Towards Predicting Expressed Emotion in Music from Pairwise Comparisons

Jens Madsen, Bjørn Sand Jensen, Jan Larsen and Jens Brehm Nielsen
Technical University of Denmark,
Department of Informatics and Mathematical Modeling,
Asmussens Allé B321
2800 Kongens Lyngby, Denmark
{jenma,bjje,jl,jenb}@imm.dtu.dk

ABSTRACT

We introduce five regression models for the modeling of expressed emotion in music using data obtained in a two alternative forced choice listening experiment. The predictive performance of the proposed models is compared using learning curves, showing that all models converge to produce a similar classification error. The predictive ranking of the models is compared using Kendall's τ rank correlation coefficient which shows a difference despite similar classification error. The variation in predictions across subjects and the difference in ranking is investigated visually in the arousal-valence space and quantified using Kendall's τ.

1. INTRODUCTION

The possibility to recommend music which express a certain mood or emotion has recently gathered increasing attention within the Music Information Retrieval (MIR) community.

Typically the recommendation is approached using computational methods, where music is represented using structural features, such as features based on the audio signal that mimic some functions of the human auditory perceptive system, and possibly features representing even higher aspects of the human cognitive system. Research is ongoing in finding what features can capture aspects in the music that express or induce emotions see e.g. [1]. Furthermore, it is well known that there is a clear connection between lyrics and the audio in music [2] and lyrical features have equally been shown to produce good results [3]. Even contextual information about music can be utilized for the prediction of emotions in music using social media contents [4].

Despite the many meaningful audio features and representations, most computational models are supervised and rely on human participants to rate a given excerpt. These ratings are mapped using supervised machine learning approaches under the assumption that the model is the same for all musical excerpts, thus the projection into feature space is based on the same model for all excerpts and typically also for all participants. Instead of obtaining decisions from subjects, unsupervised methods have recently been proposed which can be used to find emotional categories of excerpts [5]. The decision of what machine learning method to apply is tightly connected to the chosen music representation and what emotional representation [6], and in this work we consider the supervised setting.

Expressed emotions in music are typically rated based on simple self-reporting listening experiments [7] where the scales are adapted to quantify for example the categorical [8] or dimensional [9] models of emotion. Although there is not one simple way of doing this and numerous different approaches have been made to obtain these ratings e.g. using majority ruling, averaging across ratings, etc. in both domains even using combinations of the emotional models [10]. Another aspect to take into account when creating computational models of emotion, is that it is well known that emotional expression in music changes over time which could further refine a recommendation method. Two main direction has been followed in obtaining time depend ratings. The first is based on post ratings of excerpts in the 15-30 s range under the assumption that within this frame the emotional expression is approximately constant. Machine learning techniques can then be used to create models making predictions on a smaller time scale using the post ratings of larger excerpts [11]. The other direction is to continuously measure expressed emotions in music directly in e.g the arousal and valence space (AV space) [12] and subsequently model this.

In [13] we proposed an alternative way of quantifying the expressed emotion in music on the dimensions of *valence* and *arousal* by introducing a two alternative force choice (2AFC) post rating experimental paradigm. Given the relative nature of pairwise comparisons they eliminate the need for an absolute reference anchor, which can be a problem in direct scaling experiments. Furthermore the relative nature persist the relation to previous excerpts reducing memory effects. We use 15 s excerpts to minimize any change in expressed emotion over time, and large enough not to cause mental strain on subjects. We proposed a probabilistic Gaussian process framework for mapping the extracted audio features into latent subspaces that is learned by the comparisons made by participants of musical excerpts evaluated on the dimensions of valence and arousal. The underlying assumption is that given the features, the

projection made by the model mimic the cognitive decision making by participants in making the pairwise comparison. We investigated how many comparisons are needed per excerpt to reach acceptable level of performance by obtaining all possible unique comparisons for 20 excerpts and furthermore to investigate the individual subjective differences. In [14] they proposed a greedy algorithmic approach converting pairwise comparisons into a ranking of excerpts and modeling this using a RBF-ListNet algorithm. They focused on the case of few comparisons for many excerpts, using comparisons from multiple participants aggregating to one large dataset, neglecting the individual differences between subjects. On the other hand, our results showed a great difference between participants which the framework and approach accounts for along with noise on the pairwise judgments.

These individual differences are further investigated in this paper using the well known arousal and valance scores in a 2D space. Furthermore, we introduce five models for the modeling of the pairwise comparisons, where an extension to the existing framework is made using linear and squared exponential kernels. Moreover, we compare the Gaussian process model to three versions of a Generalized Linear Model (GLM) namely the standard version and two regularized versions using L1 and L2 norms. Learning curves are computed as a function of the misclassification error and the number of (randomly chosen) pairwise comparisons in order to elucidate the difference between the five models. The differences between models and the resulting ranking of excerpts is further illustrated using Kendall's τ rank correlation learning curves.

2. EXPERIMENT & DATA

2.1 Experiment

A listening experiment was conducted to obtain pairwise comparisons of expressed emotion in music using a 2AFC experimental paradigm. 20 different 15 second excerpts were chosen from the USPOP2002 [1] dataset, so that, 5 excerpts were chosen to be in each quadrant of the AV space. The selection was performed by a linear regression model developed in previous work. A subjective evaluation was performed to verify that the emotional expression of each excerpt was as constant as possible.

