging, on the other hand, takes away information pertaining to the relevance of the individual utterances. We aim to exploit this information by fitting a model tailored to the discrete data.

Classification is the problem of choosing one out of several and finitely many, alternatives, e.g., [4]. Designing a good classifier can be a challenging task when the number of choice categories is high. Classifiers for application in QA need to take into account that preference ratings exhibit a large spread over the set of possible choices [5]. It is, therefore, imperative to use probabilistic models, i.e., systems that output the probabilities with which a given data point belongs to each category.

The data we work with [5], is collected using a subjective testing protocol based on five ordered categories, ranging from ”bad” to ”excellent”. This protocol is applied to the evaluation of signal quality after compression and transmission [6] and is extensively used for training QA models [2], [3]. The relatively low number of choice categories makes the design of a classifier an attractive approach. For higher resolution protocols, and depending on the amount of available training data, it may be more attractive to consider a regression model.

The architecture of the proposed system is based on a discrete choice model (DCM) [7]. In particular, we work with an ordered logit model, suitable within a maximum likelihood parameter estimation setting, and an ordered probit model, suitable within a Bayesian parameter estimation setting. Ordered DCMs, as the name suggests, perform classification when the choice categories are arranged in an order, which makes them suitable for application to QA. Fitting a DCM to the data, effectively fits the distribution of the discrete ratings for each training data point. As a result, within the limitations of the model architecture, a DCM can be used to predict both the mean and the variance of the test data. Experimental results reveal high performance both in terms of conventional performance measures such as correlation coefficient and root mean-square (RMS) error [2] as well as Kullback-Leibler (KL) divergence between the predictive and the empirical preference distributions.

This paper is organized as follows. Theory is presented in Section 2, considering both the ML and the Bayesian inference paradigms. The experimental setting, including the feature set and the data, and experimental results are presented in Section 3. We conclude with Section 4.

2. Theoretical Background

DCM for ordered data, such as the output of quality rating experiments, have a hierarchical structure consisting of a latent variable, called *utility*, and a set of decision thresholds, as illustrated in Figure 1, [7]. The utility, a scalar and probabilistic random variable, can be viewed as an internal representation of the perceived quality. As the utility is not observable, its char-

acteristics can only be inferred in view of the choices revealed by the data. The utility of utterance i is defined with a general probabilistic model of the form

$$u_i = \mathrm{s}\left(\theta_i\right) + \epsilon_i, \tag{1}$$

where θ are features extracted from the signal, $\mathrm{s}\left(.\right)$ is some appropriate function and ϵ is noise. The probability of observing a particular category choice for a given utterance, is obtained by integrating the utility within the thresholds identifying this category.

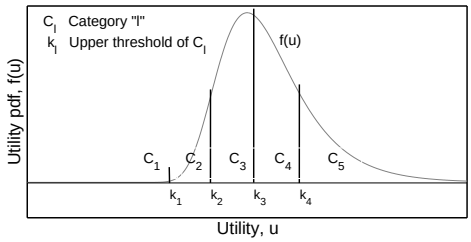

Figure 1: A five-category ordinal DCM.

For convenience, it is assumed that $\epsilon_i, \forall i$ are independent and identically distributed (i.i.d.). The part of the assumption concerning independence can easily be motivated and is used throughout the preference modeling literature [7]. The notion of identity of the distributions is the more restrictive part of the assumption as it implies that conditional on the parameters of $\mathrm{s}\left(.\right)$, the utility has a constant variance. This assumption appears reasonable for a formal test with selected utterances but need not hold in general. The advantages of tractable mathematical analysis, together with high performance results, however, favor the i.i.d. assumption.

With respect to the function $\mathrm{s}\left(.\right)$, we chose to work with a linear model of the form

$$\mathrm{s}\left(\theta_i\right) = \mathbf{w}^{\mathrm{T}}\theta_i, \tag{2}$$

where $\mathbf{w}$ is the vector of model coefficients. This choice is motivated by the simplified analysis of the resulting models. Next we review the specifics of the ordered ML and Bayesian DCMs.

2.1. Maximum Likelihood Set-up

In an ML set-up, the thresholds $\mathrm{k}_l, l \in \{1, 2, 3, 4\}$ and the coefficients $\mathbf{w}$ are assumed to be deterministic parameters. To obtain a closed-form expression for the likelihood of each decision we choose to work with a Gumbel distribution, a particular case of the generalized extreme value distribution, for ϵ, effectively deriving an ordered logit model.

