# Integration of top-down and bottom-up information for audio organization and retrieval

**Jensen, Bjørn Sand; Larsen, Jan; Hansen, Lars Kai**

*Publication date:*
2012

*Document Version*
Publisher's PDF, also known as Version of record

# Integration of top-down and bottom-up information for audio organization and retrieval

Bjørn Sand Jensen

Kongens Lyngby 2012
IMM-PHD-2012-291

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

# Summary

The increasing availability of digital audio and music calls for methods and systems to analyse and organize these digital objects. This thesis investigates three elements related to such systems focusing on the ability to represent and elicit the user's view on the multimedia object and the system output. The aim is to provide organization and processing, which aligns with the understanding and needs of the users.

Multimedia, including audio and music, is often characterized by large amount of heterogeneous information, and the first element investigated in the thesis concerns the integration of such heterogeneous and multimodal information sources based on latent Dirichlet allocation (LDA). The model is used to integrate bottom-up features (reflecting timbre, loudness, tempo and chroma), meta-data aspects (lyrics) and top-down aspects, namely user generated open vocabulary tags. The model and representation is evaluated on the auxiliary task of genre classification.

Eliciting the subjective representation and opinion of users is an important and challenging element in building personalized systems. The thesis contributes with a setup for modelling and elicitation of preference and other cognitive aspects with focus on audio applications. The setup is based on classical regression and choice models placed in the framework of Gaussian processes, which provides flexible non-parametric Bayesian models. The setup consist of a number of likelihood functions suitable for modelling both absolute ratings (direct scaling) and comparative judgements (indirect scaling). Inference is typically performed by analytical approximation methods, including the Laplace approximation and expectation propagation. In order to minimize the cost of the often expensive

and lengthy experimentation, sequential experiment design or active learning is supported as an integrated part of the setup. The setup is applied in the field of music emotion modelling and optimization of a parametric audio system both with high-dimensional input spaces.

The final element considered in the thesis, concerns the general context of users, such as location and social context. This is important in understanding user behavior and in determining the users current information needs. The thesis investigates the predictability of the user context, in particular location, based on information theoretic bounds and a particular experimental approach based on context sensing using the ubiquitous mobile phone.

# Resumé (in Danish)

Den stigende tilgængelighed og brug af digitale medier kræver metoder og systemer til at forstå og organisere sådanne digitale objekter og det ideelt på en måde, der er i tråd med brugernes forståelse, forventninger og behov. Denne afhandling undersøger tre elementer i sådanne systemer som alle vedrører systemets evne til at repræsentere og frembringe brugerens syn på objektet eller systemets output.

Multimedie, inklusiv lyd og musik, er ofte karakteriseret ved store mængde af heterogene informationskilder, og det første element, der er undersøges i afhandlingen, er integration af sådanne informationskilder ved hjålp af Latent Dirichlet Allocation (LDA). Modellen anvendes til at integrere bottom-up aspekter (timbre, loudness, tempo og chroma features), metadata aspekter (lyrik) samt top-down aspekter i form af brugergenererede annotationer. Modellen og repræsentationen evalueres blandt andet på sin even til at repræsentere genre.

Modellering og eksperimentel frembringelse af en brugers interne repræsentation og forståelse af for eksempel musik er en generel udfordring i multimediesystmer og andre applikationer. Afhandlingen bidrager med en opsætning til modellering og frembringelse af præference og andre kognitive aspekter. Opsætningen er baseret på klassiske regressions- og beslutningsmodeller i rammerne af Gaussiske processer, hvilket resulterer i en række ikke-parametriske Bayesianske modeller. Opsætningen består af en række likelihood funktioner, der er egende til at modellere både brugeres absolutte vurderinger eller parrerede sammenligninger. Inferens i disse modeller er typisk udørt gennem analytiske approksimationsmetoder såsom Laplace-approksimationen og expectation propagation. For at minimere den ofte kostbare eksperimentelle forsøgtid, er sekventiel eksperimentel

design understøttet som en integeret del af opsætningen. Metoderne anvendes inden for modellering af følelser i musik og brugeroptimering af et parametrisk audio system med høj-dimensionelle data.

Det sidste aspekt, der er undersøgt, relaterer sig til brugerens generelle kontekst, som placering og social kontekst, hvilket er vigtigt for forståelsen af brugeradfærd og til afdækning af brugernes aktuelle informationsbehov. Afhandlingen undersøger forudsigeligheden af brugerens kontekst, navnlig placering, baseret på informationsteoretiske grænser og en bestemt eksperimentel tilgang baseret på indsamling af data fra den allestedsnærværende mobiltelefon.

# Preface

This thesis was prepared at The Department of Informatics and Mathematical Modeling (IMM), The Technical University of Denmark (DTU), in partial fulfillment of the requirements for acquiring the Ph.D. degree at DTU.

The project was funded by DTU, initiated April 2009 and completed December 2012. Throughout the period, the project was supervised by associate professor Jan Larsen and by co-supervisor professor Lars Kai Hansen.

The thesis reflects the research part of the project. It consists of an summary report in combination with a collection of published and submitted research papers written during the period and published during the project period (or immediately thereafter [1]).

The project is motivated by the challenges involved in processing, modelling and organization of multimedia in particular in systems where users plays an integral role. This is a highly cross-disciplinary field including elements from digital signal processing, human-computer interaction, cognitive modelling and machine learning. The thesis therefore consist of contributions originating in three different research fields of course with some overlap. It is therefore the aim of the summary report to give a coherent and general overview of the contributions from a system perspective. Hence, this summary report is therefore not an exhaustive walk-through of all applied methods and detailed derivations, but an attempt to place the contributions in an overall and general context of user driven machine learning systems.

[1] Note that the final version of the report has been updated with the published versions of the papers originally indicated as *submitted*

The report further refrains from describing well-known methods such as Expectation Maximization, Support Vector Machines and K-means, and simply provide textbook reference for standard methods well-described in textbooks or elsewhere. As a consequence it is assumed that the reader is familiar with basic probability theory and its application in machine learning.

Bjørn Sand Jensen
January 14th, 2013

# Dissemination

**Papers (peer-reviewed)**

A Bjørn Sand Jensen, Jakob Eg Larsen, Kristian Jensen, Jan Larsen, and Lars Kai Hansen. Estimating Human Predictability from Mobile Sensor Data. IEEE International Workshop on Machine Learning for Signal Processing, Pages) 196-201, 2010. DOI:10.1109/MLSP.2010.5588997

C Bjørn Sand Jensen, Jens Brehm Nielsen, and Jan Larsen. Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes. IEEE International Workshop on Machine Learning for Signal Processing. Pages 1-6, 2011. DOI:10.1109/MLSP.2011.6064616

D Bjørn Sand Jensen, Javier Saez Gallego and Jan Larsen. A Predictive Model of Music Preference using Pairwise Comparisons. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Pages 1977-1980, 2012. DOI:10.1109/ICASSP.2012.6288294

F Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen. Modeling Expressed Emotions in Music using Pairwise Comparisons. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR). Pages 526-533, 2012.

E Jens Madsen, Bjørn Sand Jensen, Jan Larsen and Jens Brehm Nielsen. Towards Predicting Expressed Emotion in Music from Pairwise Comparisons. 9th Sound and Music Computing Conference. Pages 350-357, 2012.

G Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen. Pseudo Inputs For Pairwise Learning With Gaussian Processes IEEE International Workshop on Machine Learning for Signal Processing, Pages 1-6, 2012 DOI:10.1109/MLSP.2012.6349812

H Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen and Lars Kai Hansen Towards a universal representation for audio information retrieval and analysis International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Pages 3168-3172, 2013. DOI:10.1109/ICASSP.2013.6638242.

I Jens Brehm Nielsen, Bjørn Sand Jensen and Toke Jansen Hansen Personalized Audio Systems - a Bayesian approach Audio Engineering Society Convention 135, Pages 1-10, 2013. [2]

J Bjørn Sand Jensen, Jens Brehm Nielsen and Jan Larsen. Bounded Gaussian Process Regression. IEEE International Workshop on Machine Learning for Signal Processing, Pages 1-6, 2013 DOI:10.1109/MLSP.2013.6661916. [3].

**Workshop Contributions (peer-reviewed)**

B Bjørn Sand Jensen, Jan Larsen, Lars Kai Hansen, Jakob Eg Larsen, and Kristian Jensen. Predictability of mobile phone associations. European Conference on Machine Learning : Mining Ubiquitous and Social Environments Workshop (ECML-MUSE), Pages 1-15, 2010, Peer Reviewed.

**Miscellaneous**

Other papers and software prepared during the project, but not part of the thesis

[4] Tommy S. Alstrøm, Bjørn S. Jensen, Mikkel N. Schmidt, Natalie V. Kostesha and Jan Larsen. Haussdorff and Hellinger for Colorimetric Sensor Array Classification IEEE International Workshop on Machine Learning for Signal Processing, Pages 1-6, 2012. DOI:10.1109/MLSP.2012.6349724

· Bjørn Sand Jensen & Jens Brehm Nielsen
Matlab® Toolbox for Preference Learning with Gaussian Proceses
- with documentation, 2013

[2]Note that this paper was originally included in the thesis as *submitted* (and included in the thesis) as: Jens Brehm Nielsen, Bjørn Sand Jensen and Toke Jansen Hansen, Fast and flexible elicitation of preference in complex audio systems, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, but subsequently published as I

[3]Note that this paper was originally included in the thesis as a *technical report* entitled: On Bounded Regression with Gaussian Processes but was subsequently published as J

# Acronyms

| | |
|---|---|
| **AIC** | Akaike information criterion |
| **ANOVA** | Analysis of variance |
| **ARD** | Automatic Relevance Determination |
| **BIC** | Bayesian Information Criterion |
| **BTL** | Bradley Terry Luce |
| **BFGS** | Broyden Fletcher Goldfarb Shanno |
| **CCA** | Canonical Correlation Analysis |
| **CV** | Cross-validation |
| **DCT** | Discrete cosine transform |
| **DACE** | Design and analysis of computer experiments |
| **EVOI** | Expected Value of information |
| **EM** | Expectation Maximization |
| **EM** | Expectation Propagation |
| **FFT** | Fast fourier transform |
| **FI(T)C** | Fully Independent (Training) Conditional |
| **GLM** | Generalized Linear Models |
| **GMM** | Gaussian Mixture Model |

| | |
|---|---|
| **GP** | Gaussian Process |
| **GEVIO** | Gradient of EVIO |
| **HB** | Hierarchical Bayes |
| **HDP** | Hierarchical Dirichlet process |
| **IVM** | Informative Vector Machine |
| **KL** | Kullback-Leibler |
| **LSA** | Latent Semantic Analysis |
| **LDA** | Latent Dirichlet Allocation |
| **EP** | Expectation Propagation |
| **KNN** | k-nearest neighbor |
| **LOO** | Leave-One-Out |
| **LOO-CV** | Leave-One-Out Cross validation |
| **LZ** | Lempel-Ziv |
| **MSE** | Mean Squared Error |
| **MAE** | Mean Absolute Error |
| **MAP** | Maximum-a-Posteriori |
| **MAP-II** | Maximum-a-Posteriori Type-II (estimation) |
| **MCMC** | Markov Chain Monte Carlo |
| **MI** | Mutual Information |
| **MIR** | Music Information Retrieval |
| **MKL** | Multiple Kernel Learning |
| **MFCC** | Mel-frequency cepstral coefficients |
| **ML** | Maximum Likelihood |
| **ML-II** | Maximum Likelihood Type-II (estimation) |
| **MSD** | Million Song Dataset |
| **MTK** | Multi-task kernel |
| **NN** | (Artificial) Neural Network |

| | |
|---|---|
| **NMF** | Non-negative matrix factorization |
| **NMI** | Normalized Mutual Information |
| **PCA** | Principal Component Analysis |
| **PI(T)C** | Partial Independent (Training) Conditional |
| **pLSA** | Probabilistic Latent Semantic Analysis |
| **P/L** | Probit/Logit) |
| **PLP** | Perceptual linear predictive (analysis/encoding) |
| **PJK** | Pairwise Judgement Kernel |
| **PPK** | Probability Product Kernel |
| **SSK** | Semi-Supervised Kernel |
| **SVM** | Support Vector Machine |
| **SVMrank** | Support Vector Machine for ranking |
| **RVM** | Relevance Vector Machine |
| **THOMP** | Thompson (sampling) |
| **UCB** | Upper Confidence Bound |
| **VB** | Variational Bayes |
| **VOI** | Value of information |
| **VQ** | Vector Quantization |

# Notation

The following contains a list of common notation and symbols used throughout the report, which may differ slightly from the contributions in order to provide a coherent notation across different focus areas. Variations and specialized use may occur which is made clear from the context.

**Various sets of variables and observations**

| | |
|---|---|
| $\mathbb{R}$ | The reals. |
| $\mathbb{Z}$ | Integers. |
| $\mathbb{N}$ | Natural numbers (including zero). |
| $\mathbb{X}$ | Domain of the input variable / input space. |
| $\mathbb{Y}$ | Domain of the output variable / output space. |
| $\mathcal{X}$ | A set. Typically of input instances from $\mathbb{X}$. Typically index with $n$ |
| $\mathcal{Y}$ | A set. Typically of outputs from $\mathbb{Y}$. Typically indexed with $k$ |
| $\mathcal{V}$ | A set. Typically used to denote a vocabulary of words indexed by $v$. |
| $\mathcal{D}$ | A set. A joint collection of inputs and outputs, $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, so a dataset. |
| $\mathcal{C}$ | A (choice) set of inputs. A subset from $\mathcal{X}$ used in a specific likelihood. |
| $\mathcal{E}$ | An experiment set. A subset from $\mathcal{X}$ |

**Size/count parameters**:

| | |
|---|---|
| $D$ | Dimension of input space $\mathbb{X}$ |
| $N$ | Number of inputs, $N = \lvert\mathcal{X}\rvert$ or as or general count, typically with an informative subscript. |
| $K$ | Number of observations/experiment $K = \lvert\mathcal{Y}\rvert$. . |
| $C = \lvert\mathcal{C}\rvert$ | Size of choice set. Typically indexed with $k$. |
| $M$ | Number of data modalities or as a general count. Typically indexed with $m$ |
| $V = \lvert\mathcal{V}\rvert$ | Size of a given vocabulary, i.e. V words in the vocabulary. |
| $E = \lvert\mathcal{E}\rvert$ | Size of (candidate) experiment set. |
| $Z$ | Number of latent components, e.g. number of Gaussian components or topics. Related index variable is $z$. |
| $S$ | Number of songs, typically indexed with $s$. Note that $S$ is also used to denote various information theoretic measure such as entropy, however never in the same context as song. |

**Variables and observations:**

| | |
|---|---|
| $y$ | A output variable/observation (used in supervised context only) |
| $\mathbf{y}$ | A multidimensional output variable / observation (used in supervised context only) |
| $X$ | A random variable (used when required to differential between the variable itself and the outcome which is clear from context) |
| $\mathbf{x}$, $\mathbf{w}$ | (Multi-dimensional) input (or outcome when required to differentiate between outcome and variable which is clear from context) |
| $\mathbf{X}$, $\mathbf{W}$ | A collection of inputs $N \times D$ |
| $x^*/\mathbf{x}^*$ $w^*/\mathbf{w}^*$ | A test input |
| $y^*/\mathbf{y}^*$ | A test output |

**Probabilities and distributions**

The notation does not distinguish between probability and probability densities, where it is clear from the context. Capital letters is used for the variable itself where it is advantageous to differentiate (Chapter 4) and in following list to differentiate between variable and outcome).

| | |
|---|---|
| $P(X = x)$ | The probability that the random variable $X$ takes on a given value $x$. $P(X) \in [0; 1]$. |
| $p(x\|\boldsymbol{\theta})$ | A probability density parameterized by elements in $\boldsymbol{\theta}$ |
| $P(X = x\|Y = y)$ | A conditional probability probability where $X$ is condition on another random variable, $Y$ |
| $p(y\|x)$ | A conditional probability density function (or condition probability, which is clear from context) |
| $\mathbb{E}(X)$ | Expectation of a random variable, $X$. |
| $\mathbb{V}(X)$ | Second order moment of the random variable $X$, i.e. variance. Also used to denote covariance. |

**Distributions, processes and related functions:**

A stochastic process is a collection of random variables $X_i$ indexed by i $\mathbf{X} = \{X_1, X_2, ..., X_N\}$. In the report a Gaussian process considered from a function viewpoint a the following notation is employed, and using an informal notation, a Gaussian process is then denoted $f = \{f_1, f_2, ..., f_N\}$ or $f(\cdot) = \{f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_N)\}$, thus $f_i \triangleq f(x_i)$ being the individual random variables. It is noted that the Gaussian process is defined for all $\mathbf{x} \in \mathbb{X}$, i.e. in principle infinite, but we usually consider a finite subset through a vector $\mathbf{f}$ of the function evaluated a finite set of inputs, i.e., $\mathbf{f} = [f(x_1), f(x_2), ..., f(x_N)]$, resulting in a tractable finite multi-variate Gaussian distribution.

| | |
|---|---|
| $\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$<br><br>$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | Normal/Gaussian Distribution (used interchangeably) [26, App. B] |
| $\boldsymbol{\Phi}(z)$ | Cumulative Gaussian, with mean 0 and standard deviation 1 |
| $\boldsymbol{\Phi}^{-1}(z)$ | Probit function. Inverse cumulative Gaussian. |
| $Beta(\alpha, \beta)$ | A standard two parameter beta distribution |
| Categorical($\lambda$) | A categorical distribution [26]. A multinomial distrbution with only one draw. Parameterized by the probabilities $\lambda$. |
| Dirichlet($\boldsymbol{\alpha}$) | A Dirichlet distrbution parameterized by the concentration parameter $\boldsymbol{\alpha}$ [26, App. B] |
| TG($\mu, \sigma$)<br><br>Truncated G. | Standard truncated Gaussian distribution usually with support in $[0, 1]$ |
| k($x, x'$) | Covariance function, co-defining a Gaussian process |
| m(x) | Mean function, co-defining a Gaussian process |
| $\mathbf{K}$ or $\mathbf{K_{XX}}$ | Covariance marix with elements k($x, x'$) between all training inputs. |
| $\mathbf{k}_r$ | Covariance vector between all traning inputs and a test input $\mathbf{x}_r$. |

**Information theory**

| | |
|---|---|
| $S(X)$ | (Shannon) Entropy of the random variable $X$ |
| $S(X\|Y = y)$ | Conditional entropy of the random variable $x$ conditioned on the random variable $Y$ taking a particular value |
| $S(X\|Y)$ | Conditional entropy of the random variable $S$ conditioned on the random variable $Y$ - or a dataset in some cases. |

# Contents

Chapter 1

# Introduction

The growth of the digital multimedia world, in terms of both scale and use, makes it increasingly important to create systems to process and organize digital multimedia objects, such as text documents, books, video and audio objects with the aim to increase the productivity and the satisfaction of the users. The emphasis on the user is especially important in multimedia systems, where the information itself has a profound perceptual and cognitive influence on the users, such as music which has the ability to both move, repel and even heal us [185]. It has therefore become an important engineering task to design and build systems to store, process and organize multimedia information, designed to be well aligned with the user's needs and expectations.

