Pre-Print: Personalized Audio Systems - a Bayesian Approach

Jens Brehm Nielsen (1,2), Bjørn Sand Jensen (1), Toke Jansen Hansen (1), and Jan Larsen (1)

(1) Technical University of Denmark, DTU Compute, Matematiktorvet 303B, 2800 Lyngby, Denmark
(2) Widex A/S, Nymøllevej 6, 3540 Lynge, Denmark

Abstract

Modern audio systems are typically equipped with several user-adjustable parameters unfamiliar to most listeners. In order to obtain an optimal system setting, the listener is nevertheless forced to perform high-dimensional optimization with respect to the user's own objective. In the present paper, a general inter-active framework for performing robust personalization of such audio systems is proposed, which addresses the problems associated with traditional trial and error methods. The framework builds on Bayesian Gaussian process regression in which the *belief* about the user's *objective function* is updated sequentially. The setting to be evaluated in a given trial is then carefully selected by sequential experimental design. A modified Gaussian process model is suggested that assumes adjacent parameters to be correlated, which shows better modeling abilities compared to a standard model. We further demonstrate the framework in an interactive loop, where twelve test subjects obtain a personalized setting in a five-band constant-Q equalizer. The proposed approach is able to find a significantly better solution than obtained with random experimentation.

1 Introduction

The ever increasing number of features and processing possibilities in many modern multimedia systems, such as personal computers, mobile phones, hearing aids and home entertainment systems, has made it possible for users to customize these devices significantly. A downside in this trend is the large number of user adjustable parameters which makes it a daunting and complex task to actually adjust/optimize the devices optimally. For audio systems, the optimization is further complicated by perceptual and cognitive aspects of the human auditory and cognitive system, which result in a significant spread in subjects's opinions concerning the adjustment of a particular device. As a consequence, users often have to navigate in a high-dimensional parameter space,

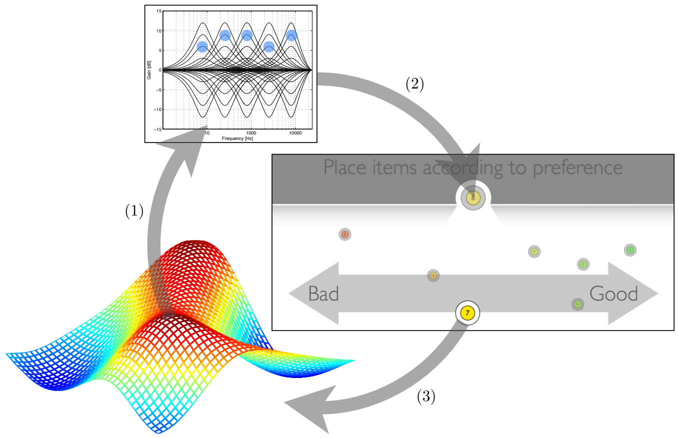


Figure 1: A conceptual overview of the interactive system. At step (1) we draw a new EQ from the current estimate of the subjects objective function. Next, at step (2) this particular EQ is associated with a *ball*, in this case number *eight*, in the visualized user interface. Finally, after the user has rated the new EQ, the objective function is updated to reflect current positions of all previous *balls*, this update occurs at step (3). We emphasize that the user at any time may select between previously sampled EQ by clicking the *balls*, making the current song play through the newly selected EQ.

which makes it extremely difficult for users to find even a local optimum. It is therefore of great interest to find and evaluate fast and flexible tools for optimizing user adjustable parameters with the aim to rapidly obtain a truly personalized audio system setting.

A prime example of such complex audio systems is hearing aids, where hundreds of parameters make up a unique and personal experience. It is therefore natural that this field has considered ways to efficiently learn an optimal setting based on preference (Kuk *et. al.* [7] and Baskent *et. al.* [1]) based on non-probabilistic methods. Recently—and the closest related to our approach—Birlutiu *et. al.* [3] have proposed a Gaussian process approach.