A sound booth provided neutral surroundings for the experiment and the excerpts were played back using headphones to the 8 participants (2 female, 6 male). Written and verbal instructions were given prior to each session to ensure that subjects understood the purpose of the experiment and to ensure that each subject were familiar with the two emotional dimensions (valence and arousal). Each participant compared all 190 possible unique combinations. For the arousal dimension, participants were asked the question *Which sound clip was the most excited, active, awake?* For the valence dimension the question was *Which sound clip was the most positive, glad, happy?*. The two dimensions was rated individually and the presentation of the 190 paired excepts was randomized. The details of the experiment is available in [15].

[1] http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

No.	Song name
1	311 - T and p combo
2	A-Ha - Living a boys adventure
3	Abba - Thats me
4	Acdc - What do you do for money honey
5	Aaliyah - The one i gave my heart to
6	Aerosmith - Mother popcorn
7	Alanis Morissette - These r the thoughts
8	Alice Cooper - Im your gun
9	Alice in Chains - Killer is me
10	Aretha Franklin - A change
11	Moby - Everloving
12	Rammstein - Feuer frei
13	Santana - Maria caracoles
14	Stevie Wonder - Another star
15	Tool - Hooker with a pen..
16	Toto - We made it
17	Tricky - Your name
18	U2 - Babyface
19	UB40 - Version girl
20	ZZ top - Hot blue and righteous

Table 1. List of songs/excerpts.

2.2 Audio Representation & Features

In order to represent the 15 second excerpts in later mathematical models, each excerpt is represented by standard audio features, namely Mel-frequency cepstral coefficients (MFCC) (30 dimensional), that describes the log transformed short-term power spectrum of the musical signal. Furthermore a total of 9 features are included namely spectral-flux, roll-off, slope and variation and 5 features describing the temporal music signal including zero crossing rate and statistical shape descriptors.

These features are extracted using the YAAFE toolbox [1] for 512 sample frames with 50% overlap, thus for each excerpt we obtain a 39x1292 feature matrix $\mathbf{X}$. We create a vector representation by first standardizing the features and then estimating the mean, $\mu(\cdot)$ and the variance of the matrix $var(\cdot)$ over the frames and then applying the following vectorization, $\mathbf{x} = [\mu(\mathbf{X}), \mathrm{var}(\mathbf{X})]$. This (row) vector representation can directly be used in standard modeling tools and serves as a common ground for comparisons.

3. MODELS FOR PAIRWISE COMPARISONS

The pairwise observations presented in Section 2 poses a special challenge since each output now depends on two inputs and standard regression and classification tools do not immediately apply since they are typically formulated in a one to one relationship between inputs and outputs. The modeling aspect will thus necessarily play an integral part of this section, and we will initially outline the general framework.

[1] http://yaafe.sourceforge.net/

The audio excerpts presented in Section 2 are assembled in the set $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ with $n = 20$ distinct excerpts, each described by the feature input vector $\mathbf{x}_i$. For each of the test subjects the dataset comprises of all unique $m = 190$ combinations of pairwise comparisons between any two distinct excerpts, u and v, where $\mathbf{x}_u \in \mathcal{X}$ and $\mathbf{x}_v \in \mathcal{X}$. Formally, we denote the output set as

$$\mathcal{Y} = \{(y_k; u_k, v_k) | k = 1, ..., m\},$$

where $y_k \in \{-1, 1\}$ indicates which of the two excerpts that had the highest valence or arousal. $y_k = -1$ means that the u_k'th excerpt is picked over the v_k'th and visa versa when $y_k = 1$.

The main assumption in our setup is that the pairwise choice, y_k, between the two distinct excerpts, u and v, can be modeled as a function of the difference between two functional values, $f(\mathbf{x}_u)$ and $f(\mathbf{x}_v)$. The function $f : \mathcal{X} \rightarrow \mathbb{R}$ hereby defines an internal, but latent absolute reference of e.g. valence or arousal as a function of the excerpt represented by the audio features.

In order to model noise on the decision process we consider the logistic likelihood of the functional difference. The likelihood of observing a discrete choice thus becomes:

$$p(y_k | \mathbf{f}_k) \equiv \frac{1}{1 + \mathrm{e}^{-y_k (f(\mathbf{x}_{u_k}) - f(\mathbf{x}_{v_k}))}}, \qquad (1)$$

where $\mathbf{f}_k = [f(\mathbf{x}_{u_k}), f(\mathbf{x}_{v_k})]^T$. The remaining question is how the function is modeled and how we in turn regard the problem as a special regression problem. In the following we consider two different frameworks, namely Generalized Linear Models (GLM) and a flexible Bayesian non-parametric approach based on the Gaussian process (GP). In all cases we assume that the likelihood factorizes over the observations i.e., $p(\mathcal{Y}|\mathbf{f}) = \prod_{k=1}^{m} p(y_k|\mathbf{f}_k)$.