By introducing a constant feature into the feature set we can assume, without loss of generality, that ϵ is zero-mean. As a result, the probability and cumulative density functions of the utility, for a given utterance, are obtained of the form:

$$f_G(u_i) = \frac{e^{-\frac{u_i - \mathbf{w}^{\mathrm{T}}\theta_i + \beta\gamma}{\beta}} e^{-e^{-\frac{u_i - \mathbf{w}^{\mathrm{T}}\theta_i + \beta\gamma}{\beta}}}}{\beta} \tag{3}$$

$$F_G(u_i) = e^{-e^{-\frac{u_i - \mathbf{w}^{\mathrm{T}}\theta_i + \beta\gamma}{\beta}}}, \tag{4}$$

where the parameter β determines the variance of the utility as $\frac{\pi^2}{6}\beta^2$ and γ is the Euler-Mascheroni constant. Note that since we already assumed that the variance of the utility is constant, it is acceptable to associate β with any feasible numerical value. The model is scale invariant as scaling the variance is compensated for by scaling the separation among the thresholds.

Let $\mathrm{p}\left(\mathrm{C}_j|\theta_i, \mathbf{w}, \beta, \mathbf{k}\right)$ denote the probability of choosing category C_j for utterance i given the model. Then

$$\mathrm{p}\left(\mathrm{C}_j|\theta_i, \mathbf{w}, \beta, \mathbf{k}\right) = F_G(\mathrm{k}_j|\theta_i, \mathbf{w}, \beta) - F_G(\mathrm{k}_{j-1}|\theta_i, \mathbf{w}, \beta) \tag{5}$$

and the log-likelihood of the observed data becomes

$$\mathrm{LL} = \sum_{i=1}^{N}\sum_{j=1}^{J} M_{ij} \log\left\{\mathrm{p}\left(\mathrm{C}_j|\theta_i, \mathbf{w}, \beta, \mathbf{k}\right)\right\}, \tag{6}$$

where N is the number of utterances, $J = 5$ is the number of choice categories and M_{ij} is the number of times that category C_j was selected for utterance i. The optimization problem of training the model is finally formulated as

$$\begin{aligned} \mathrm{argmax}_{\mathbf{k},\,\mathbf{w}}\mathrm{LL} \quad & \mathrm{s.t.} \\ \mathrm{k}_1 - \mathrm{k}_2 \quad &\leq \quad 0 \\ \mathrm{k}_2 - \mathrm{k}_3 \quad &\leq \quad 0 \\ \mathrm{k}_3 - \mathrm{k}_4 \quad &\leq \quad 0, \end{aligned} \tag{7}$$

where the constraints impose the ordering of the thresholds. The gradient and the hessian of the above objective function, with respect to the model parameters are readily derived in closed-form. We note that while the objective is not convex, experimental results obtained with feasible initial values are consistently good. Lack of convexity, however, precludes feature selection from a larger set of features based on L_1 norm constraint [8]. This limits the use of the ML-based model to a preselected feature set.

2.2. Bayesian Set-up

In a Bayesian set-up, all the unknown model parameters are assumed to be random variables. This allows for taking into account the uncertainty about their values in view of the data. For the Bayesian ordered DCM, we assume $\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$. The reason lies with the computational advantage achieved by introducing an augmented likelihood function [9]. While it is not feasible to derive a closed-form solution, efficient sampling from the conditional posteriors of the unknown model parameters is possible due to the augmentation procedure [9], [10].

Before describing the model we introduce the following notation. Let the vector $\bar{\mathbf{C}}$ contain all the category choices that are observed in the training data. As there is a fixed number R of ratings for each utterance, where $R = 24$ in [5], we have $\dim\left(\bar{\mathbf{C}}\right) = RN$. The augmented likelihood includes explicitly the utility. We use $\bar{\mathbf{u}}$ to denote the vector of utilities that correspond to $\bar{\mathbf{C}}$. Thus for the category choice in $\bar{\mathbf{C}}_m$ there is an utility in $\bar{\mathbf{u}}_m$ that lies in the range between the thresholds of this category.

Using the above notation leads to the following expression for the joint posterior distribution:

$$\begin{aligned} \mathrm{p}\left(\mathbf{w}, \mathbf{k}, \bar{\mathbf{u}}|\bar{\mathbf{C}}\right) \quad &\propto \quad \mathrm{p}\left(\bar{\mathbf{u}}, \bar{\mathbf{C}}|\mathbf{w}, \mathbf{k}\right)\mathrm{p}\left(\mathbf{w}, \mathbf{k}\right) \\ &\propto \quad \mathrm{p}\left(\bar{\mathbf{C}}|\bar{\mathbf{u}}, \mathbf{w}, \mathbf{k}\right)\mathrm{p}\left(\bar{\mathbf{u}}|\mathbf{w}, \mathbf{k}\right) \cdot \\ &\qquad \mathrm{p}\left(\mathbf{w}\right)\mathrm{p}\left(\mathbf{k}\right), \end{aligned} \tag{8}$$