In audio reproduction systems—like home entertainment and professional mixing equipment—such preference learning approaches are relatively unknown for efficient personalization, despite the clear evidence that personalization may be beneficial in for example equalization (Paterson [10] and Zhang *et. al.* [17]). Existing approaches are based on non-probabilistic approaches such as Reed [12], Pardo *et. al.* [9], and Sabin *et. al.* [13].

In the present paper, we consider the audio reproduction scenario and focus on the task of optimizing the parameters of a five-band constant-Q equalizer (EQ). We propose and consider a combination of robust Bayesian modeling, an engaging user interface for user feedback and global optimization techniques in an interactive loop visualized in Fig. 1.

The loop constitutes a general framework where the inherent uncertainty in user feedback is addressed from a Bayesian viewpoint in which the belief in the user's (unknown) objective function is modeled with (warped) Gaussian process (GP) regression [14]. Since equalizers typically process signals in (overlapping) frequency bands—each with an associated set of parameters—the total set of parameters ends up constituting a high-dimensional space. However, parameters associated with adjacent frequency bands will typically impose correlation between them which should be exploited in the regression model to obtain better modeling abilities and thus more effective optimization. We therefore suggest a specific model which assumes correlation between specific input parameters/dimensions.

The framework uses an intuitive and simple graphical user interface for obtaining user ratings, which let the user listen to previously rated settings thus serving as anchors/references for future ratings. In contrast to standard practice, we do however not only let the user listen to previous settings, but we also let the user change the ratings of the previous settings, if for some reason a new setting would change e.g. the span of the scale. This is possible, since we are constantly updating our regression model to reflect the belief about the user's objective function given the ratings obtained so far.

Finally, we propose to use a sequential optimization technique to rapidly find a (possibly local) optimum of the user's objective function. The sequential design takes advantage of the Bayesian formulation by including the belief about the user's objective function. This significantly reduces the required number of settings that the user should rate in order to find an optimum .

Through model comparison, we first show that the model with assumed correlation between input parameters improves the modeling abilities compared to a traditional GP model without assumed correlation. The analysis is performed on real-world data, where 21 subjects have rated different randomly chosen settings of the EQ. Even for this EQ with relative few bands—which is thus perceptually well separated—we would expect the gains in adjacent bands to be somewhat correlated with regards to the user's objective. Secondly, we evaluate the usefulness of the entire framework in a real-world experiment where personalization of the EQ have been conducted for twelve test subjects. As the EQ has over fifty-nine thousands unique settings, the hypothesis is that the preferred setting will be hard to find without an efficient sequential design approach and correspondingly good modeling abilities. The results from the real-world listening experiments focusing on the statistical difference between random experimentation and sequential experimental design, show a clear advantage of the sequential design approach.

Our contribution is thus three fold: First in Sec. 2, we propose a general personalization framework with an intuitive user interface (Sec. 2.3), a principled modeling approach using warped Gaussian processes extended to expect correlation between adjacent input parameters (Sec. 2.1) and a sequential design method (Sec. 2.2). Secondly in Sec. 3.2, we show that the GP model extension provides better modeling abilities for our specific purpose. Thirdly, we evaluate the entire framework by a listening experiment in a real world interactive scenario and outline the results in Sec. 3.3. A discussion is provided in Sec. 4 and the paper is concluded in Sec. 5.

2 Personalization Framework

The proposed personalization approach uses an interactive loop to discover the user's preferred setting of a particular audio device, where we as an example use the EQ. The interactive loop is visualized in Fig. 1. The loop can conceptually be divided into three parts: a preference modeling part, a sequential design part and an interface part. The preference modeling part entails how to learn a user's objective function over EQ settings based on user ratings. The sequential design part covers how to choose new EQ settings to be rated based on what the model currently predicts. Finally, the interface part covers the design of the graphical user interface, such that it is both intuitive and easy to use for the users. The three parts are described in the following three sections.