3.1 Generalized Linear Models

Generalized Linear Models are powerful and widely used extensions of standard least squares regression which can accommodate many types of observed variables and noise models. The canonical example in this family is indeed logistic regression, and here we extend the treatment to the pairwise case. The underlying model is a linear and parametric model of the form $\mathbf{f}_i = \mathbf{x}_i \mathbf{w}^\top$, where $\mathbf{x}_i$ may be extended in a different basis but the base model is still linear in $\mathbf{w}$.

If we now consider the likelihood defined in Eq. (1) and reasonably assume that the model, i.e. $\mathbf{w}$, is the same for the first and second input i.e. $\mathbf{x}_{u_k}$ and $\mathbf{x}_{v_k}$. Which results in a projection from the audio features $\mathbf{x}$ into the cognitive dimensions of valence and arousal given by $\mathbf{w}$ which is the same for all excerpts. We can then write

$$p(y_k | \mathbf{w}, \mathbf{x}_{u_k}, \mathbf{x}_{v_k}) = \frac{1}{1 + \mathrm{e}^{-y_k \left((\mathbf{x}_{u_k} - \mathbf{x}_{v_k}) \mathbf{w}^\top \right)}}. \qquad (2)$$

The resulting cost function, $\psi(\cdot)$, is given by the log likelihood

$$\psi_{GLM}(\mathbf{w}) = \sum_{k=1}^{m} \log p\left(y_k | \mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \mathbf{w}\right).$$

Thus, the problem reduces to a standard logistic regression problem only working on the difference in input space as opposed to the standard absolute input. This means that standard optimization techniques can be used to find the maximum likelihood solution, such as Iterated Reweighed Least Squares (IRLS) or other more general non-linear optimization method.

3.1.1 Regularized Extensions

The basic GLM formulation in Eq. (2) does work quite well for many problems, however has a tendency to become unstable with very few pairwise comparisons. We therefore suggest to regularize the basic GLM cost with L1 and L2 which are of course similar to standard regularized logistic regression (see [16]). The L2 regularized cost is as usual given by

$$\psi_{GLM-L2}(\mathbf{w}) = \sum_{k=1}^{m} \log p\left(y_k | \mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \mathbf{w}\right) - \lambda \left\|\mathbf{w}\right\|_2^2,$$

where the regularization parameter λ is to be found by cross-validation. This cost is still continuous and is solved with a standard Newton method. The L1 regularized cost is

$$\psi_{GLM-L1}(\mathbf{w}) = \sum_{k=1}^{m} \log p\left(y_k | \mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \mathbf{w}\right) - \lambda \left\|\mathbf{w}\right\|_1.$$

This discontinuous cost function (in $w_i = 0$) is solved using the active set method presented in [17]. The L1 regularization effectively results in a sparse model where certain features are potentially switched off. We will not interpret this property in detail but simply use the models as a reference.

3.2 Gaussian Process Framework

The GLM framework represents the simplest - but often effective - models for many regression and classification problems. An obvious extension is to treat the problem and the likelihood in a Bayesian setting which is presented in this section and further adhere to a non-parametric principle in which we model the $\mathbf{f}$ directly such that the posterior over $\mathbf{f}$'s can be written

$$p(\mathbf{f}|\mathcal{Y}, \mathcal{X}) = p(\mathcal{Y}|\mathbf{f}) \, p(\mathbf{f}|\mathcal{X}) / p(\mathcal{Y}|\mathcal{X}). \qquad (3)$$

While many relevant priors, $p(\mathbf{f}|\mathcal{X})$, may be applied we will consider a specific prior, namely a Gaussian Process (GP) prior. A GP is typically defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [18]. By $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$ we denote that the function $f(\mathbf{x})$ is modeled by a zero-mean GP with covariance function $k(\mathbf{x}, \mathbf{x}')$. The consequence of this formulation is that the GP can be considered a distribution over functions, i.e., $p(\mathbf{f}|\mathcal{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Bayes relation leads directly to the posterior distribution over $\mathbf{f}$, which is not analytically tractable. Instead, we use the *Laplace Approximation* to approximate the posterior

with a multivariate Gaussian distribution [2]. The GP was first considered with a pairwise, Probit based likelihood in [20], whereas we consider the logistic likelihood function.

3.2.1 Predictions

To predict the pairwise choice y_t on an unseen comparison between excerpts r and s, where $\mathbf{x}_r, \mathbf{x}_s \in \mathcal{X}$, we first consider the predictive distribution of $f(\mathbf{x}_r)$ and $f(\mathbf{x}_s)$ which is given as $p(\mathbf{f}_t|\mathcal{Y}, \mathcal{X}) = \int p(\mathbf{f}_t|\mathbf{f})\, p(\mathbf{f}|\mathcal{Y}, \mathcal{X})\, d\mathbf{f}$, and with the posterior approximated with the Gaussian from the Laplace approximation then $p(\mathbf{f}_t|\mathcal{Y}, \mathcal{X})$ will also be Gaussian given by $\mathcal{N}(\mathbf{f}_t|\boldsymbol{\mu}^*, \mathbf{K}^*)$ where $\boldsymbol{\mu}^* = \mathbf{k}_t^T \mathbf{K}^{-1} \hat{\mathbf{f}}$ and $\mathbf{K}^* = \mathbf{K}_t - \mathbf{k}_t^T (\mathbf{I} + \mathbf{WK})\, \mathbf{k}_t$, where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from the Laplace approximation (see [19]) and $\mathbf{k}_t$ is a matrix with elements $[\mathbf{k}_t]_{i,2} = k(\mathbf{x}_i, \mathbf{x}_s)$ and $[\mathbf{k}_t]_{i,1} = k(\mathbf{x}_i, \mathbf{x}_r)$ with $\mathbf{x}_i$ being a training input.