where the assumption of constant $\mathrm{p}\left(\bar{\mathbf{C}}\right)$ is used broadly in the framework of Bayesian inference [4]. The expression in (8) reflects the assumption of prior independence between the set

of the model coefficients and the set of the category thresholds. From the model definition it follows that:

$$\mathrm{p}\left(\bar{\mathbf{C}}|\bar{\mathbf{u}},\mathbf{w},\mathbf{k}\right) = \prod_{m=1}^{\dim(\bar{\mathbf{C}})} \mathrm{I}\left(\mathrm{k}_{\bar{\mathbf{C}}_m-1} \leq \bar{\mathbf{u}}_m \leq \mathrm{k}_{\bar{\mathbf{C}}_m}\right) \quad (9)$$

$$\mathrm{p}\left(\bar{\mathbf{u}}|\mathbf{w},\mathbf{k}\right) = \prod_{m=1}^{\dim(\bar{\mathbf{C}})} N\left(\mathbf{w}^{\mathrm{T}}\theta_m, \sigma_\epsilon^2\right), \quad (10)$$

where $\mathrm{k}_{\bar{\mathbf{C}}_m-1}$ and $\mathrm{k}_{\bar{\mathbf{C}}_m}$ represent the lower and the upper thresholds of the category choice in $\bar{\mathbf{C}}_m$, and the indicator function $\mathrm{I}(.)$ equals one when the condition inside the brackets is satisfied and zero otherwise. The above expressions illustrate the underlying assumption of independence among the individual rating tasks. As the likelihood function is defined in terms of the Normal density, conjugate priors can be defined for the unknown model parameters:

$$\begin{aligned} \mathrm{p}\left(\mathbf{w}\right) &= N_L\left(\mathbf{w}_0, \Sigma_0\right) & (11)\\ \mathrm{p}\left(\mathbf{k}\right) &\propto N_{J-1}\left(\mathbf{k}_0, \Lambda_0\right)\mathrm{I}\left(\mathrm{k}_1 \leq \mathrm{k}_2 \leq \mathrm{k}_3 \leq \mathrm{k}_4\right), & (12)\end{aligned}$$

where L is the dimensionality of the feature space and J was defined as the number of categories. A proportionality sign was used in (12) instead of equality together with the appropriate scaling factor to simplify the expression. The role of the indicator function in (12) is to enforce the constraint that the thresholds are ordered.

The use of conjugate priors leads to conditional posteriors in closed-form. The a-posteriori utility becomes

$$\begin{aligned} \mathrm{p}\left(\bar{\mathbf{u}}_m|\,\mathbf{w},\mathbf{k},\bar{\mathbf{C}}_m\right) &\propto N\left(\mathbf{w}^{\mathrm{T}}\theta_m, \sigma_\epsilon^2\right)\cdot \\ &\quad \mathrm{I}\left(\mathrm{k}_{\bar{\mathbf{C}}_m-1} \leq \bar{\mathbf{u}}_m \leq \mathrm{k}_{\bar{\mathbf{C}}_m}\right), & (13)\end{aligned}$$

where the constant of proportionality is trivial to obtain but not needed for sampling [10]. For the thresholds, assuming prior independence, it can be shown that

$$\begin{aligned} \mathrm{p}\left(\mathrm{k}_l|\mathbf{w},\mathbf{k}_{\neq l},\bar{\mathbf{u}},\bar{\mathbf{C}}\right) &\propto N\left(\mathrm{k}_{0,l},\Lambda_{0,ll}\right)\mathrm{I}\left(\mathrm{k}_l^{\inf} \leq \mathrm{k}_l \leq \mathrm{k}_l^{\sup}\right)\\ \mathrm{k}_l^{\inf} &= \max\left(\mathrm{k}_{l-1}, \max\left(\mathrm{u}_{(l)}\right)\right)\\ \mathrm{k}_l^{\sup} &= \min\left(\mathrm{k}_{l+1}, \min\left(\mathrm{u}_{(l+1)}\right)\right), \quad (14)\end{aligned}$$

where $\max\left(\mathrm{u}_{(l)}\right)$ and $\min\left(\mathrm{u}_{(l+1)}\right)$ denote the largest utility value drawn from category l and the smallest utility value drawn from category $l+1$. The indicator function uses the two scalars $\mathrm{k}_l^{\inf}$ and $\mathrm{k}_l^{\sup}$, easily obtained conditional on the utility, to constrain each threshold to its permissible range.