2.1 Preference Modeling

We represent each system setting as a $d = 5$ dimensional vector of parameters, $\mathbf{x} = [x_1, ..., x_d]^\top$. Next, we assumed that the user's objective is an unobserved real-valued stochastic function (or process), such that each unique setting $\mathbf{x}_i$ has a corresponding real-valued function value, $f(\mathbf{x}_i)$, expressing the user's preference for the particular setting. This function is to be learned—and subsequently maximized—trough a number of experiments where we observe the user's expressed preference by a rating on a bounded scale, $y \in]0;1[$, where 0 is *Bad* and 1 is *Good* (see interface (2) on Fig. 1). At some point the user has evaluated n such distinct system settings $\mathbf{x}_i \in \mathbf{X}$ collected in $\mathbf{X} = \{\mathbf{x}_i | i = 1, ..., n\}$, with a related set of n responses denoted $\mathbf{Y} = \{y_i | i = 1, ..., n\}$.

We model the function mapping from settings, $\mathbf{x}_i$, to ratings, y_i, by a so-called *warped* Gaussian process [14]. A standard Gaussian process (GP) is a stochastic process defined as a collection of random variables, any finite subset of which must have a joint Gaussian distribution [11]. In effect, the GP is placed as a prior over any finite set of functional values $\mathbf{f} = [f_1, f_2, ..., f_n]^T$, where $f_i = f(\mathbf{x}_i)$, resulting in a finite multivariate Gaussian distribution over the set as $\mathbf{f} | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where each element of the covariance matrix $\mathbf{K}$ is given by a covariance function $k(\cdot, \cdot)$ such that $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. The GP prior can be used in non-parametric Bayesian regression frameworks where the likelihood function can be parameterized by a smooth and continuous function $f(\cdot)$.

However, our regression setup is special due to the bounded nature of the ratings. We therefore use a warped Gaussian process in which the original ratings in $\mathbf{Y}$ are transformed into a form where the data is modeled by a traditional Gaussian noise model [11]. Several warping functions would apply, but a natural choice is the inverse cumulative Gaussian (probit) $\Phi^{-1}(\cdot)$—with zero mean and unity variance—such that observations are warped as $z_i = \Phi^{-1}(y_i)$.

The final model is defined by,

$$\sigma_s|\theta_s \sim \text{half student-t}$$
$$\sigma_\ell|\theta_\ell \sim \text{half student-t}$$
$$f_i|\sigma_s, \sigma_\ell \sim \mathcal{GP}\left(m\left(\mathbf{x}_i\right), \mathrm{k}\left(\mathbf{x}_i, \cdot\right)_{\sigma_s, \sigma_\ell}\right)$$
$$z_i|f_i \sim \mathcal{N}\left(f_i, \sigma\right) \tag{1}$$
$$z_i = \Phi^{-1}\left(y_i\right), \tag{2}$$

where σ_ℓ is the length scale of the covariance function and σ_s is the variance of the latent function. We have placed hyper priors over the covariance parameters in order to provide a robust inference and prediction scheme, especially in the sequential setup with relatively few observations. These hyper priors are half student-t distributions [4, 15] with parameters, $\theta = \{\xi, \upsilon\}$, where ξ is the degree of freedom, and υ is the scale. These priors are *weakly informative* and have the effect of avoiding the GP model to (wrongly) fit hyperplanes with only few observations. With more observations available the effect of the hyper priors effectively vanishes. We note that the observation noise, σ, can be included in the covariance function.