In this paper we are only interested in the binary choice y_t, which is determined by which of $f(\mathbf{x}_r)$ or $f(\mathbf{x}_s)$ that dominates [3].

3.2.2 Covariance Functions

The zero-mean GP is fully defined by the covariance function, $k(\mathbf{x}, \mathbf{x}')$. In the emotion dataset each input instance is an excerpt described by the vector $\mathbf{x}$ representing the mean and variance of the audio features. A standard covariance function for this type of input is the squared exponential (SE) covariance function defined as $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{\sigma_l^2} \|\mathbf{x} - \mathbf{x}'\|_2^2\right)$, where σ_f is a variance term and σ_l is the length scale, in effect defining the scale of the correlation in the input space. As a reference we also consider the linear covariance function given as $k(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}'\mathbf{x}^\top + 1\right)/\sigma^2$.

3.2.3 Hyper-parameters

An advantage of the Bayesian approach is that the hyper parameters may be found in a principled way namely by evidence maximization or maximum likelihood II estimation. The hyper-parameters collected in θ can thus be found by $\hat{\theta} = \arg\max_\theta \int p(\mathcal{Y}|\mathbf{f})\, p(\mathbf{f}|\theta) d\mathbf{f}$.

There is therefore in principle no need to use cross-validation to find the parameters. As with the posterior over f, the evidence also requires an approximation and we reuse the Laplace approximation to obtain the hyper-parameter estimate. We furthermore allow for a regularizing prior on the hyper-parameters which is similar in spirit to the regularized Expectation Maximization (EM) algorithm.

3.3 Alternative Models

The two modeling frameworks considered above are not the only options for modeling the pairwise relations. An obvious intermediate model is the GLM put in a Bayesian setting with (hierarchical) (sparsity) priors on $\mathbf{w}$ which we consider an intermediate step towards the full non-parametric GP model. Also Neural Networks can easily be adapted to handle the pairwise situation, such as [21]; however, the GP will again provide a even more flexible and principled model.

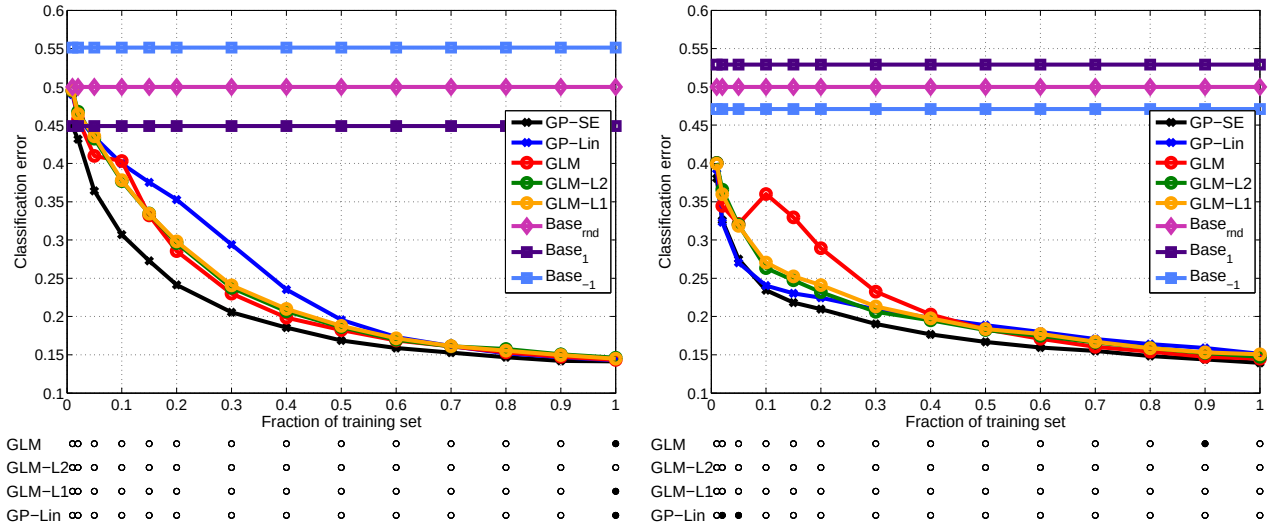
[2] More details can be found in e.g. [19].

[3] With the pairwise GP model the predictive distribution of y_t can also be estimated (see [19]) and used to express the uncertainty in the prediction relevant for e.g. sequential designs, reject regions etc.