The systems considered in this thesis ranges from relatively simple reproduction devices like personal media players or hearing aids, to more complex organization and retrieval systems, such as search engines and recommendation services. In the first case, the system task is typically to provide a optimally processed versions of the input conditioned on system parameters. In the second case, the system task is essentially to define similarity between user query and objects to be able to return relevant search results. In order to produce a particular system result and output - such as search results, processed audio files or recommendations - a given system often make use of a so called computational representation or mathematical models. This representation essentially defines the similarity and relations between either system parameters or the objects. The only as-

sumption regarding a system in this context; is that these are designed to be utilized and operated by human subjects: the users. Such users have conscious or unconscious information needs (see e.g. [144][211]), a subjective understanding of objects relations and certain expectations to the system response. We threat such user related and specific aspects under the general notion of a **user representation** encompassing the general state of the user, possibly affected by the current environmentally context of the user such as location and social context. If the computational representation and the user's representation is not aligned suboptimal system performance is typically encountered. We here define this misalignment as the **semantic gap** [198]. The overall system goal is to minimize this semantic gap by ensuring that the computational representation is well-aligned with the user's representation and needs.

Modern computational representations for organization and retrieval include the latent semantic view of multimedia objects in text mining [55] and music [240, 239, 105], which in a unsupervised manner extracts the latent semantics of the data in order to provide grouping and organization. The main goal is to provide representations well aligned with general human representations, for which reason the term 'cognitive' components [83] is sometimes preferred. Such purely content based and unsupervised representations is here defined as the **bottom-up** view or a computational low-level representation. Examples of such bottom-up based systems include content based recommendation in for example image retrieval and music recommendation/similarity [127].

An unsupervised bottom-up representation may provide general alignment with generic perceptual and cognitive representations evaluated over the general population. However, representation of objects and/or the system output is typically highly subjective and depends on the general state of the user. To obtain a computational representation aligned with the representation of the individual user, it is thus necessary to obtain information regarding the user view on objects, system output and possibly the user's environmental context. In the simplest case this information may be as simple as providing labels for the objects, which results in classical supervised machine learning setting (see e.g. [26]. Other examples include reproduction systems such as personal entertainment systems, where system parameters are adapted based on user's indication of preference. Such consciously expressed information is here defined as a **top-down** and a computational representation based solely on this view is referred to as high-level representation. Examples of purely top-down driven systems include collaborative filtering, relying only on user's ratings such as Netflix [166] and Amazon [6]. Slightly more subtle top-down driven systems include music recommendation services like last.fm [119]. Here user provided tags can be seen as a conscious wish to express the user's view on the object or system output. The high-level computational representation can, for example, be obtained by latent semantic analysis [125], classification or regression models for represent-

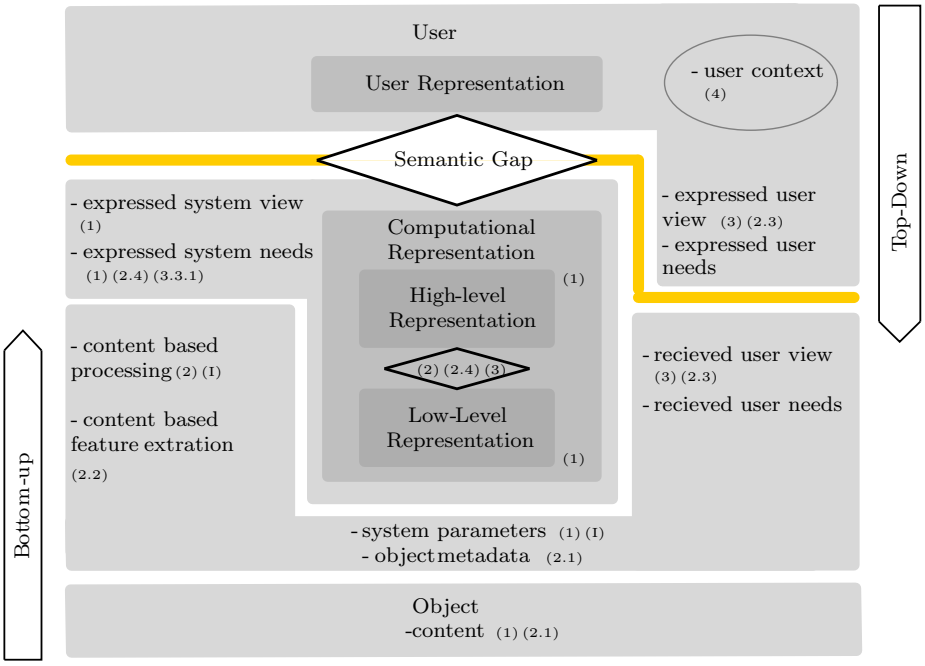



Figure 1.1: A simplified view of a general information processing system. A user's understanding, needs and expectations is represented by a (instantaneous) *user representation*, and the consciously expressed versions of this representation towards the objects or a given system output. The system consists of the objects either processed for e.g. based on some system parameters (like in reproduction) or as raw objects (retrieval and recommendation). The system considers *bottom-up* information in the form of content based features, object meta-data (context), and top-down information in the form of user generated information, e.g. *expressed user view* and common knowledge. The system models and integration via a *computational representation*, possibly consisting of a separate low-level representation , and a possibly separate high-level representation. The system presents the results and information though an *expressed system view*, and can request information through *expressed systems needs*. The actual human-computer-interface between the system and user is represented by the yellow bar, but not considered in the thesis.

ing aspects such as categorization (e.g. genre) and preference. Such top-down driven systems has the possibility to determine a representation based on the user's expressed views, namely the user generated data. However, such user generated data is typically a noisy expression of the true user representation due to subject inconsistencies and drift in the user representation itself, for example due to change in user context. The high-level computational representation, therefore, only reflects the expressed version of the user representation; not the representation its self. Furthermore, a purely top-down approach suffers from the cold-start problem, in which some ratings may be missing for a subset or all objects (or system outputs).

With the pros and cons of both bottom-up and top-down views, a potential solution towards fully bridging the semantic gap seems to be the combination of both. In combination, the two views is potentially robust to missing and noisy user generated view, and the lack of individualization in purely bottom-up driven systems. So far the definition of computational representations has been divided to a low- and a high-level representation based on information on which they were based. This simply illustrates that such separation is sometimes possible and logical on its own. In a combined view, however, simple representation can, for example, include standard classification and regression models. In this case, the bottom-up view is simply used as a input (feature) to the representation leading to a top-down aspect without the need or requirement for separate bottom-up and top-down representations[1]. More subtle systems, which combines the two separate computational representations, include recommendation systems which takes into account both content and annotations (for example in the form of two separate or combined LSA representations). A number of such systems has been proposed for multimedia systems, for examples in image retrieval [85][187], and in various form in the music information retrieval (MIR) community [188][43][208][240][206].

An important part, not addressed in the so far static description of a system, is the actual dynamic use of the system, i.e., the interaction taking place between the system and the users. From the **user perspective** this interaction is the process of using the system to obtain a given (potentially unconscious) need, i.e, find relevant webpages or listening to music with the best possible reproduction performance. From the **system perspective** this is ideally a combined process of **a)** satisfying the user's information needs and **b)** learning and/or adapting the representations in order to achieve the first part. Part a) is given by the main purpose (as seen from the users) of the system, for example retrieval/recommendation or reproduction. Part b) is the process of optimizing the computational representation (of both users and the objects) in an optimal

[1] One may argue that for example kernel methods require a low-level representation of the inputs in the form of a kernel function which on its own defines low-level similarity and relations, i.e. a representation.

fashion by obtaining the users top-down view of the objects, object relations, user relation or system results. And potentially (decoding) the context of the users. Thus, part b) relates to fulfilling the information needs of the system. Fulfilling the systems needs can be done in many ways, i.e, explicitly querying the users by requesting annotations - or implicitly by observing the user's response to an result such as click-through rates and implicit feedback [10]. In combination, we refer to part a) and b) as the **interactive learning** process or simply **active learning** when the focus is on the algorithmic and not the system aspects.

The defining aspect of the systems (and algorithms) is the user. While there are many open purely technical and algorithmic problems towards the general and ultimate machine learning system, many challenges relates directly to the users of the system. This thesis will focus on a few of the many aspects outlined above which are outlined in the following section.

## 1.1 Focus Areas

The ambition in many application domains is to eventually design and construct systems with many of the appealing properties discusses above. This thesis will focus on three different, but equally important elements of the general challenge, namely a combined computational representation of music, preference learning with focus on audio and music - and finally human predictability with relevance to user context and characterization of human behavior. These aspects are conceptually outlined below while a more detailed overview and technical aspects are covered in individual chapters 2 3 4 presenting short introductions to the different fields and the contributions.

### Computational Representations of Music

Multimedia and in particular music has an immense impact on the modern society in terms of well-being [51][185] and commercial value (e.g. [95][112]). The large boost in music consumption created by the personal and portable devices, such as MP3-players and smartphones, has created the need and demand for services which organize, recommendation and stream the actual content.

However, few if any of the current music recommendation and retrieval services has truly managed to provide relevant and personal music recommendation and search results [43]. This is for example reflected in the way many users still

discover new music, namely through the 'non-interactive' radio [171]. Potential solutions towards a successful music recommendation and retrieval system (see e.g. [208][43]) is based on the general system outlied in Fig.(1.1). This implies computational music representations which are based both on bottom-up audio analysis and user based top-down views reflecting multiple high-level aspects of the music, such as emotion, preference and categorization. In the optimal case this should be combined with interactive process ensuring that the representation is updated based on changes in environmental and social context.

The aim of this thesis is to investigate a subpart of this 'ideal' music system. The goal is specifically to investigate models and representations which can combine the bottom-up view - based on audio content analysis - with a top-down view based on an expressed user view obtained through user annotation annotations. In particular, the aim is to evaluate the representation on the recently published Million Song Dataset which will ensure both scalability of the methods and generalization of the results. Secondly, the goal of the thesis is to examine robust and flexible way to represent and elicit cognitive aspects such as preference and emotion in music, based on the methods provided by the next focus area.

## Preference Learning with Gaussian processes

Explicit, top-down user ratings and judgements is a major element in many modern information processing systems such as movie and book recommendation, satisfaction of online web services, implicit user feedback via click-through rates in online advertising or search engines [10] and skipping behavior in music playlists [174].

Such ratings, judgement and feedback represent a general desire from the system to elicit and understand perceptual and cognitive aspects of the users in order to optimize system performance. However, robustly eliciting such aspects is often complicated by the aspects themselves, such as preference, which are inherently difficult to elicit and represent due to the profound and diverse cognitive effect for example audio and music has on the users. These issues can often be framed in terms of standard signal detection theory [230] where concepts such as internal noise, bias and drift can be used to describe the challenges in obtaining robust ratings and judgements.

Experimental psychology has dealt with such issues in decades and often rely on relative simple experimental protocols for examining one (or few) effects in well-controlled situations and analyses the results with standard statistical test. The sensormetric field, often applied in the food sciences for optimizing

products [163], also deals with similar aspects. However this field typically focus on discriminative testing and post-analysing a few, well-defined effects and in particular, one (or a few) fully controllable variables. In the system view of this thesis, such methodologies are available only in situations where the aim is to post-optimize explicit system parameters in for example reproduction systems. This does not usually allow for application in interactive systems with many partly controllable variables as in the case of general preference learning of music and online system optimization.

The goal in this thesis is therefore to investigate and realize a general setup based on flexible Gaussian processes, applicable in real interactive systems. It should be both flexible and robust in eliciting and representing different perceptual and cognitive aspects, such as preference and cognitive aspects of audio. The aim is therefore is to provide support for different response types, including discrete choices and continuous ratings, suitable for modelling the particular observations originating from different experimental paradigms. To optimize the learning rate of the system, it is the aim to investigate effective paradigms including extended versions of classic forced choice paradigms. To further optimize the learning rate sequential experimental design should also be supported for optimal experimentation in real applications. It should, in a flexible way, support the many representations relevant to audio and music, including multiple heterogenous data sources potentially modeled by probability density functions.

The realized setup [2] is presented in details in section 3, and documented and applied in multiple contribution listed in Sec.1.2.

## Predictability of User Context

The context in which users interacts with systems and objects has a major influence on both the users representation in terms of needs and understanding of the objects and system output. One major aspect of the user context is location (others include social context), which can typically be used to improve the computational representation of the objects of interest.

A intriguing aspect of human context, and in particular location, is to what degree it can be predicted based on the past which determines the so-called predictability. From an engineering viewpoint this has relevance to resource allocation and general system optimization, however, the predictability of location provides a basic view into human behavior by quantifying the repetitive patterns of human life.

[2] Developed in collaboration with co-authors.

The aim of the thesis is to investigate methods for characterizing and examining the predictability of subjects, in particular the location aspect sampled using the ubiquitous mobile phone. The goal is to quantify the fundamental predictability of humans mobility, which characterises each individual user. The methods for accomplishing this are described in Chapter 4 and the dataset and results are presented in [B] and [A].

## 1.2 Contributions

The academic contributions follows the three areas defined above, thus

- The **first** contribution relates to finding semantic representation of music by modelling music data by multi-modal Bayesian topic models which provides a relatively simple, but widely applicably probabilistic model. Contribution [H] is a study of the million song dataset [20] analysing the alignment between top-down open vocabulary tags with the bottom-up representation. It furthermore examines the predictive power of the joint model for genre and style prediction - a classic task MIR task.

- The **second** and primary group of contributions is in the field of ranking, rating and preference learning. The thesis contributes with a number of modelling extensions and applications of a flexible probabilistic setup relying on Gaussian process priors.

  In terms of experimental paradigms relying on relative comparisons between objects/system output, the thesis contributes with the proposal and realization of a new likelihood model for pairwise ratings with continuous observations based on the Beta distribution [C]. The thesis furthermore contributes with a sparse/pseudo-input extension for the classic pairwise model setting. This allows scaling of the pairwise model to larger problems [G] than previously feasible.

  In terms of experimental paradigms relying on absolute ratings; contribution [C] proposes and realize a likelihood model based on Beta and Truncated distributions designed for responses with bounded support.

  In contribution [I] active learning or sequential design is investigated for system optimization, where elements of the setup is applied for active preference learning with absolute (bounded) responses in a real-world and interactive application.

  The modelling part of the individual contributions are realized in a general setup supporting various paradigm for ranking and rating, which has been

applied in the field of music emotion modelling [F][E], music preference [D], and independently by co-authors in audio preference learning [170]. Multiple kernel and generative kernel elements of the setup was furthermore applied in [4] supporting sensor fusion for binary classification.

- The **third** group of contributions is related to the field of human context prediction and in particular one-step ahead predictability of location. The thesis contributes in [B] and [A] with studies relating to and supporting previous work in the field of human predictability relying on nonparametric predictability bounds derived from information theory.

## 1.3 Structure

The report continuous with a general introduction to the three subareas previously identified and outlined. The chapters are not an exhaustive walk-through of every aspects of the methodology and modelling methods, but aims at introducing the reader to the general area in the respective fields and focusing on placing the listed contributions in a broader context of the respective research areas.

- Chapter 2 provides an overview of computational aspects of audio and music with focus on the content based bottom-up view and the users top-down view including related tasks in the field of music information retrieval (MIR). Based on [H] the chapter then describes a setup and model for integration multiple views in a single joint semantic space based on multi-modal topics models, which provides background and motivation for contribution [H] .

- Chapter 3 is an introduction to preference learning, ranking and elicitation of perceptual and cognitive aspects primarily in the audio/music domains based on a Gaussian processes. The chapters takes a holistic view on preference learning with Gaussian process, and considers a general Bayesian setup consisting of four elements: observations, prior, inference and sequential design.

- Chapter 4 describes methods for estimating users predictability based on information theoretic bounds. It gives an motivation for the approach and an introduction to the methods with a simple proof of the bounds applied in [B,A].

- Chapter 5 summarizes and concludes the thesis based on the summary report and contributions related to each of the three focus areas.

The contributions are included as pre-prints in the appendix. They are grouped in the three focus areas as outlined in the introduction,

**Computational Representation of Music**

- Appendix [H] contains a pre-print of the paper: "Towards a universal representation for audio information retrieval and analysis"

**Preference Learning with Gaussian Processes**

- Appendix [C] contains a pre-print of the paper: "Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes" (MLSP 2011)
- Appendix [D] contains a pre-print of the paper: "A Predictive model of music preference using pairwise comparisons" (MLSP2011)
- Appendix [F] contains a pre-print of the paper: "Modeling Expressed Emotions in Music using Pairwise Comparisons" (CMMR2012)
- Appendix [E] contains a pre-print of the paper: "Towards Predicting Expressed Emotion in Music from Pairwise Comparisons" (SMC 2012)
- Appendix [G] contains a pre-print of the paper: "Pseudo Inputs For Pairwise Learning With Gaussian Processes" (MLSP 2012).
- Appendix [I] contains a pre-print of the paper: "Personalized Audio System - a Bayesian Approach"
- Appendix [J] contains a pre-print of the paper: "Bounded Gaussian Process Regression".

**Predictability of User Context**

- Appendix [A] contains a pre-print of the paper: "Estimating human predictability from mobile sensor data" (MLSP2010)
- Appendix [B] contains a pre-print of the paper: "Predictability of mobile phone associations, European Conference on Machine Learning" (MUSE2010)

Chapter 2

# Computational Representation of Music

Music and audio plays an important and large part in the modern society in terms of well-being [51] and commercial importance [95]. One of the challenging issues, from a signal processing and computational point of view, is the many aspects which influences the way a user perceives and understands a particular song or even a subpart of the song. This is both an effect of the complex auditory system and high level cognitive aspects influenced by cultural and personal memory [57]. This makes it both challenging to represent and reproduce music, analyse single songs, organize millions of songs and create music and audio services for retrieval and recommendation.