Given this model, the main question remains regarding the covariance (or kernel) function, which effectively defines the smoothness of the function. We consider two covariance functions based on the general form of the squared exponential kernel [11]

$$\mathrm{k}\left(\mathbf{x}, \mathbf{x}'\right) = \sigma_s \exp\left(-\frac{1}{\sigma_\ell}\left(\mathbf{x} - \mathbf{x}'\right)^\top \mathbf{\Lambda}^{-1}\left(\mathbf{x} - \mathbf{x}'\right)\right) \tag{3}$$

In the first case, $\mathbf{\Lambda}$ is the identity matrix leading to the well-known (isotropic) squared exponential covariance function $\mathrm{k}_{\mathrm{iso}}\left(\mathbf{x}, \mathbf{x}'\right) = \sigma_s \exp\left(-\frac{1}{\sigma_\ell}\|\mathbf{x} - \mathbf{x}'\|^2\right)$. In the second case, $\mathbf{\Lambda}$ is a general positive semi-definite matrix defining a correlation between parameters (input space) as explicit prior information. We will denote this variant as the Mahanalobis covariance function, $\mathrm{k}_{\mathrm{mah}}\left(\mathbf{x}, \mathbf{x}'\right)$. The effect of the two options on the EQ example will be evaluated with reference to the standard case as `iso` and the Mahanalobis case as `mah`.

We turn to a standard GP inference scheme [11] in which the covariance parameters, σ_s, σ_ℓ, are approximated by point estimates by maximizing the marginal likelihood (or evidence) using a BFGS method and where the posterior $p(\mathbf{f}|\mathbf{Y}, \mathbf{X})$ is analytical tractable [14]. Extra terms are added to the standard evidence scheme [11] due to the student-t hyper priors. The predictive mean and (co)variance of the latent function, $\mathbb{E}(\mathbf{f}^*)$ and $\mathbb{V}(\mathbf{f}^*)$, are given in standard form [11] as

$$\mathbb{E}\left\{\mathbf{f}^*\right\} = \mathbf{K}_{\mathbf{X}\mathbf{X}^*}^\top\left[\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_i^2\mathbf{I}\right]^{-1}\Phi^{-1}\left(\mathbf{Y}\right) \tag{4}$$
$$\mathbb{V}\left\{\mathbf{f}^*\right\} = \mathbf{K}_{\mathbf{X}^*\mathbf{X}^*} - \mathbf{K}_{\mathbf{X}\mathbf{X}^*}^\top\left[\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_i^2\mathbf{I}\right]^{-1}\mathbf{K}_{\mathbf{X}\mathbf{X}^*} \tag{5}$$

where $\mathbf{K_{AB}}$ is the kernel matrix containing either evaluations between training inputs, $\mathbf{A} = \mathbf{B} = \mathbf{X}$, test inputs, $\mathbf{A} = \mathbf{B} = \mathbf{X}^*$, or between training and test inputs, $\mathbf{A} = \mathbf{X}, \mathbf{B} = \mathbf{X}^*$.

The predictive distribution and in particular the predictive uncertainty is a clear advantage of the probabilistic GP framework, since the predictive mean and predictive (co)variance can be used to determine the information gain in including a new candidate point into the model as considered in the next section.

2.2 Sequential Experimental Design

Classical experimental designs such as Latin Square or random experimentation [8] become increasingly infeasible in high dimensions. As an alternative, we propose to use sequential design methods which, by greedy selection of the most informative next sample, potentially achieves much faster convergence than fixed designs [6].

The main purpose is to define a selection criterion which finds the optimal of the (unknown) objective function. The applied criterion is a slightly modified version of the so-called *Expected Improvement* (EI) [6], a known criterion in the design of computer experiment (DACE) community. The expected improvement is for each candidate point, $\mathbf{x}_j$, defined as,

$$\mathrm{EI}(\mathbf{x}_j) = \sigma_{EI} \cdot \mathcal{N}\left(\frac{\mu_{EI}}{\sigma_{EI}}\right) + \mu_{EI} \cdot \Phi\left(\frac{\mu_{EI}}{\sigma_{EI}}\right), \tag{6}$$

where $\mathcal{N}(\cdot)$ is the standard Normal distribution and $\Phi(\cdot)$ is the standard cumulative Gaussian as before. Given the predictive distribution the EI is given by,