4. EXPERIMENTAL RESULTS

4.1 Learning Curves

We use learning curves to compare the five models described in Section 3, namely the Logistic Regression model and two regularized version using the L1 and L2 norms and finally the Gaussian Process model using a linear and a squared exponential kernel. The learning curves are evaluated for individual subjects using 10-fold cross validation (CV) in which a fraction (90%) of the total number of pairwise comparisons constitutes the complete training set. Testing all possible combinations of e.g. 17 comparisons out of 171 when using 10% of the training set is exhausting. Therefore each point on the learning curve is an average over 10 randomly chosen equally-sized subsets from the complete training set, to obtain robust learning curves. Three different baseline error measures have been introduced, corresponding to a random choice of either of the two classes in each fold and two obtained by choosing either class constantly. Thus taking into account that the data set is not balanced between the two outcomes of $[-1; 1]$. In Figure 1 we show the learning curves as an average across all subjects. Using the entire dataset the models converge to similar classification errors of 0.14 and 0.15 for valence and arousal, respectively. On the valence dimension we see that using a fraction of the training data, the GP-SE model shows a clear advantage over the other models at e.g. 30% of the training data, producing a classification error of 0.21 whereas the GLM models produce around 0.23 and the GP-Lin at 0.29. The learning curves for the arousal dimension show a slightly different picture when comparing the different models. It is clear that using regularization on the GLM model greatly improves the classification error when training with up to 30% of the training data by as much as 0.10. The two GP models perform similar up to the 30% point on the learning curve but converges at a lower classification error than that of the GP-SE. Since all models converge to a similar classification errorrate we want to test whether they are the same on a classification level. We use the McNemar's paired test [22] with the $Null$ hypothesis that two models are the same, if $p < 0.05$ then the models can be rejected as equal on a 5% significance level. We test the GP-SE against the other four models pooling data across repetitions and folds for each point on the learning curve. For the valence data the GP-SE model is different in all points on the learning curve besides when using the entire trainingset for the GLM, GLM-L1 and GP-Lin model. For arousal data the GP-Lin model and the GP-SE cannot be rejected as being different when training on 2% and 5% of the training data and for the GLM model trained on 90% of the training data.



(a) Valence (b) Arousal

Figure 1. Classification error learning curves as an average across all subjects for 10-fold CV on comparisons comparing five models. A Gaussian Process model using a linear kernel ($GP\text{-}Lin$) and a squared exponential kernel $GP\text{-}SE$, logistic regression model (GLM) and two regularized versions using the L1 ($GLM\text{-}L1$) and L2-norms ($GLM\text{-}L2$). Three different baseline error measures have been introduced, corresponding to a random choice of either of the two classes in each fold denoted $Base_{rnd}$ and two obtained by choosing either class constantly denoted $Base_1$ and $Base_{-1}$. The circles below the figure show the McNemar's paired test with the $Null$ hypothesis that two models are the same, if $p < 0.05$ then the models can be rejected as equal on a 5% significance level. The test is performed between the $GP\text{-}SE$ model and the GLM, $GLM\text{-}L2$, $GLM\text{-}L1$ and $GP\text{-}Lin$. Non-filled circles indicate $p < 0.05$, and filled cirlces indicate $p > 0.05$.

4.2 AV Space

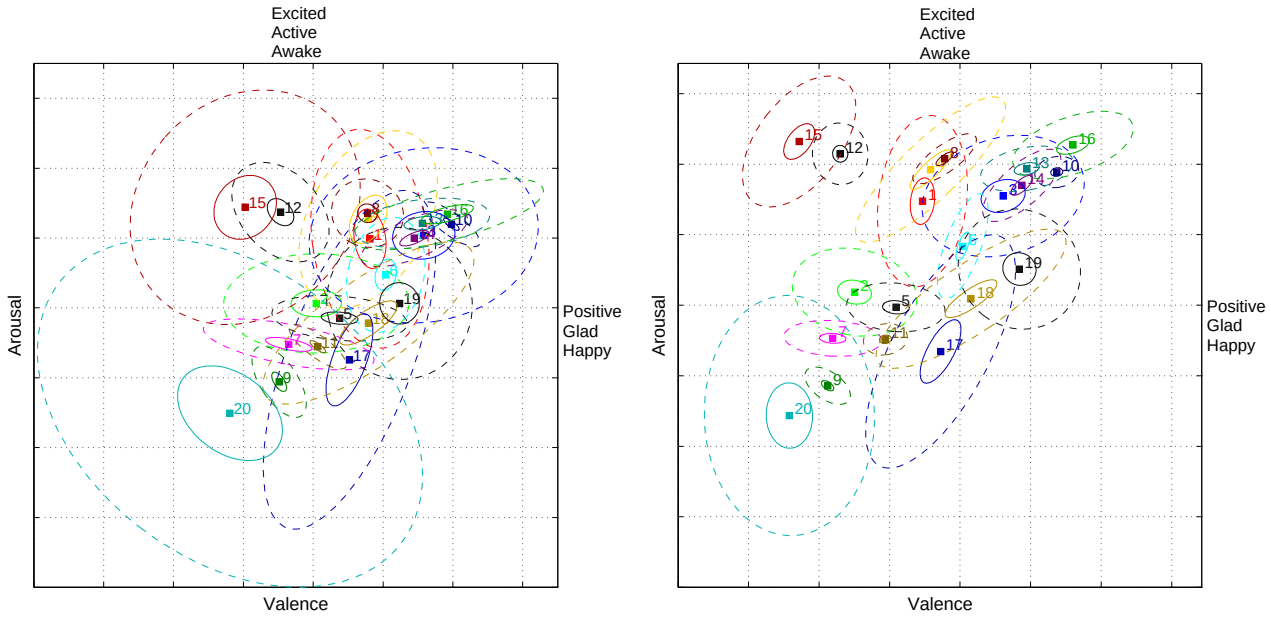
The learning curves show the performance of the models when predicting unseen comparisons. However, it may be difficult to interpret in terms of the typical AV space as one know from direct scaling experiments. To address this we show that both the GLM and the GP models can provide an internal, but unit free representation of the AV scores using the latent regression function $f(\mathbf{x}_t)$ in the case of the GP model, and by $f(\mathbf{x}_t) = \mathbf{x}_t\mathbf{w}^\top$ for the GLM models.