The field of computational audio and music goes back to at least the fifties, where the CSIRA computer was the first computer to play computer generated audio [54]. The audio (and image) domain acted as perfect application and motivator for the spur of digital signal processing developed in the last half of the last century. The field often came up with new challenges specifically related to music synthesis, reproduction and analysis, here focusing on the latter aspect. Another important aspect of the field is the understanding of the human auditory systems and in particular development of simple perceptual auditory models (see e.g. [243][76]), for example loudness models and mel-frequency filterbanks applied in many modern music analysis systems. This allowed the

computer to analyse the audio in a way similar to the listener.

The combination of digital signal processing and perceptual auditory models allowed the music information retrieval (MIR) field to automate tasks previously limited to human analysis. This included chord recognition, source separation, transcription, tempo estimation, segmentation, best extraction, timbre detection, instrument classification, gender recognition, artist recognition and cover song detection. These tasks are objective and **bottom-up** driven in the sense that they are not depending on subjective representations (experience, cultural background, memory etc.) and can in principle be deduced from the signal and metadata of the song.

While these objective tasks are certainly interesting, this thesis focuses on the **top-down** driven tasks related to organization and recommendation, which are highly influenced by the users representation and understanding. This representation is typically considered a complex result of biological, environmental, cultural background and the current context of the individual, including the social context) (e.g. [52]). Such engineering and system tasks include genre classification, emotion recognition (both expressed and induced), perceived similarity based on timbre, rhythmic, harmonic and melodic aspects with application to for example recommendation, (auto)tagging and preference elicitation.

The mentioned top-down tasks are related to general organization and retrieval systems. In the other setting considered, i.e., reproduction, the top-down task amounts to finding the optimal system parameters (such as filter characteristics) to obtain the best possible alignment with for example the users preference[1].

**Outline**: This section continues with an overview of the many domain specific aspects and object representations of audio and music also relevant to the contributions in Sec.3. First, a brief overview of content based features and representations is provided in Sec.2.2. Secondly, top-down aspects of music and music systems are revised in Sec. 2.3. Sec. 2.4 gives an introduction to the aggregation of several music representations into a multi-modal model based on probabilistic topic models, in particular multi-modal Latent Dirichlet Allocation (LDA).

[1]Such tasks are not considered from a modelling perspective in this chapter but addressed in Chapter 3

## 2.1 The Audio Object: Signal and Metadata

Audio objects, in particular music objects, are here characterized by two overall aspects,

- The signal: A time-domain signal indirectly representing the physical sound pressure level created when a system reproduces the audio from the signal.

- The metadata: (sometimes referred to as the object context) covers aspects objectively connected to the audio/music object. This includes artist, period/year, duration, sampling rate, and possibly some categorization. In context of a system, textual lyrics are also considered part of the metadata, since it is typically available in external form to the time domain signal and not derived from the signal itself.

The time-domain **signal** as illustrated in Fig. 2.1, is the base object considered in audio analysis, processing and reproduction. For the task considered later on, we typically consider basic transforms of the data to for example the frequency domain using the discrete (fast) fourier transform (FFT). Given the non-stationary of music, this and other analysis methods are often based on an equal-length and possibly overlapping analysis frames, significantly shorter that the full signal, and typically less than a second, as illustrated in Fig. 2.1. Alternatively, some analysis approaches uses frames of non-equal length based on the events in the signal such as the Echonest [60]. These analysis frames, regardless of the length, is the basic temporal entity on which the bottom-up feature extraction works

**Metadata** regarding the audio/music object includes purely objective aspects such as origin and period. The texttual lyrics are also considered metadata since it often enters as a separate set of (objective) data and not extracted from the audio. Lyrics has been analyzed in a number of studies, such as [177], finding clear patters in the way the temporal course of the emotion of lyrics change in music. For qualitative classification, lyrics have been evaluated in for example [94] showing that lyrics can aid in bottom-up based audio modeling. Contribution [H] uses lyrics in evaluating genre prediction and alignment with top-down aspects showing that lyrics is not a particular good feature for genre and style classification as compared with audio.

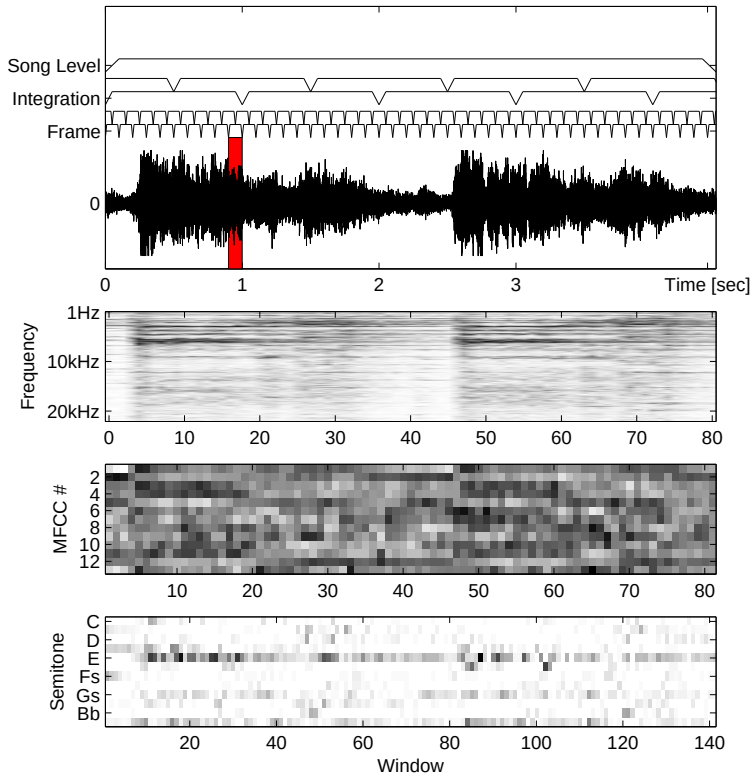



Figure 2.1: The top panel shows the audio signal by its time-magnitude representation. The top panel furthermore shows the temporal frames which is normally used to extract bottom-up features. The shortest frames (in red) shows the basic analysis windows, the intermediate frames illustrates the process of temporal integration in which multiple is used in the estimation of a representation on longer time scale. The longest frame illustrates temporal integration operating on the full signal length. The **second panel** shows the spectrum of the signal. The **third** panel shows the mel-frequency cepstrum coefficients (MFCC) extracted on the basic analysis frame. The **fourth (bottom)** panel shows the chromagram.

| Frequency Domain (see e.g [176][76]) | Time Domain (see e.g. [176][160]) | Unsupervised features | Others descriptors |
|---|---|---|---|
| FFT / DCT<br><br>Energy / ratios<br><br>Rolloff, Flatness<br>Flatness, Centroid<br>Flux, Bandwidth<br>Slope, Spread<br>. . . | Onsets<br>Zero-crossings<br>Envelope aspects<br>Autocorrelation<br>Duration<br>. . . | NMF (e.g. [175])<br>Deep NN ([124][82])<br>. . . | Information dynamics [3]<br>. . . |

| Timbre / Perceptual Spectrum Encoding | Tonal/Hamony/Melody (see e.g. [109]) | Rythmic ([109][160][176]) | Loudness [57] |
|---|---|---|---|
| MFCC [76]<br>PLP [87]<br>LPC [76]<br>. . . | Pitch (e.g. [109][169])<br>Chroma(gram)/PCP [69][15]<br>Chords (e.g. [99])<br>Melody features (e.g. [186])<br>Noisiness<br>Inharmonicity<br>. . . | Pulse / Tatum<br>Beats<br>Bars / Measue<br>Tempo<br>. . . | Loudness<br>Sharpness<br>Spread<br>. . . |

Table 2.1: A non-exhaustive but representative overview of common bottom-up audio features. The top row lists a number of base features and elements used in describing the musical features listed in the second row.

## 2.2 Bottom-Up View

The music analysis and retrieval community has considered hundreds of different content based descriptions of music. Generally these can roughly be grouped into partly musically meaningful categorizes as indicated in Tab. 2.2. Some content based descriptors are purely statistical and other such as loudness and the timbre motivated mel-frequency cepstrum coefficients (MCFF) are related to - or at least motivated by - directly to the human auditory processing.

The thesis applies a variety of the most popular bottom-up features outlined in Tab. 2.2, which is shortly described in terms of the music aspects (the second row in Tab. 2.2).

**Loudness** is the perceptual understanding of strength which allows a person to rank sounds from quiet to loud [76]. It is not a mathematical quality, such as energy and power, and loudness is often calculated based on perceptual models of loudness [159][243]. Loudness can also be considered top-down aspect since it depends on context, however since extracted from the audio, we here consider it a part of the bottom-up view.

Explicit loudness features are considered in contribution [H], based on the Echonest feature set [60].

**Timbre** is the quality of music which allows human to differentiate between different instrument playing the same note with same pitch and loudness [193]. The mel-frequency cepstral coefficients (MFCC) is a computational approach [133][153][76][56] to extract at least some aspect of this quality [213][9] by encoding the spectrum in a perceptual way. The extraction on each analysis frame is (typically) performed as follows [56][76]:

$$\text{Windowing} \rightarrow \text{FFT}(\cdot) \rightarrow \text{Abs}(\cdot) \rightarrow \log(\cdot) \rightarrow \text{Mel-FilterBank} \rightarrow \text{DCT}$$

where the FFT denotes the Fast Fourier Transform, and DCT denotes the discrete cosine transform. The main aspect to consider is the Mel frequency filter bank which aims at performing spectrum analysis in line with the human auditory system [76]. Various implementations of MFCCs such as [33][38][219] allows for different filter banks, windowing function and importantly allows for different number of filter bands distributed across the full frequency range. Furthermore, an arbitrary subset of the resulting coefficients is typically parsed on to the model stage, which altogether makes the notion of MFCC a vague concept, highly dependent on implementation and application.

Other features carry information timbre and a large number of both temporal and spectral aspects has typically been included in the aim to describe timbre, such as spectral flux and zero-crossing rate (ZCR).

In contributions [D][E][F] standard timbre features was applied (MFCCs, ZCR and spectral descriptors), whereas contribution [H] used the Echonest version of timbre, which are based on a linear projection of MFCC-like features into a 12 dimensional subspace [60].

**Tonal and Harmonic**

**Pitch** is defined [109] as "a perceptual property that allows the ordering of sounds on a frequency-related scale". Is not the same as the fundamental frequency in spectrum analysis, since the auditory system highly influences the pitch perception [57]. However, the actual fundamental frequency is often applied as a proxy for pitch.

**Chroma** (or pitch class profile) features has become a popular representative for the tonal and harmonic content of music (see e.g. [61]). The chromagram is based on the chromatic scale, thus the representation consist of twelve bin frequency spectrum. Each bin contains the aggregation of energy all bins in the fullrange frequency spectrum which are closest to the note defined by the twelve bins invariant to a particular octave. Hence, it is a twelve dimensional representing of the energy/intensity of the twelve pitch classes. It is a coarser representation than the pitch itself or fundamental frequency, but is also more robust in terms of estimation

and noise. The chroma(gram) is typically (see e.g. [160]) extracted in a frame based manner based on a set of constant-Q filters per octave (12, i.e. 1 semitone per bin, or 36, i.e. 1/3 semitones per bin).

Contribution [I] makes use of the chroma features from the Echonest [60] implementation, however beat aligned to obtain a music meaning full temporal alignment - and furthermore normalized per beat segment in line with previous work [18].

The **Melody** is an even higher level representation and focuses on the sequence of pitch and chords and can simply be defined, as follows: "Melody is the dominant individual pitched line in a musical ensemble" [173]. As noted in [186] it may be considered a top-down aspects as it is cultural and context dependent.

**Rhythmic**

Rhythm in its most general form refers to the temporal aspects of music such as tatum, beats and bars [77][161]. Often we consider the primary or generic tempo, i.e. the rate at which a standard listener would tap the foot when listening to music. This is typically represented as the number of beat per second. Tempo is applied in contribution [H].

### 2.2.1 Song Level Representation

The features described above can often (alone or combined) be considered a sequence of vectors in some high-dimensional space which constitutes the representation of the song on the frame level. However, for many modeling we are interested in a representation on the song level so this section will outline standard ways of finding a song level representation.

**Pre-processing** is often applied in order to increase the interpretability of the frame based representtaion, or ease the computational load of the often high-dimensional feature spaces and/or large music databases. A large variety of methods can been applied in the audio domain for reducing the dimensionality, such as the pallet of multi-dimensional scaling (e.g. ISOMAP, Laplacian Eigenmaps), however, the primary preprocessing tools is the simple principle component analysis (PCA) [26]. It is typically calculated via the SVD decomposition and was applied in contribution [D,F] as preprocessing with the aim of reducing the dimensionality defined by common timbre features.

**Audiowords** is a term used to describe a representation in which the feature space is quantized into a number of prototypical audio words [196][91][210][146]. Based on a vocabulary of audiowords, $\mathcal{V}$ of size $V \times 1$, each frame in all the

audio songs can be assigned in a hard manner to a single audioword corresponding to the well-known technique of vector quantization (VQ). In a song, $s$, a particular word from the vocabulary of audiowords at position/frame $i$ is denoted, $\mathbf{w}_{s,i}$, and the sequence becomes a vector of integer indexes, i.e., $\mathbf{w}_s = \left[w_{s,1}, w_{s,1}, .., w_{sN_s}\right]^\top$. There are various strategies towards finding the audiowords, and possibly the simplest is to apply a K-means algorithm [26] with a fixed number of K centers. Obviously any (spectral) clustering (hard or soft), sensible decomposition (such as normalized matrix factorization) may be used to define the audio words (followed by hard assignment). In contribution [H] an online version of the K-means algorithm [106] was applied for scalability on the Million Song Dataset (MSD) [20].

**Temporal considerations:** Many of the bottom-up features outlined above - or their low-rank and vector quantized version - are typically based on the frame based analysis as outlined in Fig. 2.1, thus a sequence of vectors. The obvious temporal aspect of audio has spurred a vast amount of work in representing and modelling temporal. The approaches can roughly be categorized as follows:

**Temporal Independence - bag-of-frames** In the simplest possible - but widely applied approach - we assume that the frames are independent and obtain a so called bag-of-frames approach.

- Mean-Variance / Gaussian:
  A simple statistical representation of a set of vectors in a song, $s$, is the multi-dimensional mean vector, $\boldsymbol{\mu}_s$, and variance, $\sigma_s$, of the multi-dimensional observations. This can naturally be generalized to a probability, and given the often continuous features, the natural distribution is a standard Gaussian, i.e.the representation for a song, $s$ is simply
  $$p\left(\mathbf{x}|\boldsymbol{\theta}_s\right) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)$$
  where $\boldsymbol{\theta}_s = \{\mu_s, \boldsymbol{\Sigma}_s\}$. $\mu_s$ is the mean of the distribution and $\boldsymbol{\Sigma}_s$ is the covariance matrix. While seemingly simple, this representation has been argued to be rich enough [143], at least for classification purposes.
- Gaussian Mixture Model
  The single Gaussian representation is possibly enough in many situation and systems [143], however another popular approach is to generalize the Gaussian representation to a mixture of individual Gaussian using the Gaussian Mixture Model (GMM) to define a considerably more complex density, i.e.,
  $$p\left(\mathbf{x}|\boldsymbol{\theta}_s\right) = \sum_{z=1}^{Z} p\left(z\right) p\left(\mathbf{x}|\boldsymbol{\theta}_s^{(z)}\right) = \sum_{z=1}^{Z} p\left(z\right) \mathcal{N}\left(\mathbf{x}|\mu_s^{(z)}, \boldsymbol{\Sigma}_s^{(z)}\right)$$

where $\boldsymbol{\theta}_s = \left\{\boldsymbol{\theta}_s^{(z)}\right\}^{z=1:Z}$ and $\boldsymbol{\theta}_s^{(z)} = \left\{\mu_s^{(z)}, \boldsymbol{\Sigma}_s^{(z)}\right\}$. The mixing proportions, $p(z)$, further has the constraint $p(z) \in [0,1]$ and $\sum_{z=1}^{Z} p(z) = 1$. The model is typically estimated using standard expectation maximization algorithm [26]. Determining the model complexity is generally a tricky matter and may be performed using the Akiies Information Criterion (AIC) or the Bayesian Information Criterion (BIC) [190][26], penalizing the resulting log-likelihood with a complexity term depending on the number of free parameters in the model.

- Histograms for audio word models:
  In case the audio feature space has been quantized into audio words, we may represent the song as a sequence of the audiowords. However, in the bag-of-frames assumption the order does not matter and the song may be represented counting the number of occurrences of the audiowords in the vocabulary. This may be expressed as counts of (audio)words in a given song, $n(\mathbf{w}_s = v)$ where $v$ is the index into the vocabulary, $\mathcal{V}$. This results in a bag-of-(audio)words representation where each song is represented by a $\mathbf{x} \in \mathbb{N}^V$ vector of counts $\mathbf{x}_s = [n(\mathbf{w}_s = 1), n(\mathbf{w}_s = 2), ..., n(\mathbf{w}_s = V)]^\top$

  This bag-of-frames representation obviously requires less estimation and computation than the GMM, and was used for scalability reasons in contribution [H] on the MSD. In some settings, the vector of counts may be represented as a probability mass function, i.e., as a distribution over all words for a given song. This is implicitly done in for example [7] for a topic model, and [146] for use in a kernel method.

**Temporal Integration** A number of approaches has been suggested aiming to integrate the low-level short time features into a longer frame in effect integrating temporal information into new features and representations which may then be applied in a algorithm. This approach can often be divided into a number of categories depending on what temporal level an subsequent algorithm makes its decision

- Early: Integration of information from each analysis frame can in the simplest form be conducted by stacking individual frames. A more advanced approach is the multi-variate auto-regression (AR) modelling across multiple analysis frames, resulting in intermediate level frames with for example AR coefficients (see [150] for an overview).
- Late: Another formally not a bottom-up approach is to make decisions for each analysis frame and subsequently make a single joint decision based for example on majority voting [150].