$$\mu_{\mathrm{EI}} = \mu_j - \mu_{\max}$$
$$\sigma^2_{\mathrm{EI}} = \sigma^2_j + \sigma^2_{\max} - 2\sigma_{j,\max}$$

where μ_j and σ_j is the predictive mean and variance of the test point and $\mu_{\max}$ and σ_{max} is the predictive mean and variance of the current maximum of the objective function (using the predictive mean as the predictor), i.e., the current best setting, all of which originate from Eq. 4-5. The correlation between the two function values, $\sigma_{j,\max}$, requires correlated predictions which we refrain from due to computation burden, thus $\sigma_{j,\max} = 0, \forall \mathbf{x}_j$. Hence, the selection of a new point to evaluate is given by

$$\mathbf{x}_{new} = \arg\max_{\mathbf{x}_j} \mathrm{EI}\left(\mathbf{x}_j\right)$$

which is then included in the current set of training points and evaluated by the user through the user interface. We refer to this as the `active` configuration, where the very first setting for the user to evaluate is chosen randomly. A random configuration `rnd` is included in which samples are selected randomly to provide a baseline method.

The interactive framework leaves four strategies to be investigated experimentally: `rnd-iso`, `rnd-mah`, `active-iso` and `active-mah`.

2.3 Interface

When applying absolute ratings, it is important to define anchor and/or reference points [2]. This allows subjects to compare stimuli with a fixed reference, such that each rating is *calibrated* both with respect to previous ratings, but also with respect to yet unobserved stimuli, which might redefine the end points of the rating scale. To address these two issues a graphical user interface similar to [9] is designed. Subjects can listen to previous EQ (references) and are allowed to change previous ratings based on the new one. Obviously, this means that ratings are neither directly comparable across subjects nor between iterations. However, it is not of particular interest to use ratings across subject to formulate one single optimal setting, but instead we are interested in personalized settings—one for each subject.

3 Experiment

To evaluate the different model configurations and experimental designs in a real-world scenario, an experiment was conducted, in which the five gains of the EQ are to be optimized by the 4 different versions of the proposed framework. The procedure and results are described in the following section.

3.1 Procedure

The experiment consisted of three parts: (1), (2) and (3) as visualized in Fig 3.1. During part (1), the user rates ten randomly chosen balls to learn how to use

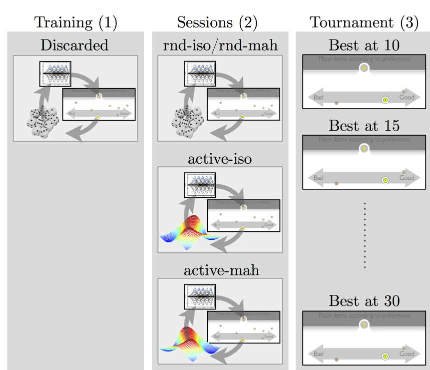


Figure 2: Visualization of the experiment with its 3 sessions: (1) Training, (2) Sessions and (3) Tournament.

the interface and to get an impression of the stimuli (EQ processed music). Part (2) consisted of three sessions for which the order of sessions was balanced across subjects. In each of the three sessions a particular model (`iso` or `mah`) and sequential design (`rnd` or `active`) are used to find a personalized setting of the EQ for the user. Finally in part (3), the preferred settings, found by each of the four combinations of models and sequential designs after 10, 15, 20, 25 and 30 presented settings, are determined by which model predicted the setting that is rated highest (in the tournament - see Fig. 3.1). Each tournament (as defined in Fig. 3.1) was repeated twice resulting in ten tournaments for which the sequence was randomized. In all parts, the sound was played back to the user through Sennheiser HD650 headphones and a FirestoneAudio FUBAR DACIII headphone amplifier at constant level. The output level was furthermore loudness normalized to the same level using a A-weighting filter, with the purpose to make the rating process easier for the test subjects, such that the listeners primarily focus on the tonal qualities—not the loudness.