Finally, the conditional posterior for the model coefficients is obtained of the form

$$\begin{aligned} \mathrm{p}\left(\mathbf{w}|\mathbf{k},\bar{\mathbf{u}},\bar{\mathbf{C}}\right) &= N\left(\mu_{\mathbf{w}},\Sigma_{\mathbf{w}}\right)\\ \mu_{\mathbf{w}} &= \left(\frac{1}{\sigma_\epsilon^2}\bar{\mathbf{u}}^{\mathrm{T}}\boldsymbol{\Theta} + \mathbf{w}_0^{\mathrm{T}}\Sigma_0^{-1}\right)\left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta} + \Sigma_0^{-1}\right)^{-1}\\ \Sigma_w &= \left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{\Theta} + \Sigma_0^{-1}\right)^{-1}, \quad (15)\end{aligned}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{RN\times L}$ is the matrix in which each row contains the features associated with an individual rating task.

Expressions (13), (14) and (15) are all based on the Normal distribution and facilitate Gibbs sampling [10]. The algorithm loops through the individual posteriors, drawing samples by conditioning always on the latest values of the parameters. After an initial burn-in period, the Gibbs sampler converges to the true posterior of the model parameters and produces dependent draws from these distributions. The draws can be used to evaluate the predictive distribution over the choice categories for a previously unseen data point, through stochastic integration, with selectively high precision. In practice, a limited number of independent draws is sufficient for a good approximation.

3. Experiments

The experimental set-up and the validation results are presented in this section. The feature set and the database used in the experiments are described in Section 3.1. The predictive performance of the models, measured in terms of correlation, RMS error and KL divergence, is illustrated in Section 3.2.

3.1. Experimental Set-up

The feature set used in these experiments was proposed in [11]. A subset of band-based modulation spectrum (MS) features was combined with a subset of the source and vocal tract descriptors from [2]. The modulation spectrum is broadly used for feature extraction in QA as it allows partial separation of the characteristics of the speech and the noise signals. We extracted the raw MS features on a per-frame basis and averaged along the time dimension for active frames only. The resulting features are global, i.e., they are representative of the whole utterance. The final feature set contains forty-seven features and was selected from the larger original set through side experiments.

We note that while preliminary feature selection is important to reduce overfitting, it mostly affects the ML case. In the Bayesian case, uncertainty for the model parameters is accounted for by the model and the risk for overfitting is reduced in a natural way. Apart from adding new features, it is possible to further increase the flexibility of the model, and correspondingly its performance, by using a polynomially extended version of the base feature set. This approach, which can, e.g., emphasize rare occurrences, was considered in [11]. We limit our analysis to a second order extension.

The database used for validating the models, [5], consists of seven data sets prepared by four different laboratories. Four languages and a large number of distortion conditions, including but not limited to additive noise and coding artifacts are represented in the data. The motivation to work with this database is twofold: i) it is freely available facilitating validation and comparison among proposed methods and ii) it is highly heterogenous and, as such, a natural choice for validation of QA algorithms. The total number of utterances is 1328. Models are trained on six data sets at a time and validated on the seventh.

The following settings were used for the Gibbs sampler in the Bayesian implementation. The burn-in was set to 20000 draws from each conditional posterior. The number of collected samples, used to evaluate the statistics of interest was 5000. A separation of five samples between two collected samples was imposed to reduce dependencies. Arbitrary but feasible values were chosen for algorithm initialization. One complete loop of the Gibbs sampler effectively produced one model realization. The predictive distributions from each of these models for a given test point were averaged to obtain a single distribution, which was then used for performance evaluation.

3.2. Validation Results

The model performance in terms of per-condition Pearson correlation coefficient ρ_{pc} and RMS error r_{pc} between the mean

Table 1: Per condition (pc) Pearson correlation and RMS .

data set	P.563		ML DCM		Bayesian DCM	
	ρ_{pc}	r_{pc}	ρ_{pc}	r_{pc}	ρ_{pc}	r_{pc}
BNR-X3	0.911	0.345	0.935	0.298	0.945	0.268
CNET-X3	0.888	0.378	0.840	0.377	0.873	0.340
CSELT-X3	0.798	0.398	0.851	0.482	0.851	0.498
NTT-X3	0.902	0.329	0.882	0.322	0.883	0.323
BNR-X1	0.867	0.349	0.901	0.355	0.912	0.324
CNET-X1	0.843	0.457	0.887	0.368	0.910	0.321
NTT-X1	0.923	0.270	0.901	0.291	0.924	0.260
mean	**0.876**	**0.361**	**0.885**	**0.356**	**0.900**	**0.333**