**Temporal Modelling** The most elaborate approach to include temporal information is a 'real' modelling of the temporal dynamics which can be done using for example hidden Markov models, or the more suitable dynamic texture mixture model [12][11].

The **song level representation** applied in the contributions is typically the Gaussian Mixture Model. There exist many proposals and evaluations [98][9] of various measures of unsupervised similarity measures between densities. Based on a single feature vector - and possibly the mean of the Gaussian - simple distance measure such as the Euclidian, cosine or Mahanalobis distance may be applied to define similarity between audio. More elaborate and popular similarity measures include the (Symmetrized) Kullback-Leibler (KL) divergence between densities. This can be computed in closed form with single components, however must be evaluated by stochastic simulation for general mixtures. The Earth Mover distance [134] is an approach to get around the KL divergences problem with more than one component. The Hellinger distance [98] is an alternative measure between distributions, which can be generalized to mixtures in analytical form [97]. A particular intriguing approach is based on the Bayesian, non-parametric Hierarchal Dirichlet Process approach [90] effectively encoding all songs with a a-prior assumption of infinite number of Gaussians common to all songs. Each song is then coded as a mixture of these common Gaussians, and similarity is given by the mixing proportions corresponding to $p(z)$.

**Tasks** which rely on such music representations and similarities, include purely bottom-up driven tasks like content based recommendation. However, for the specific task of mapping to some label, e.g, genre, we often turn to supervised machine learning algorithms (including the ones considered in Chapter 3), which are based directly or indirectly on metrics or similarity functions. Well-known algorithm include K-nearest neighbor [26] and kernel machines [197][26] such as the Support Vector Machine (SVM), widely popular in the MIR community. There is in general no requirements for the similarity function used in the K-nearest neighbor algorithm, however any algorithm based on kernels, formally require the kernel to be positive semi definite (PSD) (disregarding the notion of conditional PSD kernels). Provided that the defined similarity is a valid metric (e.g. [84]) it may directly be converted into a valid kernel function [81] [2].

The kernel based algorithms provide flexible alternatives to vector space algorithms, since the only requirement is a valid kernel (perhaps derived from a valid metric), which can be formulated for many objects such as strings, text and distributions. In particular, contribution [D] makes use of the probability product kernel (PPK) [97] (described in more detail in Sec.3) for defining correlation between the bottom-up audio features in terms of each songs density given by

[2] As considered in the co-authored paper [4]

the GMM. The PPK has gained some attention in the MIR community and has been applied in [152][150][14] with GMM and MAR, and most recently in [146] using a histogram representation. Furthermore, contribution [F,D] utilizes the PPK based on the Gaussian Mixture Model representation. Alternatives in the MIR community includes the symmetric Kullback-Leibler divergence based kernel (see e.g [151]), which is formally not a PSD kernel. And the much simpler option of applying a squared exponential on a vectorized version of the Gaussian representation (or the mixture) is often a simpler alternative. Such a representation is applied in contribution [E].

## 2.3 Top-Down View

The bottom-up view as defined by the content based analysis and metadata, has been a primary driver of music similarity in the early years. However, in recent years many successful music studies and services are based not on the content itself, but on the ratings and opinions of users, in a so called collaborative filter setups. This is the text book example of a purely top-down based system, where the users expressed views defines the possibly subjective similarity between music songs.

Such collaborative filters are certainly not the only form of user generated data suitable for including in a top-down view of music and audio systems. A number of such aspects (and related datasets) has been considering in the music information retrieval and analysis community), and we here provide an overview of the approaches and relevant studies.

**Preference - ratings and rankings**

Subjective audio and music preference is a major aspect in recommendation and understanding of user representation and needs. Such preference ratings is the basis of collaborate filtering systems (see e.g. [43] for a general overview) and has been investigated in numerous studies such as [42].

Collaborative filters usually suffers from very sparse data and have in the MIR field been found to be biased by popular artists and popularity in general [43]. Another approach is a fully personalized system which was considered in contribution [D], where we proposed and investigated a paradigm, where users compare two music songs to elicit their music preference based on partial rankings of the music objects.

Audio preference was also considered in contribution [I], which attempts to efficiently eliciting the preference in a music reproduction system where

the music signal is altered by a linear filter operation parameterized by system settings. The system task is to optimize the system parameters based on a internal representation of the users preference.

**Listening patterns**

Listening patterns of users is an indirect preference rating method, where the frequency is a representative of preference. This is, however, prone to popularity as outlined in [43]. The million song dataset and the associated TasteProfile dataset [2] - with more than one million users and more than 48 million entries - offers the possibility of evaluating the usefulness of listening patters for preference learning and recommendation.

**Music emotion - expressed and induced**

That music has an emotional effect on humans is hardly a surprise and probably one of the reason why music is such an important part of the modern life. Naturally, audio and music systems should take this important aspects into consideration, and in doing so, it is custom to differentiate between the expressed emotion and the induced emotions of the music.

**Expressed** emotions of music is the objective emotional expression the music is believed to carry, i.e., it can be seen as the expression the composer is aiming to express [104]. Aimed to be an objective evaluation of the emotion which the music expresses; it is still influenced by the cultural background of the listeners [89][103]. The **induced** emotions refer to the affect that the music has on the listener personally, i.e., highly dependent on internal representation (memories, experience etc) and context. Whether the distinction between expressed and induced emotion is feasible is certainly an issues [183][104], however not often discussed in the MIR community where the focus has been on the expressed emotion, thus, disregarding the affect the music has on the individual and focusing on the (homogeneous) population used to examine it.

The experimental setup typically applied for examining the expressed emotion are categorical approaches [88] or dimensional approaches [184]. The dimensional approach is the most applied in the MIR field [108], and the preferred dimensions are arousal and valance denote the AV space. Valance spans a range from highly positive (happy) to highly negative (sad), whereas arousal ranges from calm/passive to excited/active [184].

From an application point of view, the aim is to model and predict the emotional expression in the AV space based on the bottom-up features which can then be used in recommendation and retrieval systems. A number of approaches has been proposed in the MIR community (see [108] for a review). These are mainly based on so-called direct scaling methods in which users are asked to assign a single absolute value on the AV scale, representing the emotion expressed by a particular piece of music. This is often highly susceptible to the users lack of understanding of the scale

[140], due to the complex concepts of arousal and valance. This is in addition with the cognitive difficult in separating the expressed emotion from the induced.

Contribution [E,F] consider the task of predicting the (expressed) emotion of music based on pairwise comparisons between individual songs in the AV space, which is an alternative but more robust experimental paradigm than absolute rating. It has not yet fully been explored in the MIR community, although previously considered in [237].

Induced emotion is highly personal and even context dependent and has seemingly not been investigated within the MIR field, possibly due to the experimental difficulties in obtaining robust ratings and representations [104]. It is, however, crucial to include such knowledge into truly personal systems where individuals are affected differently by the music. This personal aspect should optimally be elicited and represented by a given system.

**Annotation - categories and tags**

**Categorization** of music, such as genre [68] is often a result of cultural understanding and grouping of music [52], and is therefore considered a top-down aspect in this context. The task of classifying music into genre based on bottom-up features has been a defining task of MIR in many years (see [209] for a comprehensive review).

**Fixed vocabulary** based on a certain taxonomy/ontology, is a simple form of tagging where genre based annotation can be seen as a special case. However, while genre is typically considered unique and exclusive, the general annotation setting allows multiple annotations per objects (and possibly users). Examples of such fixed annotations in include [221] where the vocabulary is based on the CAL500 vocabulary [39], and a recent proposal of an ontology for music annotation [181].

**Open vocabularies** also known as a folksonomies, provides a setting, where the annotation themselves is the result of a conscious decision to freely express the individual user representation towards an object or system result (see [118] for a general review). Thus, users are allowed to enter free text or even sentences expressing their view. A number of research and real-world datasets has been collection in the MIR field, including Magnatagatune [120], MajorMinor [142], last.fm [1] and CAL500 [39]. Descriptive studies of open vocabulary annotations, include [125] using probabilistic latent semantic analysis (pLSA), finding clear semantic patterns. Similarly [72] uses open vocabulary tags from last.fm to evaluate the similarity between artists also based on last.fm. [203] examines the alignment between genre and tags based on a fixed expert vocabulary and last.fm. Prediction of these tags from e.g. the audio features is a common task in MIR (e.g. [155][19][91]), often motivated from a recommendation

and retrieval point of view [43] in which the tagging induces a particular top-down similarity.

Contribution [I] considers open vocabulary tags based on last.fm dataset. This is combined with the bottom-up view to show alignment between representations and tags in a descriptive fashion.

**Music similarity - explicit and implicit**

**Explicit** ratings of similarity of music objects is not a common approach in the MIR community, but has gained some momentum with the MagnaTagAtune dataset [120]. Various approaches [207][234][235] been applied to learn the similarity based on the odd-one-out experimental paradigm, i.e., selecting the object which is most dissimilar to the two others [3]. A similar approach was formulated in [62] considering ground truth for artist similarity.

Other approaches to similarity include explicit relevance feedback, where users indicate relevance of the returned search result. This as for example been proposed in [110] in which the user's query is altered based on indicated relevance, and in [45] reweighing feature dimensions based on indicated relevance of search results.

The direct scaling of similarity is conjectured to be an important modelling aspect in future systems, since it provides a direct way to manipulate computational representations by expressed distances, or by e.g. odd-one-out paradigms [120], possibly in an interactive learning process.

**Implicit** ratings based on the general techniques of (implicit) relevance feedback [10] amounts to observing the behavior of the user in relation to a system output, e.g., click-through behavior of music search. Such feedback implicitly defines similarity between query and objects. Examples of such ideas has been proposed in for example [174] examining the skipping behavior in recommended music playlists .

**Common knowledge and Web resources**

An particular aspect of the top-down view is common knowledge, i.e. knowledge which has been collectively agreed upon such as encyclopedias and reference works. This covers some of the annotation aspects previously mentioned, at least to some degree. Wikipedia, for example, contains a detailed description of genre and other music related aspects useable as general top-down information. Such an approach was for example investigated in [157] examining content and links in Wikipedia's music universe - and recently in [204], anchoring the patters found in an analysis of last.fm using Wikipedia's music universe. Other online sources of information are

[3] The framework presented in Sec. 3 can be used to model this dataset in a probabilistic manner unlike e.g. [235]

Twitter and similar services, which can provide top-down information regarding the objects, such as the preference covered above. It can further give detailed insight into the users general context, including activities and state of mind. Online social network such as Facebook and Twitter can be a potential source of top-down and contextual information in collaborative filtering settings, inline with the friend-of-a-friend (FOAF) ideas presented in [42][43].

## 2.4 Joining Views

The tasks related to the various views typically focuses on mapping from a bottom-up view to the top-down view in order to utilize the bottom-up (or content based) analysis to enrich the users in a given system. Such task include supervised auto-tagging [91][19], unsupervised similarity computation based on bottom-up similarity (e.g. [98]), supervised genre classification based on bottom-up features [209] and unsupervised (with auxiliary task) tag analysis [126]. Here, the focus is on the combination of views - or modalities. Modern music organization systems and services, such as last.fm, typically have access to (multiple) bottom-up and top-down views, but a major challenge is to utilize either, or both, to create an effective computational representation, which aligns well with the (individual) user representation as argued in Chapter 1.

In recent years, this has spurred a large interest in multimodal integration of views - or modalities - for a number of tasks typically within retrial and recommendation domain. Particular focus has been on hybrid recommendation based on collaborative filtering and content based features (e.g. [240]), discriminative semantic retrieval by combining kernels [14] and playlist generation by applying a hyper-graph approach, using many different modalities and learning the weight of each modality discriminatively [148]. Recently [149] applied multiway canonical correlations analysis (CCA) based on bottom-up audio features, lyrics and tags to analyze emotional aspects of music based on the Million Song Dataset [20].

A further advantage of including multiple bottom-up and top-down views in a single suitable model, is the possibility for a the single model to solve multiple tasks previously covered by many different algorithms. The work in [229] is a proposal for such a multi-task system in the MIR field. Contribution [H] is an proposal of a probabilistic and generative approach towards such a system providing a relatively simple probabilistic approach. It integrates the bottom-up view (the audio features), the meta-data (lyrics) and a top-down view based on open vocabulary tags.

### 2.4.1 Example: Multimodal Integration

A concrete example where multimodal integration is an natural and often required approach is large scale music modelling which has become possible due to the release large linked datasets such as the million song dataset (MSD). In contribution [H] a setup is proposed which considers a number of the already outlined views or modalities, in particular

**Audio**
The audio content is represented by bottom-up analysis. In particular four fundamental aspects covering all of the musical aspects outlined in Sec.2.2. I.e. timbre ($\mathbf{x}^{(timbre)} \in \mathbb{R}^{12}$), rhythmic ($\mathbf{x}^{(tempo)} \in \mathbb{R}$), harmonic ($\mathbf{x}^{(chroma)} \in [0,1]^{12}$) and loudness ($\mathbf{x}^{(loudness)} \in \mathbb{R}^{1}$). The features are extracted by Echonest and already available in the MSD, however, the setup aligns the event based segment to the beats of the song in order to obtain music meaningful analysis frames (songs with no defined beat maintain the segment based analysis frames).

The million songs in the MSD poses a computational problem due to the mere size of the dataset. Contribution [H], therefore, takes the audioword approach outlined in 2.2 in which the common feature space is vector quantized (VQ) and frames in each song assigned to a single audioword.

The song level representation is obtained by assuming that the frames are independent, i.e., no temporal integration is performed in [H]. This means that each of the million songs is represented by a count of individual audiowords. In contribution [H] the combined audio vocabulary is of size $V^{(\text{tags})} = 2112$ audio words.

**Lyrics**
The textual lyrics of a song is another fundamental modality, which is integrated in the setup. The lyric representation is in a bag-of-(lyric)-words similar to the audiowords. The lyric vocabulary is of size $V^{\text{lyric}} = 5000$ ( available in the MSD).

**Tags**
The user generated top-down is based on open vocabulary tags originating from last.fm. They are represented as a bag-of-words of size $V^{(\text{tags})} = 20.000$ including the most popular tags out of more than 500.000 unique tags.

**Evaluation**
The setup uses other top-down information which is not included in the representation itself. These are genre (15 categories/classes) and style

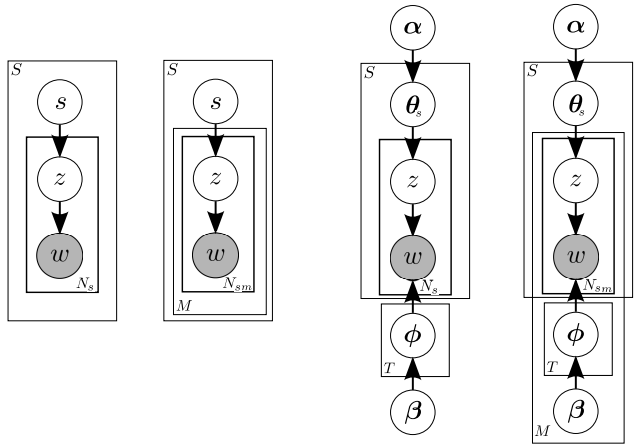



Figure 2.2: Graphical model (plate notation) of the considered versions of pLSA (left) and the two different LDA models (right).

(22 categories/classes) originating from the All Music Guide dataset associated with the MSD [189]. This categorizes are simply represented by$V^{(\text{genre})} = 15$ and $V^{(\text{style})} = 22$ word vocabulary, respectively.

The particular bag-of-words type representation allows for a symmetric and homogeneous representation of all modalities, which is advantages for scaling the systems and allowing for standard computation representations/models. The computational representation used to join these views, is in the current setup based on probabilistic topic modelling originally formulated as pLSA [92], which naturally leads to the multimodal Latent Dirichlet Allocation (LDA) formulation applied in contribution [H].

### 2.4.2 Multimodal Latent Dirichlet Allocation

Modelling text in terms of latent grouping - or topics - is a popular focus area of machine learning, with one of the most popular method being the probabilistic latent semantic analysis (pLSA) originally propose in [92]. For a song, $s$, the observation is a song-word vector, $\mathbf{x}_s \in \mathbb{N}^V$, of counts of each word, $v$, in the vocabulary, $\mathcal{V}$, originating from the sequence of words, $\mathbf{w}_s = [w_{s,1}, w_{s,2}, ..., w_{s,N_s}]$. pLSA assumes that the occurrence of a particular word $\boldsymbol{\omega}_v \in \{0, 1\}$ with index $v \in [1 : V]$ in the vocabulary $\mathcal{V}$, can be modelled as a mixture over socalled topic distributions, $\hat{\boldsymbol{\theta}}_s = p(z|s)$, and a topic specific word-topic distribution

$\hat{\boldsymbol{\phi}}_z = p\left(\boldsymbol{\omega}_{v=[1:V]}|z\right)$ Hence, the probability for observing a vocabulary word in a particular song, $s$, is given by $p\left(\boldsymbol{\omega}_v = 1|s\right) = \sum_{z=1}^{Z} p\left(\boldsymbol{\omega}_v|z\right) p\left(z|s\right)$.

Hence, the pLSA model defines $Z$ latent groups of words also referred to as topics. The graphical model of this simple mixture is shown in Fig. 2.2, which shows the underlying generative story (effectively resulting in the mixture), that for all words with position $i$ in a particular song a topic $z_{s,i}$ is drawn from $p(z|s)$ after which a word $w_{s,i} \in [1:V]$ is drawn from $p(\boldsymbol{\omega}|z)$. We of course require $\sum_{v=1}^{V} p\left(\boldsymbol{\omega}_v|z\right) = 1$ and $\sum_{z=1}^{Z} p\left(z|s\right) = 1$, such that the draw is from a Categorical distributions.