3.2 Model Analysis

The interactive loop outlined in Sec. 2 has two critical blocks which will influence the convergence of the optimization procedure; the GP model describing the user's objective function—at all possible inputs—given the observations—only the currently rated inputs—and the sequential design approach. In this section we only seek to determine which GP model that best suits our purpose without the influence of the sequential design approach. We do this by evaluating the two GP models—`iso` and `mah`—in terms of their predictive performance on random data sets for 21 subjects. In machine learning and statistics, cross-validation is typically used to get an unbiased measure of the predictive performance. Since the random data sets for each subject contain only 30 ratings, we use leave-one-out cross validation (LOO-CV) [11] to get an effectively unbiased measure of the true predictive performance.

Performance is typically defined as an error measure through a cost function, such as the sum-of-squared error function. However, such error functions only include the absolute deterministic errors made by the model on noisy data without additionally considering if the model actually fits the noise correctly. For the sequential design approach to work efficiently, the model should both fit the data and account for the noise in the data as well as possible. To capture this in the performance measure, typically, the predictive likelihood $p(y^*|\mathcal{D}, \mathcal{M})$ of the unseen data points y^* given the model $\mathcal{M}$ and the observed data $\mathcal{D}$ is used.

To be able to compare the performance of two different models, a proper Bayesian and statistical way of doing this [16, 5] is to compare the predictive likelihood ratio $p(y^*|\mathcal{D}, \mathcal{M}_{mah})/p(y^*|\mathcal{D}, \mathcal{M}_{iso})$ between the two different models—`mah` and `iso`. This is also referred to as the *Bayes factor* [5]. A (log) Bayes factor larger than zero favors the model denoted in the nominator, whereas a (log) Bayes factor less than zero favors the model in the denominator.

For each of the 21 random data sets—one for each test subject— LOO-CV is used and the (log) Bayes factor is calculated for each LOO-CV split. This gives a total of 21×30 Bayes factors estimates shown in a histogram in Fig. 3.2. We see that on average, the `mah` model performs the best probabilistic predictions of users's individual ratings and thus appears to be the most suitable model due to

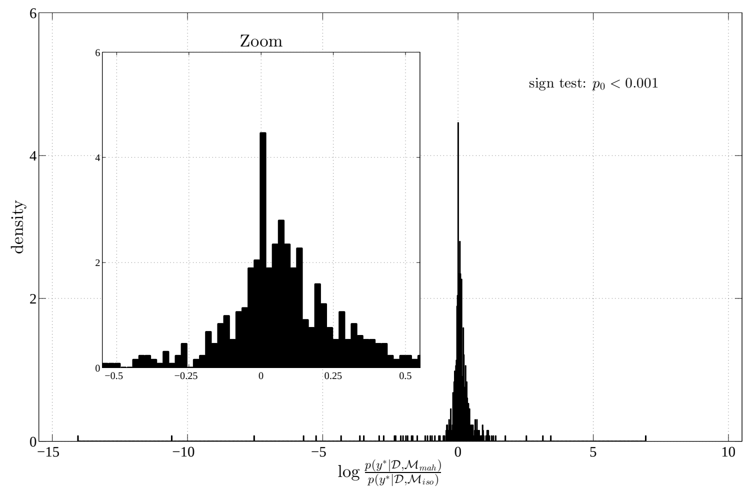


Figure 3: Predictive log-likelihood ratio (Bayes factor) over all leave-one-out cross-validation splits for all twenty-one test subjects. The p_0-value gives the probability of the null-hypothesis that the median is equal to zero (the to models are equally well) with the alternative hypothesis that the median is larger than zero (the Mahalanobis model is better than the isotropic) using an non-parametric sign test.