We first consider a model using all comparisons from all participants, thus obtaining a global mean model illustrated in Figure 2 with squares. In order to evaluate the variation across subjects, we train individual models on all comparisons from a given participant. The deviation from the global mean model is now calculated per comparison by comparing the latent difference in the global mean model with the latent difference in the individual model. The subjects deviation for a single excerpt is now evaluated as the average over all changes in latent differences for the 19 possible comparisons in which the excerpt is present. Finally, we take the variation across subjects and visualize it in Figure 2 as dashed and solid lines around each excerpt indicating the 50% and the 5% percentiles, respectively.

While the GLM and GP-SE models may seem quite different at first sight, we should focus on the relative location of the excerpts and not the absolute location in the unit free space. Comparing the relative placement of the excerpts (the center points) we see that the models are quite similar, also indicated by the averaged learning curves. In both models the relatively small variation over the subjects suggest that there despite minor subjective differences is a general consensus about the overall location of the given excerpts and the models have actually learned a meaningful representation.

4.3 Ranking Analysis

The learning curves only show the predictive classification power and does not give a clear picture as to the resulting ranking of the excerpts in the AV space. Two or more models can have the exact same classification error, but result in very different ranking of excerpts in the AV space. To quantify this difference in the ranking in the AV space we use Kendall's τ rank correlation coefficient. It is a measure of correlation between rankings and is defined as $\tau = (N_s - N_d)/N_t$, where N_s is the number of correctly ranked pairs, N_d is the number of incorrectly ranked pairs and N_t is the total number of pairs. When two rankings are exactly the same the Kendall's τ results $\tau = 1$, if the order of items are exactly opposite then $\tau = -1$ and when $\tau = 0$ they are completely different. In Figure 3 we notice that the linear models produce very similar rankings when trained on 1% with a Kendall's τ above 0.95. Between the GLM and the regularized models the Kendall's τ decreases to 0.7 at 10% of training data and increasing to 0.9 when using 50% for valence data. The largest difference in ranking lies between the GP models and both the regularized and unregularized GLM models for both valence and arousal. Using 10% of training data the comparison between the ranking of the GP-SE and GLM models produce a Kendall's τ rank correlation of 0.47 ending at 0.9 when using the entire training set for valence. Both the GLM and GLM-L2 when compared with the GP-SE lie below 0.9 us-



(a) GLM (b) GP

Figure 2. Predictions using the latent regression function for the Gaussian Process model and model parameters for the logistic regression model. The squares indicate the latent regression function values from a global mean model which is trained using all comparisons from all participants. The dashed and solid lines around each excerpt indicates the 50% and the 5% percentiles for the deviation from the global mean model calculated per comparison by comparing the latent difference in the global mean model with the latent difference in the individual model. The subjects deviation for a single excerpt is evaluated as the average over all changes in latent differences for the 19 possible comparisons.

ing the entire training set for arousal. It is noteworthy that between the 5% and 30% points on the learning curve, is where all models produce the most different rankings and as more comparisons are used they converge to similar but not same rankings.

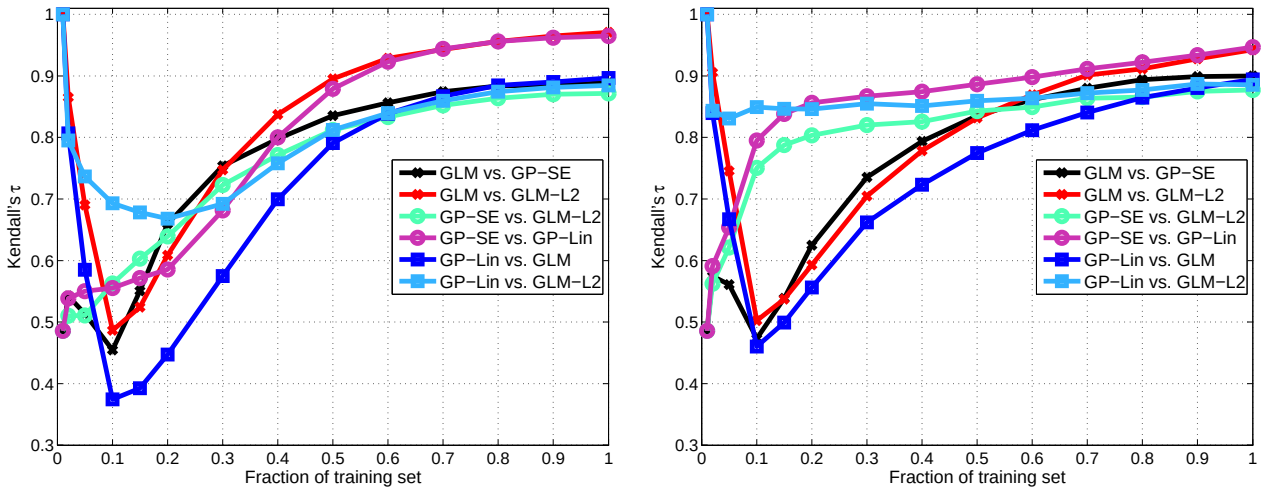
We have established that there is a difference in ranking of excerpts on the dimensions of valence and arousal given which models is chosen. As was shown in Figure 2 there is also a large difference in ranking across subjects, alternatively these individual differences can be quantified using the rank correlation. Using the GP-SE model trained on all the dataset, the Kendall's τ is computed between the predicted rankings between all subjects, which are shown in Figure 4. The ranking along the valence dimension shows a grouping of subjects where subject eight and three have the lowest Kendall's τ in average compared to all other subjects. This suggests a fundamentally different subject dependent understanding of the expressed emotion in music. Subject eight seem especially to disagree with subjects three, five, six and seven given the predicted latent regression function values. On the valence dimension subject six is very much in disagreement with other subjects, whereas subject four is in high agreement with most subjects.