Multiple modalities with each their vocabulary, $\mathcal{V}^{(m)}$, is modelled by assuming conditional independence between modalities,i.e.,

$$p\left(\boldsymbol{\omega}_v^{(1)}, \boldsymbol{\omega}_v^{(2)}, ..., \boldsymbol{\omega}_v^{(M)}|s\right) = \sum\nolimits_{z=1}^{Z} p\left(z|s\right) \prod\nolimits_{m=1}^{M} p\left(\boldsymbol{\omega}_v^{(m)}|z\right) \tag{2.1}$$

This results in a modality specific word-topic distribution $p(\boldsymbol{\omega}^{(m)}|z)$. Estimation of the model parameters, i.e. the point estimates of all the vectors denoted $\hat{\boldsymbol{\theta}}_s = p(z|s)$ and $\hat{\boldsymbol{\phi}}_z = p(\boldsymbol{\omega}|z)$, is done by observing that given a song and a modality, words are generated independently. Thus, the total data likelihood for a sequence of observed words, $\mathbf{w}_s^{(m)}$, in a modality is simply a function of the number of counts of each word indexed by $v$, thus adhering to the bag-of-words representation. The likelihood of a corpus with counts, $\mathbf{X}$, consisting of $S$ songs with $M$ modalities, is typically performed using maximum likelihood based on the total data log likelihood given as, $p\left(\mathbf{X}\right) = \prod_{s}^{S} \prod_{m}^{M} p\left(\mathbf{x}_s^{(m)}\right)$ due to the conditional independence between word and between modalities. Hence, the log likelihood becomes,

$$\log \mathcal{L}\left(\theta; \mathbf{X}\right) = \sum_{s}^{S} \sum_{m}^{M} \mathbf{X}_{s,v.m} \log \left( \sum_{z}^{Z} p\left(\boldsymbol{\omega}^{(m)}|z\right) p\left(z|s\right) \right)$$

where $\mathbf{X}_{s,v.m} = n(\mathbf{w}_s^{(m)} = v)$ denotes the count of the modality specific vocabulary words in each song, i.e. the bag-of-words representation. $\boldsymbol{\theta}$ denotes all the model parameters. The likelihood can be maximized by standard tools such as the Expectation Maximization (EM) algorithm easily derived, also for the multimodal version, and suggested in the original pLSA formulation [92]. Another, but equivalent approach is based on the easy recognizable relation to Non Negative Matrix factorization (NMF) [58] applied for music analysis in for example [7][157][4]. The equivalence is based on the direct use of Kullback-Leibler divergence as the NMF objective (as opposed to the 'indirect' use in the EM case) to derive the multiplicative updates [123][58].

[4] Although with the least-squares objectives and an appropriate normalization

Contribution [H] extends the popular (multimodal) pLSA model by considering the Bayesian version for music modelling, namely the Latent Dirichlet Allocation (LDA). LDA was introduced in [29] with the aim to provide a valid generative model for text documents/songs, which is formally lacking in the pLSA model [29], since there is no way of generating a new document, i.e. obtaining a topic distribution for it.

The LDA model accomplishes this by considering the topic distribution, $\hat{\boldsymbol{\theta}}_s$, a real random variable and not a fixed parameter given the song, as in the pLSA model. With the topic distribution being a random variable, $\boldsymbol{\theta}_s$, LDA specifically places a common Dirichlet prior over it. The generative story now holds, since a new document may be generated by drawing a topic distribution from this common prior. Now it is possible draw a particular topic , $z_{s,i}$, for a word at position $i$ in song $s$. Secondly, in the smoothed extension of LDA, a Dirichlet prior is also introduced over the word-topic distribution, such that with a particular topic $z_{s,i}$, a word at position $i$ is draw from $p(\boldsymbol{\omega}|\boldsymbol{\phi}_{z_{s,i}})$, a categorial distrbution. The graphical model for the smoothed LDA model is shown in Fig. 2.2.

The multimodal LDA, mmLDA, is a straightforward extension of standard LDA topic model [29], as shown in Fig. 2.2. The model is easily understood by the way it generates a new song by the different modalities. The generative process, which defines the model, is given by

- For each topic $z \in [1; Z]$ in each modality $m \in [1; M]$
  Draw $\boldsymbol{\phi}_z^{(m)} \sim \mathit{Dirichlet}(\boldsymbol{\beta}^{(m)})$.
  This is the $z^{th}$ topic's distribution over vocabulary $\mathcal{V}^m$ in modality $m$.
- For each song $s \in [1; S]$
  - Draw $\boldsymbol{\theta}_s \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$.
    This is the $s^{th}$ song's distribution over the topics $[1; Z]$.
  - For each modality $m \in [1; M]$
    - For each word position in the song $i \in [1; N_{sm}]$
      - Draw a specific topic $z_{s,i}^{(m)} \sim \mathit{Categorical}(\boldsymbol{\theta}_s)$
      - Draw a word $w_{s,i}^{(m)} \sim \mathit{Categorical}(\boldsymbol{\phi}_{z^{(m)}}^{(m)})$

The important aspect to notice is that each song has a particular distribution over topics drawn from the common Dirichlet prior, and that each modality has its own word-topic distribution. The joint probability of observations, $\mathbf{w}_s$, and the latent variables, $\mathbf{z}_s$, for each song and modality in the full corpus $\mathbf{W}$ and $\mathbf{Z}$

can thus be written

$$p\left(\mathbf{W},\mathbf{Z},\mathbf{\Theta},\mathbf{\Phi}|\alpha,\beta\right) = \tag{2.2}$$

$$\prod_{s}^{S} p\left(\theta_s|\alpha\right)\left[\prod_{z}^{Z}\prod_{m}^{M} p\left(\boldsymbol{\phi}_z^{(m)}|\beta\right)\left[\prod_{i}^{N_{sm}} p\left(\mathbf{W}_{i,s}^{(m)}|\boldsymbol{\phi}_{z=\mathbf{Z}_{s,i}}^{(m)}\right) p\left(\mathbf{Z}_{s,i}|\theta_s\right)\right]\right] \tag{2.3}$$

where $\mathbf{\Theta} = \{\theta_s\}^{s=1:S}$ and $\mathbf{\Phi} = \left\{\boldsymbol{\phi}^{(m)}\right\}^{m=1:M}$ denotes the model parameters. Inference of these model parameters is intractable and contribution [H] resorts to Gibbs sampling. This is done by first collapsing the likelihood, i.e.,$p\left(\mathbf{W},\mathbf{Z}|\alpha,\beta\right) = \int\int p\left(\mathbf{W},\mathbf{Z},\mathbf{\Theta},\mathbf{\Phi}|\alpha,\beta\right) d\mathbf{\Theta} d\mathbf{\Phi}$. Noting that the Dirichlet prior is conjugate to the Multi-nominal (and hence the Categorical)) distribution; it is after some manipulating found that the Gibbs procedure yields [5] updates close to and generalizing the original LDA Gibbs updates [78]. The Gibbs sampler then amounts to sampling the topic assignment for each observed word [78] in each modality based only the counts of the individual words, not the sequence, hence still adhering to the a bag-of-words representation as in pLSA. The hyperparameters, $\alpha$ and $\beta$, controlling the concentration of the Dirichlet priors, is in contribution [H] estimated by optimizing the marginal likelihood [228].

**Evaluation** of the resulting model is in [H] performed by taking the best sample from the Markov chain (after burn-in), based on the marginal likelihood. Evaluating the performance of topic models is somewhat of a problem (see e.g. [227] for a discussion). It has for example been argued that only a subpart of the test song observation should be used for estimating $p(\boldsymbol{\theta}_{s*}|\alpha)$, in order to avoid over estimating the predictive log-likelihood. In contribution [H] folding is performed on the full set of words in the test song, $s^*$, by implicitly fixing the training versions of $\mathbf{\Phi}$ and $\Theta$ based on the 'MAP' sample. Then sampling of the topic assignments for the test song is performed with the other parameters fixed. This allows estimation of the topic distribution, $\boldsymbol{\theta}_{s*}$, for the test song by sampling and (re)assigning topics to each word in the test song in all the modalities. A 'MAP' estimate of the test song may then be obtained by again selecting the best sample from the 'test' chain. A point estimate, $\hat{\boldsymbol{\theta}}_{s*} = p(z|s^*)$ and $\hat{\boldsymbol{\phi}}_z^{(m)} = p(\boldsymbol{\omega}^{(m)}|z)$, is then obtained by taking the expectation of the corresponding Dirichlet distributions.

**General alignment** between the latent sematic structure - as defined by the topic distribution - and top-down and interpretable aspects is an important element in evaluating topic models for real-world applications. In music retrieval and prediction, tagging models such as [91] are often evaluated based on pre-

[5] It is beyond this report to derive the Gibbs updates, however, the derivation closely follows the standard derivation (see e.g. [8])

cision/recall and mean average precision [144], which focuses on the number of correctly retried tags or possibly songs based on query tags. In contribution [H], a different approach is suggested which aims at evaluating the general alignment of the computational representation with a given (not included) view/modality, $m^*$. The suggested approach makes use of the mutual information between the computational representation given by the LDA model, in particular $p(z|s^*)$ and $p(\boldsymbol{\omega}^{(m)}|z)$. Given an auxiliary task/modality, $m^*$, we calculate the mutual information between individual words $\boldsymbol{\omega}_v^{m*} \in \{0,1\}$, indicating the occurrence of a particular word in the test modality or not, and the topic indicator variable $z \in [1:Z]$. Thus,

$$\mathrm{MI}\left(\omega_v^{(m*)}, z|s^*\right) = \mathrm{KL}\left(p\left(\omega_v^{(m*)}, z|s\right) || p\left(\omega_v^{(m*)}|s\right) p\left(z|s\right)\right) \tag{2.4}$$

$$= \sum_{\boldsymbol{\omega}\in\{0,1\}} \sum_{z=1}^{Z} p\left(\omega_v^{(m*)}|z\right) p\left(z|s^*\right) \log \frac{p\left(\omega_v^{(m*)}|z\right)}{\sum_{z'} p\left(\omega_v^{(m*)}|z'\right) p\left(z'|s^*\right)} \tag{2.5}$$

The probability of the particular word not occurring, i.e., $\boldsymbol{\omega}_v^{m^*} = 0$, is simply $p\left(\boldsymbol{\omega}_v^{(m^*)} = 0, z|s^*\right) = 1 - p\left(\boldsymbol{\omega}_v^{(m^*)} = 1, z|s^*\right)$.

The mutual information is normalized and averaged over all songs, thus the final measure of alignment is the average mutual information, $\mathrm{avgNMI}(\boldsymbol{\omega}_v^{(m^*)})$, for each test word $\boldsymbol{\omega}_v^{(m^*)}$.

This relatively simple approach supports evaluating the alinement of a representation with auxiliary top-down views by ranking the auxiliary words by avgNMI, thus finding the test words which are best explained by the representation. This is for example analysed in [H], where the alignment of a multimodal music representation (based on bottom-up features) is evaluated against the top-down view (as represented by open vocabulary tags).

### 2.4.3 Discussion and Extensions

The joint representation relies on information originating from the top-down view, i.e., data generated from users expressing their conscious or unconscious view on the object or the system output. This is usually a time consuming and expensive process to obtain these user generated data, and does seldom provide any direct and obvious benefit to the users. An example includes the last.fm in which it is possible to tag each song, however the immediate gain in doing so is not apparent. However, the combined tagging space is presumable used

to recommend new songs. There is a number of machine learning elements, which may be employed to address this issue of experimental cost and perceived burden. The first step is to turn the information finding process or system optimization into a interesting game or application [13], which entices the user to use and interact with the system in a way where the uses expresses view is obtained without any particular perceived extra burden. Such games or simply interesting user interfaces has been an ongoing area in the MIR community with suggest such as [120][107][221].

The second element is to ensure that the system - based its needs - only requests annotations, which are useful in learning or adapting the computational representation. This is the field of active learning [194] (or experimental sequential design), which is specifically addressed in the setting of (discriminative) preference learning in Sec. 3. For the specific (generative) topic models (pLSA/LDA) only a limited body of studies exists [115][114]. It is speculated that this user driven adaption of topic models becomes an important part of further research in the field of MIR, as indicated by recent interest in the field of adaptive music representations [208].

Contribution [H] is a relatively simple extension of existing music topic models, but a required step towards more advanced model. A more advanced model should consider correlated topic models [27][63] for allowing different topic distributions in each modality. Supervised and discriminative topic models [117][28] for task and prediction driven learning, and potentially non-parametric extensions for added flexibility based for example on the hierarchical Dirichlet process (HDP) [212]. Finally, extensions to the base model should potentially support continuous observations, i.e., allowing for Gaussian observations typically encountered in music. This is for example considered [90] in the HDP setting, and as been evaluated for the pLSA model as well [239].

## 2.5 Summary

This chapter gave a overview of standard audio representations based on the bottom-up and top-down view of audio and music information systems. The chapter outlined popular bottom-up features of which some has been applied in contributions. The chapter furthermore outlined a number of top-down aspects, which are found relevant to music information retrieval, analysis and organization. The chapter further gave a short introduction to multi-modal integration in the music information retrieval field. Finally, a specific and realized system for multi-modal integration was described based on contribution H and the Million Song Dataset.

Chapter 3

# Preference Learning with Gaussian Processes

Ranking and rating of objects by human subjects is a common source of top-down information in many multimedia systems such as online movie recommendation services [166], book stores [6] and music services [1]. A noticeable characteristic of these examples is the human and cognitive implication since all examples depends on some subjective understanding of the objects and what relevance, similarity and preference is in relation to the those objects. We will generally consider such ranking and rating scenarios under the common term of preference learning, however the principle has applications much beyond preference learning and modelling.

**The response** is given by subjects who are asked to convey some notion of preference in the form a response, $y$, towards one or more objects from the (choice[1]) set $\mathcal{C} = \{\mathbf{x}_i | i = 1...C, C < \infty, \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is a set of $N$ input instances, $\mathcal{X} = \{\mathbf{x}_i | i = 1...N, \mathbf{x} \in \mathbb{X}\}$. This can for example be a ranking of multiple objects, a rating of similarity between two objects, or directly assigning a noisy absolute value to the object. The resulting set of $K$ such observations is denoted $\mathcal{Y} = \{(\mathbf{y}_k; \mathcal{C}_k) | k = 1...K, y \in \mathbb{Y}\}$[2].

[1] Only called a choice set in case of discrete choices

[2] The notation $(y_k; \mathcal{C}_k)$ is used to indicate that $y_k$ is dependent on the choice set, not part of the set as such.

This response- or observation types, can along with many others, be divided into two fundamentally different groups, namely

- **Relative (or indirect scaling [16])** Comparing (a finite set) of objects and ranking them in order (discrete choice / permutation model, e.g. [220][17]) or assigning value to the similarity between them. Depending on the setting, $y_k \in \{-1, 1\}$ $y_k \in [1 : C]$, $\mathbf{y}_k \in \mathbb{P}$ where $\mathbb{P}$ denotes all possible permutations of the $C$ objects, e.g. $\mathbf{y}_k = \{4, 1, 3, 2\}$.
- **Absolute (or direct scaling [16])** Assigning an absolute value to each object (regression model), e.g. $y_k \in \mathbb{R}$ or $y_k \in [0, 1]$.

The main assumption is, regardless of the response type, that each object $x \in \mathcal{X}$ has a latent value $f(\mathbf{x})$ and the response given by the subject is dependent on these latent values, whether this being in terms of one value (the absolute case) or in terms of multiple values, e.g. difference or rations between values (relative case).

The **relative (or indirect)** case typically calls for models based on discrete choice models in which either: a) one and an only one outcome is selected or b) a specific permutation of $C$ objects is chosen, indicating the ranking of the objects. Modelling such discrete choices was considered by Thurstone [216] in his seminal paper 'The law of comparative judgement', which lays the groundwork for the probit choice model considered in Sec. 3.2, and the basic working principle of socalled Thurstonian models. Later Bradley and Terry [35] developed the Bradley-Terry (BT) model, in effect a logit model for pairwise comparisons. Luce later extended this view with the Bradley Terry Luce (BTL) model [135] - or unordered multinomial/generalized logit - which laid the groundwork for 'Luce's axoim of choice' and the specific incarnation of the logistic function as choice model.

These discrete choice models can be extended to likelihood functions over permutations of objects, such as the Placett-Luce model originating from the BTL model [178], and a similar principle based on probit model [238], albeit the latter with much more difficult estimation/inference than the Placett-Luce. An even more general view can be obtained by considering general exponential forms for permutations such as the Mallow models [141], which is however not considered here.

A particular special case of discrete ranking models arises when only partial rankings are observed ($C < N$ and typically $C << N$), and the focus in this chapter is mainly on the special case when only two objects are considered at the time, i.e. $y \in \{-1, 1\}$. The resulting pairwise scenario results in a number

of paired observations where two different objects are ranked. Furthermore, the comparison between the two objects could potentially result in assigning a value to the difference between objects or degree of preference indicating for example beliefs or confidence in the decision. This latter aspect is considered in contribution [C].

In direct scaling paradigms i.e. the **absolute** case the subject assigns an absolute value to each object, essentially calling for regression models. Typically such ratings are given on a bounded or/and ordinal scale, in which case, a consistent modelling framework requires specialized noise/likelihood models as considered in Sec. 3.2.

**Modelling** these specialized response types is a major questions. In order to narrow down the possibilities, a number of properties is considered, often found in multimedia objects and elsewhere in machine learning, namely

I) High dimensional features spaces
II) (Relatively) few labeled examples
III) Unknown - or varying - task complexity (linear/non-linear)
IV) Multiple and heterogeneous sources of information
V) Limited time for experimentation (a cause of the few labeled samples)
VI) Possibility to predict $f(\mathbf{x}^*)$ for unseen examples,$\mathbf{x}^*$

The last aspect is not a must in a frequentist view using classical hypothesis tests. However, with the system and machine learning applications in mind, this is a requirement and immediately calls for modelling the latent preference values as a function. Aspect I-III calls for robust and flexible regression frameworks, and aspect IV calls for a flexible data representations. Aspect V calls for the ability to express (at least predictive) uncertainty such that sequential experimental design becomes feasible.

With these aspects in mind, the choice of modelling framework has fallen on Bayesian approach, and in particular a non-parametric versions based on Gaussian process priors on the function, $f(\cdot)$. In this Bayesian formulation, prior information can ensure robust and flexible modelling even given the adverse properties listed above. The non-parametric nature, furthermore, allows for arbitrary flexibility in the underlying function only assuming some notion of smoothness .