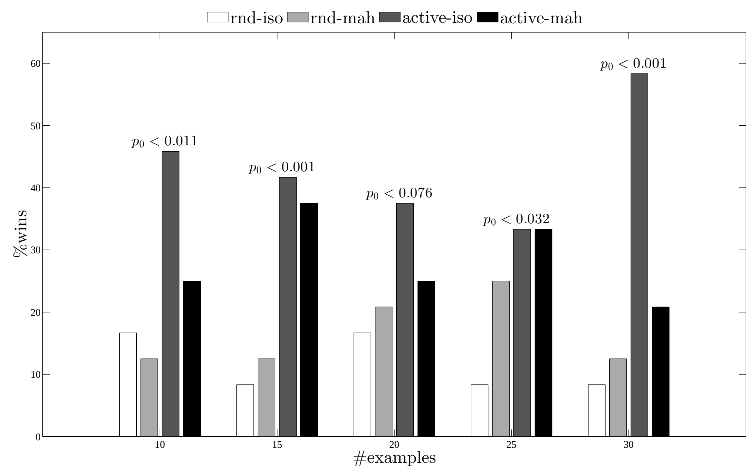
the assumed correlation between adjacent parameters. A non-parametric sign test shows that this is significant (sample size of 630).

3.3 Sequential Design Analysis

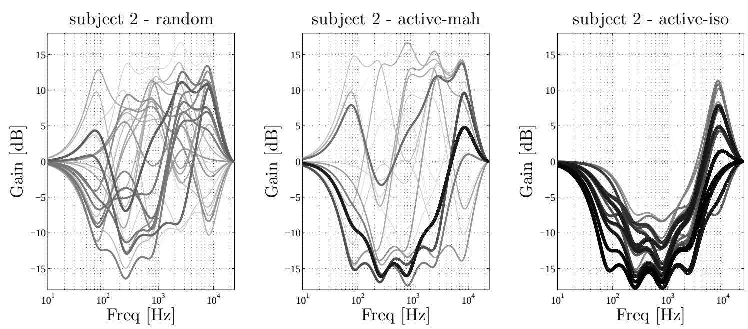
The results are summarized in Fig. 4(a). The illustrated p_0-values gives the significance level for which the hypothesis, that the total number of active wins is equal to the total number of random wins at each tournament point (#examples), can be accepted.

Averaged across subjects and repetitions, active sequential design is significantly better than random design after any given number of examples, as shown by the p_0–values. This is without differentiating between the two applied covariance functions. It demonstrates the potential of the Bayesian model and active learning methods in audio applications. It is furthermore noted that a standard fixed design will approximate the random configuration in this high-dimensional space.

The second aspect is if the more informative *Mahalanobis* (`mah`) prior results in a more accurate model with few ratings available. This is generally not the case, although the specific Mahalanobis model possesses better generalization abilities compared to the isotropic model as shown in Sec. 3.2.



(a) Learning Curve



(b) Ratings

Figure 4: (a): The percentage of times the predicted preferred setting by each of the four models wins over the other models across users at each of the five tournament points. The p_0-values is for accepting the null-hypothesis that the two active sequential design approaches is equal to the two random approaches using a binomial test. (b): Actual ratings of different EQ settings from the three Sessions for subject 2. The EQ curves are the imposed gain and the color and thickness of the EQ curves both indicate the rating, where think/dark black is a good ratings ($y \to 1$) and thin/light gray is a bad ratings ($y \to 0$).

4 Discussion and Future Work

The results presented in this paper has focused first on verifying that the proposed Mahanolobis model is suitable in this context, and secondly, demonstrating that the sequential design method actually performs as expected (and better than random). There are however many possibilities for further evaluation and development.