Table 2: Average entropy and KL divergence in bits.

data set	Entropy	KL div., ML	KL div., Bayes
BNR-X3	1.738	0.383	0.300
CNET-X3	1.501	0.667	0.518
CSELT-X3	1.483	0.814	0.997
NTT-X3	1.480	0.430	0.416
BNR-X1	1.527	0.402	0.301
CNET-X1	1.447	0.446	0.352
NTT-X1	1.510	0.328	0.297
mean	**1.527**	**0.496**	**0.454**

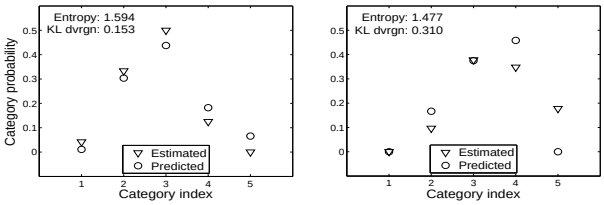

Figure 2: Predictive and empirical distributions.

of the predictive and the mean of the empirical (computed from $R = 24$ preference ratings per utterance) distributions are presented in Table 1. These measures, well-established for the performance evaluation of QA models [2], [3], were computed by using the numerical representation $\{1, 2, 3, 4, 5\}$ of the choice categories. A third order monotonic polynomial mapping was applied to the objective mean estimates before computing correlation and RMS error. The motivation for using such a mapping can be found, e.g., in [2]. To evaluate how close the predictive distribution is to the estimated distribution, we used the KL divergence. The corresponding results are presented in Table 2.

Performance in terms of correlation and RMS error illustrates that the proposed model is capable of learning complex dependencies from a relatively small number of data points. A comparison with results obtained with the model from [2], which is the current ITU-T standard for non-intrusive quality assessment, also favors the proposed approach. A comparison between the ML and the Bayesian implementations illustrates the power of the Bayesian learning paradigm. Taking into consideration the uncertainty of model parameters reduces over-fitting to the training data and results in significantly higher performance for the Bayesian model.

The entropy of the subjective distribution and the KL divergence between the subjective and the predictive distributions, averaged over all test points in the test data set provide the intuition for assessing the model performance. In Figure 2 we present the empirical and the predictive distributions for two randomly chosen utterances from the NTT-X1 data set. These results were produced with the ML model. The entropy of the empirical distribution and the KL divergence are also indicated.

The experimental results suggest that both models produce an informative fit to the distribution of the subjective data. Interestingly, this is achieved with the limited flexibility provided by optimally positioning the category thresholds. Even more interesting, in this respect, is the result that the thresholds are approximately equidistantly spaced. Similarly to the case with the other two performance measures, the Bayesian model results in better performance on average.

4. Conclusions

Discrete choice models are a natural candidate for application to QA as they are well-suited to the original data. The proposed models learn efficiently from a limited amount of training data, testifying to the appropriateness of the choice of the feature set and the statistical inference approach. Fitting the individual preference scores, rather than the average values, allows modeling of the distribution of these scores and makes it possible to predict the reliability of individual test points subject to the constraint of constant utility variance. The almost equidistant spacing among the thresholds, likely related to the constant utility variance, suggests the presence of redundancy in the model definitions. This redundancy can probably be utilized to reduce the complexity of the models.

5. Acknowledgements

The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 256230

6. References

[1] ITU-T Rec. P.862, “Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs,” 2001.

[2] ITU-T Rec. P.563, “Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications,” 2004.

[3] D. S. Kim and A. Tarraf, “ANIQUE+: A New American National Standard for Non-intrusive Estimation of Narrowband Speech Quality,” *Bell Labs Technical Journal*, vol. 12, pp. 221–236, 2007.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[5] ITU-T Rec. P.Sup23, “ITU-T Coded-Speech database,” 1998.

[6] ITU-T Rec. P.800, “Methods for Subjective Determination of Transmission Quality,” 1996.

[7] K. E. Train, *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least Angle Regression,” *The Annals of Statistics Journal*, vol. 32, pp. 407–499, 2004.

[9] J. H. Albert and S. Chib, “Bayesian Analysis of Binary and Polytomous Response Data,” *Journal of the American Statistical Society*, vol. 88, pp. 669–679, 1993.

[10] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith, *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley & Sons, Ltd, Chichester, 2002.

[11] P. N. Petkov and W. B. Kleijn, “Probabilistic Non-Intrusive Quality Assessment of Speech for Bounded-Scale Preference Scores,” in *In proc. Quality of Multimedia Experience*, 2010, pp. 188–193.