In this Bayesian framework, it is only assumed that the likelihood function can be parameterized by the function, $f(\cdot)$, and can be written in a factorized form

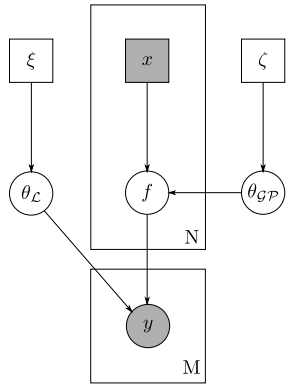



Figure 3.1: Plate model for the general setup

as $p(\mathcal{Y}|\mathcal{X}) = \prod_{k=1}^{K} p(y_k|\mathbf{f}_k, \boldsymbol{\theta}_{\mathcal{L}})$ where $\mathbf{f}_k = [f(\mathbf{x}_{k,1}), f(\mathbf{x}_{k,2}), .., f(\mathbf{x}_{k,C})]^\top$, where it is assumed that all elements in $\mathcal{X}$ are present in the union of all $\mathcal{C}_k's$.

Priors can now be placed on all the parameters, including $f$, and inference can (for now) be performed considering the full posterior,

$$p(\mathbf{f}, \boldsymbol{\theta}|\mathcal{Y}, \mathcal{X}) = \frac{p(\boldsymbol{\theta}_{\mathcal{GP}})\, p(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X})\, p(\boldsymbol{\theta}_{\mathcal{L}})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})}{p(\mathcal{Y}|\mathcal{X})}, \tag{3.1}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathcal{L}}, \boldsymbol{\theta}_{\mathcal{GP}}\}$. The marginal likelihood - or model evidence - is given as

$$p(\mathcal{Y}|\mathcal{X}) = \int\int\int p(\boldsymbol{\theta}_{\mathcal{GP}})\, p(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X})\, p(\boldsymbol{\theta}_{\mathcal{L}})\, p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})\, d\boldsymbol{\theta}_{\mathcal{GP}} d\boldsymbol{\theta}_{\mathcal{L}} d\mathbf{f}. \tag{3.2}$$

Where the prior on, $\mathbf{f}$, is a Gaussian process, which we will return to in Sec. 3.1. The graphical model in Fig. 3.1 shows the simplicity of the model, which, despite its appearance, is able to model and support the following four aspects:

- Many response/observation types (the likelihoods)
- Gaussian process prior for robustness and flexibility (the prior)
- Practical inference via analytical approximations (the inference)
- (Bayesian) active learning / sequential design (the sequential design)

The general setup includes many variants of each of these elements and Tab. 3 provides an overview of the many likelihoods functions (left), the Gaussian process prior (top), the inference methods (bottom), and the sequential design methods (right). Many of the elements are based on prior work in the field

and some are merely included to provide a coherent and complete overview and taxonomy, but not yet realized. Contributions [C],[G] and [D] can be seen as development steps towards the general setup, whereas contributions [E],[F] and [I] are direct applications of the realized setup.

**Outline:** The rest of this chapter aims to explain the elements of the overview shown in 3 starting with a short review of Gaussian processes and a basic overview of inference methods considered in Sec. 3.1. This is followed by an overview in Sec. 3.2 of how and why Gaussian processes can/should be applied in ranking and preference learning scenarios. Sec. 3.2 furthermore describes some of the contributes to the field, in particular new likelihood functions relevant for the preference and ranking modelling. Sec. 3.3 provides a combined overview of sequential design methods and sparse Gaussian processes - with focus on learning and computational complexity for preference leaning and ranking.

## 3.1 Gaussian Processes

The modelling element differentiating the preset setup from other choice modelling frameworks such as generalized linear models (GLM) [220][36] is the particular prior placed on the function values, $f(\mathbf{x})$, namely the Gaussian process (GP). The Gaussian process is simply defined as: 'A collection of random variables, any finite number of which have a (consistent) joint Gaussian distribution', [182]. It is here denoted, $\mathbf{f} \sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}\right), \mathrm{k}\left(\mathbf{x}, \cdot\right)_{\theta_{\mathrm{GP}}}\right)$.

The GP is defined by its mean function, $\mathbb{E}\left\{f\left(\mathbf{x}\right)\right\} = \mathrm{m}\left(\mathbf{x}\right)$, and a covariance function $\mathrm{k}(\mathbf{x}, \cdot|\boldsymbol{\theta}_{\mathcal{GP}})$ which - given the hyperparameters collected in $\boldsymbol{\theta}_{\mathcal{GP}}$ - defines the correlation between all individual variables, $f(\cdot)$. For a zero mean GP the correlation function is $\mathbb{E}\left\{f(\mathbf{x})f(\mathbf{x}')\right\} = \mathrm{k}\left(\mathbf{x}, \mathbf{x}'\right)_{\boldsymbol{\theta}_{\mathcal{GP}}}$. Since each random variable, $f(\mathbf{x})$, is correlated with another random variable $f(\mathbf{x}')$ though the covariance function, the function must necessarily have some smoothness. Furthermore, the 'nice' marginalising and conditioning properties of the multivariate Gaussian distribution makes the GP itself tractable. Fig. 3.2(a) shows the graphical model for a GP.

By considering the function value of any input a random variable, the GP defines a prior over functions. From this prior, we may draw a finite number of function values, $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_N)]$, and given the definition, we may continue drawing function values for an infinite number of different inputs. Fig. 3.2(b) shows a number of functions drawn from an unconditioned GP. If observation about $f(x)$ the GP is simply defined by conditioning on these observations, Fig. 3.2(c) shows functions drawn from a conditioned GP.

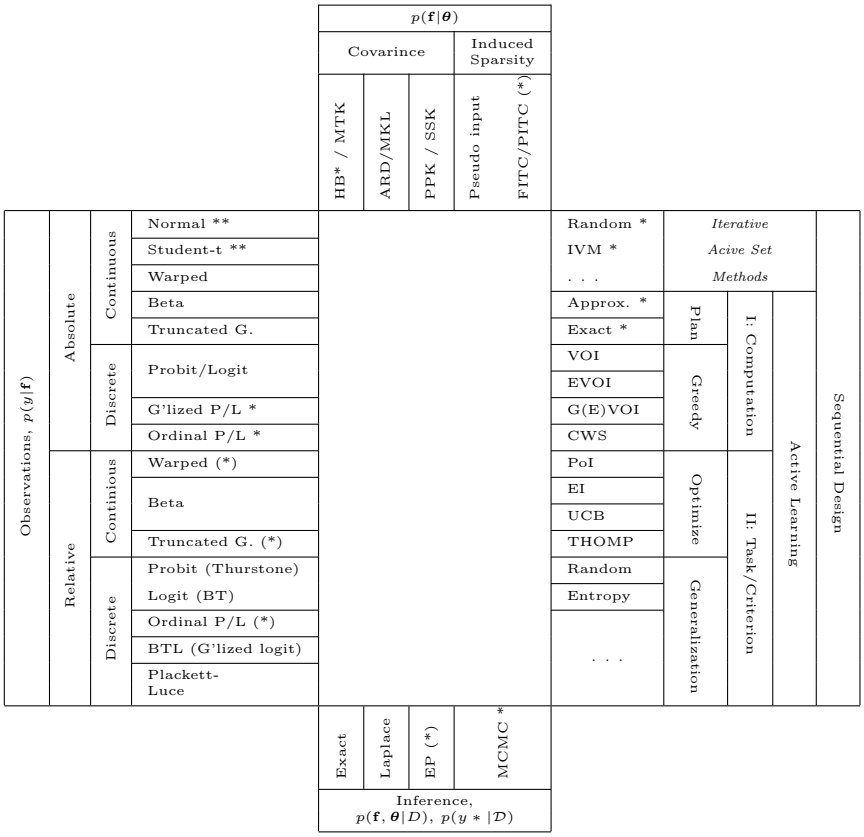
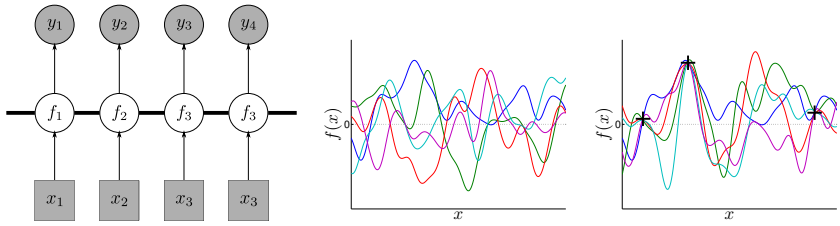



Table 3.1: General preference learning setup. **Left**: Likelihood functions **Top**: The GP prior specification **Bottom**: Inference methods **Right**: Sequential design elements. Note that elements form different directions (top/bottom/left/reight) are not exclusive, i.e., may in particular select both sparsity by the pseudo input and the PPK covariance function. And the sequential design in the active learning setting is required to select both a computational option and a task/criterion. Elements marked with ** has not been implemented in the setup but is often a part of other GP toolboxes such as the GPML toolbox [67]. Elements marked with * has not yet been implemented or only partly (*), e.g. EP only exists for pairwise probit and absolute Beta/ truncated Gaussian (Truncated G.). Variational Bayes (VB) [167] has been omitted since not considered in the thesis.

So the GP is nothing more than a prior over functions, indicating that the GP encompasses many different function classes, only assuming smoothness. Any likelihood which can be parameterized by variables corresponding to particular




(a) Prior sample (b) Prior samples (c) Conditioned samples

Figure 3.2: a) Chain model of the standard GP model showing the correlation between all f's as a solid bar. b) Samples drawn from the same zero-mean Gaussian Processes with a so-called *squared exponential* covariance function. c) Samples drawn conditioned on other variables showing how the GP effectively defines the function and is able to interpolate between observed samples. The variation of the different realizations indicate the predictive uncertainty in between the observed (and conditioning) points.

input, $\mathbf{x}$, may have its parameterizing function,$f(\cdot)$, equipped with a Gaussian process prior. Such likelihood functions include the standard Gaussian noise model[182] and the Student-t [223] for regression. GP models for classification typically rely the probit and the logit models [182]. These standard classification and regression settings of Gaussian process has been covered in a vast body of literature such as [182][139][26], and this section will only provide a minimal introduction to the subject sufficient for reading the associated contributions.

The simplest - yet highly applied - model relying on Gaussian process is regression with a standard Gaussian likelihood model, i.e., $\mathrm{p}(\mathrm{y}_k|\mathrm{f}_k) = \mathcal{N}(y_k|\mathrm{f}_k, \sigma_i)$ where $\sigma_i$ denotes the observation noise and the mean is parameterized by the function drawn from a GP. If all variables/hyperparameters in the posterior from Eq. 3.1 are kept constant except for $\mathbf{f}$, predictions for a set of test inputs collected in $\mathbf{X}^* \in \mathbb{R}^{N_{test} \times D}$ can derived for a new input $\mathbf{x}^*$ depending on previously observed responses collected in $\mathbf{y}$ as

$$\mathbb{E}\{\mathbf{f}^*\} = \mathbf{K}_{\mathbf{X}\mathbf{X}_*}\left[\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_i^2\mathbf{I}\right]^{-1}\mathbf{y} \tag{3.3}$$

$$\mathbb{V}\{\mathbf{f}^*\} = \mathbf{K}_{\mathbf{X}^*\mathbf{X}_*} - \mathbf{K}_{\mathbf{X}\mathbf{X}^*}\left[\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_i^2\mathbf{I}\right]^{-1}\mathbf{K}_{\mathbf{X}\mathbf{X}^*} \tag{3.4}$$

where $\mathbf{f}^*$ denotes the multivariate variable $\mathbf{f} \in \mathbb{R}^{N^*}$. $\mathbf{K}_{\mathbf{X}\mathbf{X}}$ is a covariance matrix with element between all training inputs, $\mathbf{X}$. $\mathbf{K}_{\mathbf{X}\mathbf{X}^*}$ is the covariance matric between training and test sets. $\mathbf{K}_{\mathbf{X}^*\mathbf{X}^*}$ is the covariance matrix between all test inputs. This leads to so-called correlated predictions which is not typically employed in standard machine learning using GP regression and classification, but highly relevant for the preference learning model considered later.

The expression for the predictive mean $\mathbb{E}(f^*)$, is highly informative in understanding the Gaussian process model. In particular the mean prediction which is simply a linear combination of the observations. This emphasizes the non-parametric form of the GP where the predictions are not dependent on any parameters, but only ont the observations themselves and the hyperparameters in the covariance function. The predictive (co)variance for this Gaussian model is further characterized by not having any dependence on the observations.

The nice properties and analytical tractability implied here, is unfortunately not a general property for the model with arbitrary likelihoods interesting to the ranking and rating applications. If the likelihood - as in many of the sections to follow - is not Gaussian, the posterior and predictive distribution are not analytically trackable - even when holding all other variables constant - and we are forced to resort to approximation of the posterior for $\mathbf{f}$ and possibly the hyper parameters. Luckily there exists a multitude of approximation methods for Gaussian processes of which we will shortly discuss the Laplace approximation, Expectation Propagation (EP) and a full Bayesian simulation/MCMC approximation in Sec. 3.1.2.

### 3.1.1 Mean & Covariance Functions

The defining part of the GP prior is the mean and covariance functions, which effectively defines how flexible the function can be. The mean function, $\mathrm{m}(\mathbf{x})$ - as the name suggest - defines a particular prior belief regarding the expectation of the latent function. A common choice is the zero-mean function, such that $\mathrm{m}(\mathbf{x}) = 0 \forall \mathbf{x}$. Other easy to handle mean functions are constants and polynomials of varying order [182].

The covariance function is a much more subtle, and in turn powerful tool, since defining the correlations between individual function values in effect results in smoothness. This lets us predict beyond the current observations as already indicated by the examples in Fig. 3.2(c). The standard covariance function in machine learning is the so-called squared exponential covariance function, which is the default in most kernel methods employed in machine learning. The generalized version of this covariance function is given by

$$\mathrm{k}(\mathbf{x}, \mathbf{x}') = \sigma_s \exp\left(-\frac{1}{\sigma_\ell}(\mathbf{x} - \mathbf{x}')^\top \boldsymbol{\Lambda}^{-1}(\mathbf{x} - \mathbf{x}')\right) \tag{3.5}$$

where $\boldsymbol{\Lambda}$ is a positive definite matrix defining the correlation between input dimensions or features.

**Automatic Relevance Determination (ARD)** Letting the matrix in 3.5 be

diagonal,i.e.,

$$\mathbf{\Lambda} = \begin{bmatrix} \sigma_\ell^{(1)} & & & \\ & \sigma_\ell^{(2)} & & \\ & & \dots & \\ & & & \sigma_\ell^{(D)} \end{bmatrix}$$

with individual $\sigma$'s leads to the so-called automatic relevance determination covariance function, by allowing individual feature dimension to be scaled individually. Thus, in the case of large dispersion between individual $\sigma_\ell$'s it makes sense to rank features according to their individual values, where small $\sigma_\ell$ indicates dimensions with larger variance (less correlation) than ones with larger $\sigma_\ell$ (high correlation).

Letting $\mathbf{\Lambda}$ be a full - yet still positive semi-definite matrix - provides a principled way of formulating metric learning[3] with Gaussian processes. However, this typically requires more involved inference than the methods considered in the following. In contribution [I] a constant $\mathbf{\Lambda}$ was proposed for the task of elicitation preference in a system with assumed correlation between input dimensions.

**Multiple Kernel Learning (MKL)** With the ARD covariance function it is often difficult to obtain robust and interpretable results on individual feature, however another way to consider relevance determination is on a group level (of features) though multiple kernel learning (MKL), in which multiple individual kernels are linearly combined to form a joint feature space. This can be written as

$$\mathrm{k}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sum_{m=1}^{M} \alpha_m \, \mathrm{k}^{(m)} \left(\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)}\right)$$

where $\alpha_m \geq 0$. This expression indicates that each individual (possibly different) covariance function may operate on the full input space - or if seen as a group level ARD kernel, each kernel can also operates on non-overlapping subset of the feature space, i.e. different dimensions of $\mathbf{x}$. The main issue is the inference of each $\alpha$, which in the Gaussian process setting may be performed using evidence maximization, which also has a strong link to the risk minimization view of MKL [168].

MKL has an important purpose in the context of this thesis, namely to fuse heterogeneous data sources in for example multi-media, as suggested in [D] for music, or sensor fusion [4] by the use of multiple kernels. The structure of these additive kernels has links to ANOVA type analysis, which in the GP framework

[3] See e.g. for an a metric learning principle with SVMs [147], where the learning takes place in kernel, i.e. Hilbert, space, in contrary to the outlined possibility which defines the metric directly in the input space.

can be obtained through the use of additive kernels [59] allowing interpretation of complex interactions between individual features.

**Multi-Task (MTK)** The GP covariance function further provides a simple, but easy and intuitive way of defining a multi-task model, without resorting to more advanced hierarchical structures such as [22]. The simplest form of such multi-task kernel was proposed in [32] and is simply given by a product kernel

$$\begin{aligned} \mathrm{k}\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \mathrm{k}\left(\left[\mathbf{x}_i^{(a)}, \mathbf{x}_i^{(b)}\right]^\top, \left[\mathbf{x}_j^{(a)}, \mathbf{x}_j^{(b)}\right]^\top\right) \\ &= \mathrm{k}\left(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(a)}\right) \mathrm{k}\left(\mathbf{x}_i^{(b)}, \mathbf{x}_j^{(b)}\right) \end{aligned} \tag{3.6}$$

This means that the kernel matrix can be written as the Kronecker product between the two individual kernel matrices, i.e. $\mathbf{K} = \mathbf{K}^{(a)} \otimes \mathbf{K}^{(b)}$ yielding a relative easy implementation. A comprehensive review of such multi-task methods is given in [5] in the setting of multiple output learning with general kernel methods.

The multitask-kernel approach provides a simple way of defining a multi-user model by correlating user by their features $\mathbf{x}^{(user)}$, as applied in contribution [J]. However, more elaborate methods has been prosed for the particular task for preference learning such as an hierarchical approach [24], and a more traditional collaborative filtering approach [93] comparing the aforementioned methods.

**Probability Product Kernel (PPK)** The PPK kernel [97] (as previously mentioned in connection with audio modelling) is an elegant method for combining the discriminative nature of the Gaussian process model with generative properties of the input data by directly defining the covariance functions as the inner product between two probability distributions, $p(\mathbf{x}|\boldsymbol{\theta})$ and $p(\mathbf{x}|\boldsymbol{\theta}')$, i.e.,

$$\mathrm{k}\left(p\left(\mathbf{x}|\boldsymbol{\theta}\right), p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right) = \int \left(p\left(\mathbf{x}|\theta\right) p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right)^{1/q} d\mathbf{x}$$

where $q \geqslant 0$. In effect this implies that each object can be represented by a probability distribution, which is highly relevant for audio where models such as Gaussian mixture models, Markov models and even Hidden Markov Models (HMM) are frequently used models of low levels audio features as outlined in Sec. 2). In contribution [F]D, the PPK was used based on GMM models for individual audio objects.