In regards to the specific prior, we believe despite the lack of evidence in the present paper, that the Mahalanobis covariance function will be found suitable in several audio applications—including the EQ example used here. We speculate that at least two additions would improve the performance of the Mahalanobis model in the suggested framework. Firstly, the modeling abilities could be improved by parameterizing the correlation structure in the Mahalanobis kernel by one parameter, which could then be inferred from data. The latter is easily accomplished in the GP framework by evidence maximization. Secondly, the sequential design criterion (Sec. 2.2) does not in its current form fully exploit the correlation between predictive function values for different settings. To include this correlation the covariance matrix between all unique settings must be calculated. Calculating these is currently computational infeasible. To overcome this and exploit the modeled correlation in the sequential design criterion, a greedy-gradient approach is currently being developed and tested with regards to find a (possible local) optimum of the (correlated) EI.

The current evaluation is based on a absolute paradigm with adjustable anchors in terms of previous ratings. It can however be quite demanding to keep track of all ratings, when there are several items (*balls*) present, which leads to inconsistent ratings. The GP based personalization framework is easily extendable with other paradigms such as pairwise comparisons or more general ranking based approaches. It is speculated that a more robust paradigm (with respect to user feedback) may further aid the optimization process.

Finally, it is the ambition to evaluate the proposed framework on a larger population, which could be accomplished by embedding the current personalization framework in a web application allowing evaluation of the approach on a larger scale.

5 Conclusion

We have proposed a method for obtaining true personalized systems—in particular audio systems—which utilizes the Bayesian probabilistic modeling approach through sequential design. This improves the high-dimensional preference optimization procedure in comparison to random (analogue to manual) experimentation. The solutions found by the sequential approach is significantly preferred by the test subjects over the solutions found by random experimentation. The results do not support any benefit in using the more informative Gaussian process prior with the Mahalanobis kernel compared to the less informative Gaussian process prior with the isotropic kernel. Supported by the demonstrated modeling benefits of the Mahalanobis kernel, it is nevertheless believed that future additions to the framework would be able to exploit the more informative Mahalanonis kernel and thus improve the performance of the framework.

Acknowledgment

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. This publication only reflects the authors' views.

References

[1] D. Baskent, C. L Eiler, and B. Edwards. Using genetic algorithms with subjective input from human subjects: Implications for fitting hearing aids and cochlear implants. *Ear and Hearing*, 28(3):370–380, 2007.

[2] S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley, July 2006.

[3] A. Birlutiu, P. Groot, and T. Heskes. Efficiently learning the preferences of people. *Machine Learning*, (July 2010), May 2012.

[4] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2004.

[6] D. R Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

[7] F.K. Kuk and N.M.C. Pape. The reliability of a modified simplex procedure in hearing-aid frequency-response selection. *Journal of Speech and Hearing Research*, 35(2):418–429, 1992.

[8] D.C. Montgomery. *Design and analysis of experiments*. Wiley, 2009.

[9] B. Pardo, D. Little, and D. Gergle. Building a personalized audio equalizer interface with transfer learning and active learning. *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies - MIRUM '12*, page 13, 2012.

[10] J. Paterson. The Preset Is Dead; Long Live the Preset. In *Audio Engineering Society Convention 130*, 2011.

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

[12] D. Reed. Capturing perceptual expertise: a sound equalization expert system. *Knowledge-Based Systems*, 14(1-2):111–118, March 2001.

[13] A. Sabin and B. Pardo. Rapid learning of subjective preference in equalization. In *Audio Engineering Society Convention 125*, 2008.

[14] E. Snelson, C. E. Rasmussen, and Z. Ghahramani. Warped gaussian processes. *Advances in Neural Information Processing Systems (NIPS)*, (16):337–344, 2004.

[15] J. Vanhatalo and A. Vehtari. Sparse log Gaussian processes via MCMC for spatial epidemiology. In *JMLR Workshop and Conference Proceedings*, volume 1, pages 73–89, 2007.

[16] A. Vehtari and J. Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142, 2012.

[17] D. Zhang, H. Xia, T. Chua, and G.A. Maguire. Impact of personalized equalization curves on music quality in dichotic listening. *Digital Audio Effects - DAFx '12*, pages 1–7, 2012.