4.4 Discussion

Five different regression models were introduced to model the expressed emotions in music directly by pairwise comparisons, as previously shown in [13] the results clearly show this is possible. Common for all models is the convergence to similar classification errors, indicating that given this limited dataset, that the underlying problem is linear and thus does not benefit from the flexibility of the non-linear GP-SE model, when using all available comparisons. But having all possible unique comparisons is an unlikely scenario when constructing larger datasets. This is the strength of the GP-SE model using only a fraction of training data for valence it is evident that it is improving predictive performance of around 0.08 comparing to a linear GP model using 30% of the training data. Which shows that it is not necessary to let participants evaluate all comparisons when quantifying the expressed emotion in music. For arousal data the GLM model benefits greatly with regularization when training with up to 40% percent of the training data with as much as 0.10 classification error. Whereas for valence all GLM models produce very similar results.

In previous work the predictions from the latent regression function was shown as a mean across subjects, here we emphasize the differences between subjects with the predictions by the model. Both the GLM and GP-SE model can produce results which show the relative position of excerpts in the AV space, and between models produce visually similar results. These differences are quantified between the ranking of the different models using Kendall's rank correlation coefficient emphasizing the fact that not only is there a difference in ranking amongst participants but also between models. This links the difference between models producing a given classification error and the resulting ranking produced by the model. Even though two models produce the same classification error they can end up with a different ranking of excerpts in the AV space.

Identifying differences between participants and their in-



(a) Valence (b) Arousal

Figure 3. Comparison of the ranking of the internal regression function predictions for different models using Kendall's τ rank correlation coefficient. Curves are an average of the Kendall's τ computed for each individual subjects predicted ranking across folds and repetitions.

ternal ranking of excerpts in the AV space can become a challenge when using a pairwise experimental paradigm to quantify the expressed emotion in music. We remedy this by using Kendall's τ computed between all users rankings provided by the GP-SE model. The results show that there is a great difference between users individual ranking producing a difference in Kendall's τ of as much as 0.55 for arousal and 0.35 for valence. Given the fact that the predictions by the models are so different for each subject this stresses the importance to distinguish between subjects. Currently we investigate individual user models which are linked/coordinated in a hierarchical Bayesian modeling framework in order both to obtain individual models and the possibility to learn from a limited set of pairwise data. In particular we see these models as a required tool in the examination of the difference between direct scaling methods and the pairwise paradigm presented in the current work. Future models will furthermore provide a principled approach for combining pairwise and direct scaling observations, thus allowing for optimal learning and absolute grounding.

5. CONCLUSION

In this paper we outlined a paradigm for obtaining robust evaluation of expressed emotion in music based on a two alternative forced choice approach. We examined five different regression models for modeling these observations all based on the logistic likelihood function extended to pairwise observations. The models ranged from a relatively simple GLM model and two regularized GLMs using the L1 and L2 norms to non-parametric Bayesian models, yet the predictive performance showed that all proposed models produce similar classification errors based on the entire training set. The true strength of the non-parametric Bayesian model comes into play when using a fraction of the dataset leaving good opportunities in constructing larger datasets where subjects do not need to evaluate all possible unique comparisons. It is left for future work to further analyze the detailed difference between the models. Furthermore we illustrated a significant difference between models and subjects in both AV space and quantified it using Kendall's τ with the conclusion that it is critical to model subjects individually.

Acknowledgments

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328.