### 3.1.2 Inference & Model Selection

The main element required for Bayesian inference is naturally the posterior defined in Eq. 3.1, which combines the observations (via the likelihood function) with the arbitrarily complex prior. In the current preference learning setup the posterior is of the same form given by,

$$p\left(\mathbf{f}, \boldsymbol{\theta} | \mathcal{Y}, \mathcal{X}\right) = \frac{p\left(\boldsymbol{\theta}_{\mathcal{GP}}\right) p\left(\mathbf{f} | \boldsymbol{\theta}_{\mathcal{GP}}, \mathcal{X}\right) p\left(\boldsymbol{\theta}_{\mathcal{L}}\right) p\left(\mathcal{Y} | \mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}\right)}{p\left(\mathcal{Y} | \mathcal{X}\right)}, \tag{3.7}$$

where we differentiate between the the parameters directly entering the likelihood - and the hyperparameters $\boldsymbol{\theta}$. In the ideal case all priors would be conjugate to the likelihood, i.e., resulting in a closed form posterior [26]. This is not often the case, especially in the current setting. Instead, we turn to approximations methods as often the case when it comes to non-trivial Bayesian models.

The interest from the machine learning community in Gaussian process models during the nineties and since has resulted in a massive body of literature on inference for Gaussian process models: from Laplace approximations [231][26], to variational methods [26], expectation propagation [154] and sampling methods [165]. This thesis takes advantage of a few of the well-established ones for the particular task of practical and robust preference learning in real-world applications, where robustness and speed is of the essence.

Eq. 3.1 defines the full posterior over all parameters which is often a very difficult distribution to deal with, and in the current we are often interested in the posterior over $\mathbf{f}$. In general a vast number of inference schema and techniques have been suggested for dealing with the full posterior including exact/numerical marginalizing over $\boldsymbol{\theta}$, approximate (possibly a factorized version) the full posterior - or focus on inferring $p(\mathbf{f}|\cdot)$ with fixed $\boldsymbol{\theta}$. Given the focus in this thesis, we will describe the the latter approach, i.e. focus on $p(\mathbf{f}|\cdot)$ with fixed $\boldsymbol{\theta}$ due to its simplicity and effectiveness.

**A two step approach:** Analytic approximations to the full posterior are typically applied in a two-stage, iterative inference approach, where the **first** level inference regarding the posterior over $\mathbf{f}$ is performed with all other parameters fixed. Subsequently a **second** level inference is performed finding point estimates, $\hat{\boldsymbol{\theta}}$, of the hyper parameter (or alternatively marginalizing over them) as discussed in Sec. 3.1.2. This conceptual leads to a two step approach w is then iterated until convergence, which often provides a simple yet effective algorithm , i.e.,

**I** Approximate first level posterior, $p(\mathbf{f}|\boldsymbol{\theta}, \mathcal{X}, \mathcal{Y})$ using Laplace or EP with $\boldsymbol{\theta}$

fixed.

**II** Find ML/MAP-II point-estimates of the hyperparemetrs $\hat{\boldsymbol{\theta}}$ based on marginal likelihood approximation, provided by the first level approximation.
... iterate until convergence of $\hat{\boldsymbol{\theta}}$ or the marginal likelihood / evidence.

Thus, the **first level** inference considers the following posterior $p(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is a set of fixed hyperparameters (e.g. in the covariance function). Given the posterior, predictions for general likelihoods easily follows from the Bayesian realm. Both the latent function value of the new inputs, $f^*$ and the final observation, $y^*$, which is dependent on one or more latent function values though the likelihood. Hence, first finding the multiple latent function values of a number of inputs (for the relative likelihoods) collected in $\mathcal{X}^*$,

$$p\left(\mathbf{f}^*|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}}, \mathcal{X}^*\right) = \int_{\mathbf{f}} p\left(\mathbf{f}^*|\mathbf{f}, \mathcal{X}^*\right) p(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \widehat{\boldsymbol{\theta}}) d\mathbf{f} \tag{3.8}$$

$$= \mathcal{N}\left(\mathbf{f}^*|\boldsymbol{\mu}_f{}^*, \mathbf{K}_f{}^*\right) \tag{3.9}$$

$\boldsymbol{\mu}_f^*$ is referred to as the predictive mean of $f(\cdot)$ and $\mathbf{K}_f^*$ is referred to as the predictive (co)variance of $f(\cdot)$. The predictions for the relative likelihoods are multidimensional and correlated due to the likelihoods being dependent on two or more $f(\cdot)$ variables, also for predictions.

Generally, we are interested in the prediction of the observed variable which is given by,

$$p\left(y^*|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}}, \mathcal{X}^*\right) = \int_{\mathbf{f}^*} p\left(y^*|\mathbf{f}^*, \hat{\boldsymbol{\theta}}_{\mathcal{L}}\right) p\left(\mathbf{f}^*|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}}, \mathcal{X}^*\right) d\mathbf{f}^* \tag{3.10}$$

$$= \int_{\mathbf{f}^*} p\left(y^*|\mathbf{f}^*, \hat{\boldsymbol{\theta}}_{\mathcal{L}}\right) \mathcal{N}\left(\mathbf{f}^*|\boldsymbol{\mu}^*, \mathbf{K}^*\right) d\mathbf{f}^* \tag{3.11}$$

which is referred to as the predictive distribution. Whether the latter integral is tractable depends on the particular likelihood.

The defining element is of course still the posterior, $p(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}})$, which needs to be approximated in many cases. It not the aim to provide a complete and detailed overview of GP inference, and in the classification case, we refer to [167] for an excellent review. Here a number of inference methods is outlined for the general preference learning, which often requires custom realizations of the standard GP methods, due to the multiple dependencies on $f(\cdot)$.

**Exact** : Exact inference, even at the first level, is only possibly in a subset of the models considered here such as the Normal likelihood. Noticeable and important cases include the so-called Warped likelihood models [200], which provides predictions of the same form as the Normal regression case.

**Laplace** : The Laplace approximation is a general tool for approximating intractable integrals. For GP based models, the starting point is the first level posterior. This posterior is approximated by a second order Taylor expansion around the mode, which provides a multivariate Gaussian approximation, i.e,

$$q\left(\mathbf{f}|\,\mathcal{X},\mathcal{Y},\boldsymbol{\theta}\right) \triangleq \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}},\mathbf{A}^{-1}\right) \tag{3.12}$$

where $\hat{f}$ denotes the mode and $\mathbf{A}$ denotes the covariance matrix of the multivariate approximation. Thus, the main task is to find the mode and the covariance matrix for this approximating distribution, which in turn becomes an optimization problem. This is a standard problem of moment matching, which is done by considering the unnormalized log-posterior and its first two derivatives. If the log-likelihood is log-concave this is a convex problem and can be solved by standard Newton step [182, Sec.3.4.1].

The standard Laplace approximation for binary classification with Gaussian processes [182, Sec.3.4.1], i.e., the model denoted as absolute-discrete-bounded-probit in Fig.3.3, assumes that the Hessian of the cost function is diagonal since the likelihood only depends on one variable, $f(\mathbf{x})$. However, in the preference learning scenario with relative models (relative-discrete-pairwise, BTL and Planket-Luce) this does not hold true, and custom implementation is required. The Laplace approximation was suggested for the pairwise probit model in [48], and the setup presented in Tab.3 contains a general Laplace implementation supporting both absolute, relative observation for both the pairwise, and the general BTL and Plackett-Luce models. For robustness and fast convergence, the setup has the option to use a damped Newton step. Furthermore, for non-log concave likelihoods the setup applies a simple but effective methods of thresholds the second derivative of the likelihood function as suggested in [232] to obtain a robust approximation, despite possibly non-unique solutions.

The Laplace approximation is also directly applied in approximate the marginal likelihood - or evidence - defined by the integral in Eq. 3.2.

**Expectation Propagation** : The Laplace approximation is a simple, but effective approximation method and relatively fast to derive for a number of new models based on the setup in Tab. 3.1, however, it simply assumes a local expansion around the mode of the posterior. The EP does not make this local assumption and considers a factorized posterior with individual approximations to each likelihood terms. The approximation can thus be written as

$$q\left(\mathbf{f}|\mathcal{X},\mathcal{Y}\right) \triangleq \frac{1}{Z_{EP}} p\left(\mathbf{f}\right) \prod t_k\left(\mathbf{f}_k|\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k,\tilde{Z}_k\right) = \mathcal{N}\left(\boldsymbol{\mu}_{EP},\boldsymbol{\Sigma}_{EP}\right), \tag{3.13}$$

where $t_k\left(\mathbf{f}_k|\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k, \tilde{Z}_k\right)$ is the approximation to each likelihood term and taken to be a (unnormalized) Gaussian distribution. $Z_{EP}$ is a global normalizing term ensuring that the global multivariate Gaussian approximation is consistent, i.e., integrates to one. In the case of relative observations with more than one $f(\cdot)$ variable in each likelihood term, the $t_k$ terms becomes an (unnormalized) multivariate Gaussian distribution.

The inference then amounts to estimating the so-called site parameters, $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{\boldsymbol{\Sigma}}_k$ which is done using the fixed-points algorithm proposed in [154] consisting of the following steps

For each observation $k \in [1:K]$

1 $q^{/k}(\mathbf{f}) \leftarrow q(\mathbf{f}|\cdot)/t_k(\mathbf{f}|\cdot) = \prod_{k \neq l} t_l(\mathbf{f}_l|\cdot)$

2 ${t_k}^{new}(\mathbf{f}_k|\cdot) \leftarrow \underset{t_k(\mathbf{f}_k|\cdot)}{\arg\min} D_{KL}\left(p(y_k|\mathbf{f})\, q^{/k}(\mathbf{f}) \,||\, t_k(\mathbf{f}_k|\cdot)\, q^{/k}(\mathbf{f})\right)$

3 $q(\mathbf{f}) \leftarrow {t_k}^{new}(\mathbf{f}_k|\cdot)\, q^{/k}(\mathbf{f})$

...repeat until convergence

where $q^{/k}$ is used to denote that the likelihood term originating from observation $k$ has been removed. The minimization of the KL divergence ($D_{KL}$) in step 2 is a matter of moment matching since using the (un)normalized Gaussian distribution as site approximation.

The setup in Tab.3 includes a custom EP implementation for the discrete pairwise probit model, which is an extended version of the standard EP models for probit models originally proposed in [46], alltough EP inference for general relative models is not currently supported.

It should be emphasized that EP can be used in a 'online' setting [53] in which the iteration only runs for a single sweep though all the data points [53]. This, however, still requires maintaining the site parameters for all the observations.

**Sampling** A different approach, not considered in detail in the thesis, is to perform inference through simulation by drawing $N_s$ samples, $\mathbf{f}^{(s)}$, from the posterior in Eq. 3.1, from which we can evaluate the moments of the posterior of both the latent variables, $\mathbf{f}$, and the hyperparameters, $\boldsymbol{\theta}$ and can in turn calculates expectations of for example the first level posterior as, $\mathbb{E}_{p(\mathbf{f}|\mathcal{X},\mathcal{Y},\boldsymbol{\theta})}\{\mathbf{f}\} \approx \frac{1}{N_s}\sum_{s=1}^{N_s}\mathbf{f}^{(s)}$.

Sampling in Gaussian process models is usually performed using variations of Markov Chain Monte Carlo (MCMC) methods as outlined in [165] possibly combined with other sampling steps such as slice sampling [162] for sampling hyperparameters. Given the applied nature of the preference learning setup, sampling is only used for evaluation of other approximation, i.e., as a golden standard.

The particular inference approach used for estimating $p(\mathbf{f}|\cdot)$ is typically dependent on the actual likelihood / model, but the applications considered in the contributions usually rely on the relatively simple Laplace approximation, which has proven a easy, reliable and robust method for the relative models.

**Hyperparameters and Model Selection** In the two step procedure outlined above, the first level inference focused (mainly) on the posterior of $\mathbf{f}$, i.e. $p(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}})$, however, the hyperparameters, $\hat{\boldsymbol{\theta}}$, are as mentioned also critical model parameters (as indicted by the full posterior in Eq. 3.1) and should in principle not be considered any different than $\mathbf{f}$ in a full Bayesian treatment. If we do not need to infer actual values/distributes for the hyperparameters, we may **marginalize** over them either as part of part of a full MCMC inference step or a combination with the level one analytical/approximation, and/or numerical methods.

However, in the two-step approach which is the focus, we may turn to a **optimization** approach which is an effective alternatively to the often computational expensive marginalization approach. We thus abandon the full Bayesian schema indicated by Eq. 3.7 and aim only a a single 'good' point estimate for the hyperparameters. This corresponds to the **second** level of inference outlined above, and is based on the notion of model evidence and Bayesian model comparison.

The model evidence - or marginal likelihood - is an important and and much debated entity in the Bayesian framework. The marginal likelihood is a expression of how likely it is to generate the given data, $\mathcal{D}$, given a particular model and model parameters [26]. So if the hyperparameters are fixed, we may conduct Bayesian model comparison by the ratio between the marginal likelihood resulting in the so-called Bayes factor,

$$\mathrm{BF} = \frac{p\left(\mathcal{Y}|\mathcal{X}, \mathcal{M}_1\right)}{p\left(\mathcal{Y}|\mathcal{X}, \mathcal{M}_2\right)} = \frac{\int p\left(\mathcal{Y}|\mathbf{f}\right) p\left(\mathbf{f}|\mathcal{X}, \mathcal{M}_1\right) d\mathbf{f}}{\int p\left(\mathcal{Y}|\mathbf{f}\right) p\left(\mathbf{f}|\mathcal{X}, \mathcal{M}_2\right) d\mathbf{f}}$$

If the true data generating distribution is contained in the model this factor can on average be shown to select the correct model [26]. Hence, the model with the highest marginal likelihood is preferred and finding the 'best' point value of the hyper parameters,$\hat{\boldsymbol{\theta}}$, can be an integrated part of the inference by directly optimizing the marginal likelihood as part of the inference. This procedure is known under different names including maximum likelihood II (ML-II), empirical Bayes [41] or in the machine learning literature as evidence optimization [136].

The **second level** inference/estimation therefore involves finding point esti-

mates of the hyperparamters in the covariance functions, $\hat{\boldsymbol{\theta}}_{\mathcal{GP}}$, and in the likelihood function $\hat{\boldsymbol{\theta}}_{\mathcal{L}}$. This can be done by approximating the marginal likelihood as previously discussed, which is readily available given the Laplace and EP approximation (or can be estimated numerically, e.g. using quadrature). The general preference learning setup and model furthermore allows for specific simple and constant hyperpriors on the covariance and likelihood parameters as indicated in Fig.3.3. In a point-estimation procedure this setup is referred to as maximum posterior II (MAP-II) estimation. Computational this is still only a part of the second level inference. At the present the preference learning setup include Gamma priors C or weakly informative half Student-t priors [73]I, the latter inspired by their use in [222].

The actual problem of optimizing the evidence is typically carried out using gradient methods or a generic BFGS method. For the Laplace approximation this requires a certain number of derivatives [182, Sec5.5] or [G]. The EP approximation is often simpler in this regard and does not require as many higher order derivatives as the Laplace approximation [182, Sec.5.5.2]. Regardless of applying the Laplace or EP approximation, the problem of evidence maximization is often highly non-convex and global solution can seldom be guaranteed, although it has been found robust in many of the applications considered in contributions.

In real applications, where systems based on Gaussian process acts as the computational representation (Fig. 1.1), the notion of model evidence is a clear advantage over its non-Bayesian relatives such as kernel generalized linear models [242] and support vector machines [197] often relying on cross-validation for model comparison. Cross-validation is typically also used in the described GP models as model evaluation, however the pure option of evidence optimization, allows for stand-alone models without the need for cross-validating the hyper parameters, as for example the case in contribution [I].

## 3.2 Rating & Ranking with GPs

| Observations, $p(y\|\mathbf{f})$ | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relative | | | | | | Absolute | | | | | | | |
| Discrete | | | | Continious | | Discrete | | Continuous | | | | | |
| General | | Pairwise | | | | Bounded | | | | | | (Std.) | |
| Plackett-Luce | BTL (G'lized logit) | Ordinal P/L | Probit (Thurstone) Logit (BT) | Truncated G. | Beta | Warped | Ordinal P/L | G'lized P/L | Probit/Logit | Truncated G. | Beta | Warped | Student-t | Normal |
| [178] [80] | [135] | | [216][48][46] [35][35][34] | | ([225]) | | [47] | [182][74] | [182] | | [66][232] | | [222] | [182] |
| | | | [G],F [E] | | [C] | | | | [4] | [J] | [J] | [I] | | |

Table 3.2: Taxonomy of likelihood functions. The **second** layer is whether the observations is a relevant or absolute (direct) scale. The **third** layer specifies whether the observations are discrete or continuous the **fourth** layer then determines the particular aspects of the observations (pairwise, bounded etc). Here *G'lized P/L* refers to Generalized probit or Logit models. In the absolute case these are also known as *unordered* P/L or simply multinomial models. In the relative case these are known as *ordered* P/L models. The **fifth** layer specifies the particular likelihood function. The **sixth** row specifies a number of non-exhaustive but relevant reference, favoring links to GP versions. The bottom row indicated which contribution makes use of the likelihood and potentially which section describes them in detail later in this section.

The standard classification and regression setting of Gaussian processes are by now well established tools in the machine learning community (happily competing with other kernel methods such as SVMs and kernel GLM) [182], and extensions and variations are being applied in a number of non-standard settings. One such non-standard setting is the ranking and preference which calls for a special kind of observational models regardless of it being a discrete choice model or a continuous bounded rating.

The related contributions applies and proposes a number observational models for Gaussian processes in the model defined in Fig. 3.1. In the following the the most often applied models are outlined and motivated.