6. REFERENCES

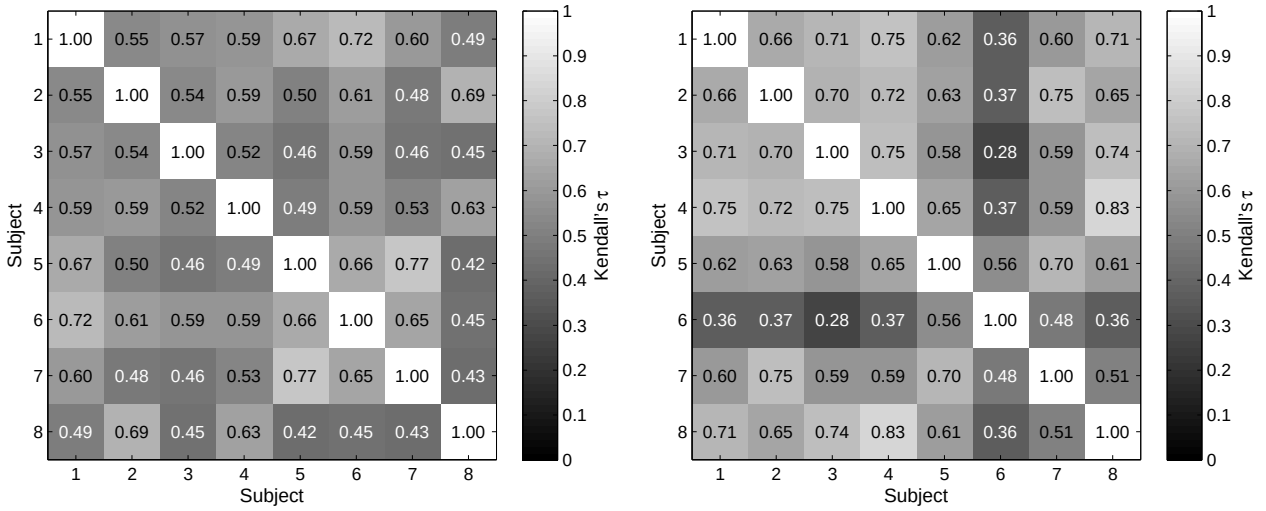
[1] K. Trochidis, C. Delbé, and E. Bigand, "Investigation of the relationships between audio features and induced emotions in Contemporary Western music," *8th Sound and Music Computing Conference*, 2011.

[2] E. Nichols, D. Morris, S. Basu, and C. Raphael, "Relationships between lyrics and melody in popular music," *10th International Conference on Music Information Retrieval (ISMIR)*, pp. 471–476, 2009.

[3] X. Hu and J. Downie, "When lyrics outperform audio for music mood classification: a feature analysis," *11th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 1–6, 2010.

[4] K. Bischoff, C. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification-a hybrid approach," *10th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 657–662, 2009.

[5] B. Schuller and F. Weninger, "Multi-Modal Non-Prototypical Music Mood Analysis in Continuous Space: Reliability and Performances," *12th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 759–764, 2011.



(a) Valence

(b) Arousal

Figure 4. Comparison of the different rankings between subjects using Kendall's rank correlation coefficient τ on the latent regression function output of the GP-SE model trained on the entire dataset. The values are averages of the Kendall's τ computed for all folds and repetitions.

[6] Y. Panagakis and C. Kotropoulos, "Automatic Music Mood Classification Via Low-Rank Representation," *19th European Signal Processing Conference*, no. Eusipco, pp. 689–693, 2011.

[7] M. Zentner and T. Eerola, *Handbook of Music and Emotion - Theory, Research, Application*. Oxford University Press, 2010, ch. 8 - Self-report measures and models.

[8] K. Hevner, "Experimental studies of the elements of expression in music," *American journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.

[9] J. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[10] Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," *11th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 255–266, 2010.

[11] E. M. Schmidt and Y. E. Kim, "Modeling Musical Emotion Dynamics with Conditional Random Fields," *12th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 777–782, 2011.

[12] E. Schubert, "Measurement and time series analysis of emotion in music," Ph.D. dissertation, University of New South Wales, 1999.

[13] J. Madsen, J. B. Nielsen, B. S. Jensen, and J. Larsen, "Modeling expressed emotions in music using pairwise comparisons," in *9th International Symposium on Computer Music Modeling and Retrieval (CMMR) Music and Emotions*, 2012.

[14] Y.-H. Yang and H. Chen, "Ranking-Based Emotion Recognition for Music Organization and Retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, May 2011.

[15] J. Madsen, "Experimental protocol for modelling expressed emotion in music," DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6246, 2012.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning / Data mining, inference, and prediction*. Springer, 2009, springer series in statistics.

[17] M. Schmidt, G. Fung, and R. Rosaless, *Optimization Methods for l1-Regularization*. UBC Technical Report, August 2009.

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[19] B. Jensen and J. Nielsen, *Pairwise Judgements and Absolute Ratings with Gaussian Process Priors*. Technical Report, DTU Informatics, "http://www2.imm.dtu.dk/pubdb/p.php?6151", September 2011.

[20] W. Chu and Z. Ghahramani, "Preference learning with Gaussian Processes," *22nd International Conference on Machine Learning (ICML)*, pp. 137–144, 2005.

[21] L. Rigutini, T. Papini, M. Maggini, and F. Scarselli, "Sortnet: Learning to rank by a neural preference function," *IEEE Transactions on Neural Networks*, vol. 22, no. 9, pp. 1368–1380, 2011.

[22] Q. McNemar, *Psychological statistics*. Wiley New York, 1969.