### 3.2.1 Relative-Discrete-Pairwise-Probit/Logit Model

One of the most popular experimental paradigms for eliciting perceptual and cognitive effects is the two alternative forced choice, i.e., the choice set consist of two objects $x_{u_k} \in \mathcal{X}$ and $x_{v_k} \in \mathcal{X}$. These are presented to a subject and asking him/her to select one, i.e., $y_k \in \{-1, 1\}$. By repeating this for many different objects results in $K$ responses on pairwise comparisons between any two inputs in $\mathcal{X}$, $\mathcal{Y} = \{(y_k; x_{u_k}, x_{v_k}) | k = 1, ..., K\}$

A classic modelling approach towards such pairwise choices is based on Thurstone's "law of comparative judgements" in which the choice is assumed to be the effect of internal but noisy utilities for each object, i.e., the utility for an object $\mathbf{x}_u$ is $f(\mathbf{x}_u) + \varepsilon_u$ and the utility for object $\mathbf{x}_v$ is $f(\mathbf{x}_v) + \varepsilon_v$ where $\varepsilon$ is the noise. The general starting point for both the probit and pairwise choice models comes from considering the probability of a subject selecting $\mathbf{x}_u$, i.e.

$$P\left(f\left(\mathbf{x}_u\right) + \varepsilon_u > f\left(\mathbf{x}_v\right) + \varepsilon_v\right) = P\left(f\left(\mathbf{x}_u\right) - f\left(\mathbf{x}_v\right) > \varepsilon_v - \varepsilon_u\right) \tag{3.14}$$

$$= \int \mathbb{I}\left(f\left(\mathbf{x}_u\right) - f\left(\mathbf{x}_v\right) > \varepsilon_v - \varepsilon_u\right) p\left(\varepsilon\right) d\varepsilon \tag{3.15}$$

where $\mathbb{I}(\cdot)$ is an indicator function, returning 1 if the argument is true; otherwise zero. The main aspects to consider is the noise distribution, $p\left(\varepsilon\right)$, on the internal utilities which effectively leads to difference choice models such as the probit or logit. In [216] Thurstone suggested a Normal distribution which results in the probit model [30], while the a pairwise logit model assumes $\varepsilon$ to be logistically distributed [30][220]. Following Thurstones approach, $p(\varepsilon)$, is normally distributed and depending on the assumptions regarding the correlated variables; we obtain a number of the case which Thurstone outlines in [216]. However, the simplest one, case 5, is derived by assuming that the noise on individual objects are uncorrelated and have equal variance, $\sigma_{\mathcal{L}}$. Fortunately, this results in the multidimensional integral in Eq. 3.14 being tractable and can be shown to result in the cumulative Gaussian (the probit model) such that the likelihood becomes

$$p\left(y_k | \mathbf{f}_k, \sigma\right) = \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma_{\mathcal{L}}}\right)$$

with $\mathbf{f}_k = \left[f\left(\mathbf{x}_{u_k}\right), f\left(\mathbf{x}_{v_k}\right)\right]^\top$. This defines the likelihood (actually the probability) of the choice, $y_k$, however the estimation of the function values themselves is

the main interest here. By assuming uncertainty on the values of $f(\cdot)$, we place the Gaussian process over the latent function values. Such as model was first considered in [48]. Since the model now has dependencies on two latent function values the graphical chain model is slightly different than in the standard case and depicted in Fig. 3.3.

The primary model applied in contributions [C],[G][D][F] is the probit choice model described above. Compared to the standard setup proposed in [48], we allow for simple priors on the hyperparameters, which has been found useful and robust in many of the applications. Assuming a squared exponential covariance function [182, Cha.4], with two hyper parameters, $\sigma_\ell$ and $\sigma_s$, the combined model can be specified as

$$
\begin{aligned}
\sigma_s|\xi_s &\sim \text{half Student-t } / \text{Gamma} \\
\sigma_\ell|\xi_\ell &\sim \text{half Student-t } / \text{Gamma} \\
\mathbf{f}_k|\sigma_s, \sigma_\ell &\sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}_\mathrm{k}\right), \mathrm{k}\left(\mathbf{x}_\mathrm{k}, \cdot\right)_{\sigma_\mathrm{s}, \sigma_\ell}\right) \\
\pi_k|\mathbf{f}_k, \sigma_\mathcal{L} &= \Phi\left(f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)\right) \\
y_k &\sim \text{Bernoulli}\left(\pi_k\right)
\end{aligned}
$$

where $\xi$ is a set of parameters in the hyperprior, for example scale and degree of freedom parameters for the half Student-t prior [73].

This model requires approximation of the posterior, and the setup currently supports both Laplace and EP. Predictions for a new choice set $\mathcal{C} = \{\mathbf{x}_r, \mathbf{x}_s\}$ is (with the Gaussian approximation) analytical in the form of a cumulative Gaussian.

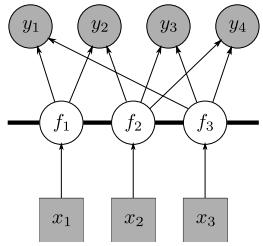



Figure 3.3: Graphical chain representation of the pairwise model. Compared to the standard GP classification model, the output variables $y$ now each depend on multiple $f$ variables.

### 3.2.2 Relative-Continuous-Pairwise-Beta Model

The classic probit likelihood is mainly chosen for its robustness compared to e.g continuous/absolute scaling experiments. However, the pairwise comparisons is potentially a slow way of learning in a high-dimensional input space and with many objects to consider. It may therefore be advantageous to exploit the extra information from continuous responses to get a faster method for preference elicitation without jeopardizing the robustness from standard binary responses. This can be done by observing a degree to which either is preferred alongside the binary decision of $\mathbf{x}_u$ and $\mathbf{x}_v$. Such a setting is some times refereed to as sureness or confidence [230]. In the preference learning scenarios, we refer to it as degree of preference (DoP).

To address this setting a continuous, but bounded response $y \in ]0;1[$ is defined, observed when comparing the two options in $\mathbf{x}_u$ and $\mathbf{x}_v$. The first option, $\mathbf{x}_u$, is preferred for $y < 0.5$. The second option, $\mathbf{x}_v$, is preferred for $y > 0.5$ and none is preferred for $y = 0.5$. Hence, the response captures both the choice between $\mathbf{x}_u$ and $\mathbf{x}_v$, and the degree of the preference. One could consider the two choices separately and assume that continuous DoP choice always follows the binary choice such that it can not change the binary choice to ensure robustness. This can however easily be shown to lead to the same modelling considerations.

Modelling this type of continuous response is slightly more involved than modelling the discrete choice in the discrete pairwise case. First of all, the likelihood must adhere to the binary choice and furthermore be bounded in its support since $y \in ]0,1[$. In contribution [C], we proposed a model using a Beta type distribution with its mean parameterized by the cumulative Gaussian function through the shape parameters $\alpha$ and $\beta$. The general model can be written

$$\begin{aligned}
\sigma_s|\xi_s &\sim \text{half Student-t / Gamma} \\
\sigma_\ell|\xi_\ell &\sim \text{half Student-t / Gamma} \\
\mathbf{f}_k|\sigma_s,\sigma_\ell &\sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}_\mathrm{k}\right), \mathrm{k}\left(\mathbf{x}_\mathrm{k},\cdot\right)_{\sigma_\mathrm{s},\sigma_\ell}\right) \\
\beta(\mathbf{f}_k) &= \nu(1-\mu(\mathbf{f}_k)), \alpha(\mathbf{f}_k) = \nu\mu(\mathbf{f}_k) \\
y_k &\sim \text{Beta}\left(\alpha(\mathbf{f}_k), \beta(\mathbf{f}_k)\right)
\end{aligned} \tag{3.16}$$

Where $\nu$ relates to the precision of the Beta distribution and is not parameterized by $f$. The shape parameters of the Beta distribution is a function of of the GP, and we applied a well-known re-parametrization of the Beta distribution [66]. The mean function of the Beta distribution is given as $\mu\left(\mathbf{f}_k,\sigma\right) = \Phi\left(\frac{f(v_k)-f(u_k)}{\sqrt{2}\sigma}\right)$. The precision term $\nu$ in Eq. 3.16 is inversely related to the observation noise on the continuous bounded responses. In general, $\nu$, can be

viewed as a measure of how consistent the scale is used in a given comparison.[4].

Approximation for this model is primarily based on the Laplace approximation. Despite the Gaussian approximation predictions are problematic due to the Beta distribution. The probability for $y_k > 1/2$ defines the binary choice of option $\mathbf{x}_u$ and must be evaluated by numerical methods, whereas the binary choice itself i.e. the hard choice of either $\mathbf{x}_u$ or $\mathbf{x}_v$ can easily be evaluated due to the symmetric likelihood. However, the expectation of $y$ can be evaluated analytical due to the specific parametrization which provides insight into the effective choice model, a topic for further investigation (see [J]). Alternatively, one may rely on maximum-a-posterior analysis of $p(f|\mathcal{X}, \mathcal{Y}, \hat{\boldsymbol{\theta}})$ which provides a closed form expression [C], but compromises the Bayesian principle.

In the applied versions of the model ([C] and [170]) the warping function, which maps the latent function to the likelihood, has been limited to a single fixed cumulative Gaussian based on the principle from standard choice models. However, a more general warping (or cumulative link) function is the (so far finite) sum of such functions, i.e., $g\left(f(\mathbf{x})\right) = \sum_{i=1}^{N_g} \Phi\left(f(\mathbf{x})|\mu_i, \sigma_i\right)$, which includes finding the parameters of the warping function by evidence optimization. Very recent advances in warped GPs has furthermore shown potential to use non-parametric functions for the warping potentially generalizing many of the outlined models relying on some kind of link or warping function [122].

### 3.2.3 Absolute-Continuous-Bounded Beta Model

Modeling absolute bounded responses, e.g. $y \in ]0;1[$, has relevance to many applications in finance, survival analysis, and here preference learning, where users are asked to rate object on an absolute scale which is inherently bounded. In [J], a Gaussian process model was proposed for modelling continuous ratings on bounded scales for a particular dataset based on Beta and Truncated Gaussian distributions. Thus, the setting strongly resembles the classical regression case with Gaussian likelihood [182] or Student-t likelihoods [223] - but with a bounded support.

[4] In memorandum, it is noted that a similar likelihood has been proposed for generalized linear models but not realized or evaluated in [225]

The model can be written

$$\sigma_s|\theta_s \sim \text{half Student-t / Gamma} \tag{3.17}$$

$$\sigma_\ell|\theta_\ell \sim \text{half Student-t / Gamma} \tag{3.18}$$

$$f_k|\sigma_s, \sigma_\ell \sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}_\mathrm{k}\right), \mathrm{k}\left(\mathbf{x}_\mathrm{k}, \cdot\right)_{\sigma_\mathrm{s},\sigma_\ell}\right) \tag{3.19}$$

$$\beta(f_k) = \nu(1-\mu(f_k)), \alpha(f_k) = \nu\mu(f_k)$$

$$y_k \sim \text{Beta}\left(\alpha(f_k), \beta(f_k)\right) \tag{3.20}$$

The idea is similar to the pairwise Beta model used for relative modelling described in 3.2.2, i.e., parameterizing the mean of the Beta distribution with the latent Gaussian process though a warping function given by the cumulative Gaussian.

The inference is based on the Laplace approximation or the EP approximation in which a numerical integration is used to evaluate the partition function (see [J]) for details. The model is in essence a non-parametric version of the model proposed in [66] also extended to support Truncated Gaussian observations by parameterizing the mean or the mode of the Truncated Gaussian distribution [J].

The models has a strong link to Copular processes (e.g. [232]) which will not be explored further in the present thesis, but is topic for further investigation.

### 3.2.4 Releative-Discrete-General Bradley-Terry-Luce and Plackett-Luce

The pairwise model applying probit and logit models is limited by the amount of information each experiment yields. It is possible to extend both the probit and the logit (Bradley Terry) model to multiple alternatives, i.e. $C$-alternative forced choice forced [220]. However, the generalized probit results in a likelihood which needs to be evaluated analytically, we therefore focus on the extension of the logit model as proposed by Luce [135], which results in analytical likelihoods. The Bradley-Terry-Luce (or ordered generalized logit) is defined as

$$p\left(y_k \in [1:C] \,|\mathbf{f}_k\right) = \frac{e^{f\left(\mathbf{x}_{y_k}\right)}}{\sum_i^C e^{f(\mathbf{x}_i)}}$$

Inference in the Bradley-Terry-Luce (BTL) model is challenging, due to the many dependencies in the likelihood, however, in the Laplace case this does not matter due to the global approximation of the posterior, i.e. there is no

approximation to each term, and the first level approximation simply requires the first and second derivative with regards to all variables entering into the each likelihood term.

The Plackett-Luce (or exploded logit) now effectively extends the BTL in defining a model over permutations given as [178],

$$p\left(\mathbf{y}_k|\mathbf{f}_k\right) = \prod_{j=1}^{C-1} \frac{e^{f\left(\mathbf{x}_{\mathbf{y}_k(j)}\right)}}{\sum_{i=j}^{C} e^{f\left(\mathbf{x}_{\mathbf{y}_k(i)}\right)}}$$

where $\mathbf{y}_k$ is a permutation of the $C$ object in the choice set, e.g. $\mathbf{y}_k = \{4, 2, 3, 1\}$. Analytical probabilistic predictions are currently made on the pairwise level, i.e. resulting in the standard pairwise logit and applying the logit approximation in [182] also used in the GPML toolbox [67]. Probabilistic predictions for a full permutation is based on simulation. Alternative MAP predictions may prove a use full, however, enumerating all possible permutations is still a cumbersome task, which requires further investigation.

The Laplace approximation provides a stating point for examining the Plackett Luce model for preference learning in real-applications, however other approximations should certainly be explored. Some work on the Plackett-Luce models has shown a potential route for such work in terms of a EP version for a non-GP version of the model [80].

### 3.2.5 Additional Models

The specific models outlined above provides a non-exhaustive, but general overview to modelling within the setup defined in Fig. 3.1.

Not yet realised observational models, include the ordinal likelihoods (also known as cumulative link models) previously considered in the GP literature [47] for absolute settings. This is highly relevant to ratings both in the pairwise and absolute case since many experimental paradigms consider e.g. the Likert scale [128], where the scale is divided into discrete intervals.

The observation/response types outlined in Tab. 3.2 covers the ones encountered and found relevant during the construction of the setup, however many more variations may be considered for modelling real-world scenarios, including triangle tests, odd-one-out and other experimental paradigms [16] where the Gaussian process approach may very well provide a flexible modelling framework.

## 3.3 Sequential Design & Sparsity

The motivation for the setup presented in Fig.3.3 is based on elicitation of various aspects of human perceptual and cognition through experimental. However, just one experiment, involving human users, can be both costly and time consumption. In real applications, it is therefore important to minimize the burden for the users and the computational time required. This can be done by minimizing the number of experiments which potentially limits the information. An alternative method is socalled sequential design - or active learning - in which experiments are proposed as the ratings come in and the models gets better. This will optimally ensure that the system does not require more experiments than needed in order to solve the task and find a appropriate model. However, while GP based classification and regression models often perform better or similar to other models regardless of the setting [182], a price is often paired in terms of the computational requirements due to an inherently $\mathcal{O}(N^3)$ scaling in terms of the number of labeled examples. These two aspects can both been seen as a way to reduce the number of inputs, but from two different perspectives outlined as follows;

**Directly:** The experimental design used to obtain the data obviously determines how many observations are obtained. Directly limiting the number of observations can roughly be performed in two ways:

a From a fixed set of labeled points include only a subset, the active set.

b From a (infinite) set of unlabeled inputs obtain only labels/values for a subset (and include in the dataset).

An overview and review is provided in 3.3.1 placing methods applied contribution [I] in context with other methods in the field.

**Indirectly:** Modelling can reduce the effective number of points which is here generally refereed to as sparsity which also was two approaches: [5].

b Inducing: Define the number of effective points (and locate them anywhere according to some measure of performance). An overview and review provided in 3.3.2 placing contribution [H] in context.

a Inferred: Infer the number of points, often a subset of the original inputs. Examples are Support Vector Machines and the Releveance Vector Machines (RVM) [217]. This setting is not considered in the current thesis.

[5] The orthogonal notion of sparsity, obtained by removing individual feature dimensions is not considered here (see e.g. the ARD kernel)

### 3.3.1 Sequential Design & Natural Sparsity

| Sequential Design | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Iterative Acive Set Methods* | Active Learning | | | | | | | | | | | | | |
| | I: Computation | | | | | | | II: Task/Criterion | | | | | | |
| | Plan | | Greedy | | | | | Optimize | | | | Generalization | | |
| Random *<br>IVM *<br>. . . | Approx. * | Exact * | VOI | EVOI | G(E)VOI | CWS | | PoI | EI | UCB | THOMP | Random | Entropy | . . . |
| [121] | | [172] | | [130] | [65] | | | [116][102] | [156][102] | [205] | [215] | | [131][138][113] | |
| | | | I | | | [170] | | [140] | I | | | | | |

Table 3.3: Overview of the sequential design aspects of preference learning. The **second layer** indicates whether we consider the sequential selection of instances from the point of view of scarification (iterative / active set methods) or a sequential experimental design approach denote active learning. The **third layer in the active learning case** illustrates two aspects of active learning, the framework under which the value of an experiment is computed, e.g. by real planning or a greedy approach. The other aspect indicate the criterion which determines the objective of the sequential design, e.g. finding the optimum of a function vs. finding the general function. The second row from the **bottom** indicates relevant references whereas the bottom row indicates contributions or work by co-authors applying principles outlined here.

The direct and simples way to reduce the $\mathcal{O}(N^3)$ complexity of a standard GP classifier is to simply limit the number of training points, the active set. This can be done in a number of ways depending on the actual setting, in particular,

- **Random design** implies randomly selecting a number of candidates from a finite/infitice candidate set (used as reference in [I]).
- **Fixed design** Before experimenting, construct an experimental design with a number of fixed experiments, $K$, using classical statistical methods

of experimental design [158]. This could be performed using factorial designs or combinational design like Latin squares [158]. For standard GP regression with fixed hyper parameters, this may even be specified beforehand by selecting the design that minimizes the posterior/predictive entropy/variance [6].

- **Iterative/Active Set methods** is the general setting where the algorithm (iteratively) either removes or adds **labeled** points based on some criterion from a fixed candidate set, such as the Informative Vector Machine (IVM) and its predecessors [121],[191][192] and the recently proposed approach considering the predictive uncertainty [86]. In general, this principle is characterized by the label being known for all candidate points.

- **Active Learning/Sequential Design** or interactive learning implies that experiments are added sequentially based on some criterion, which is dependent, in some way, on the current model posterior and/or predictive distribution. This setting is typically characterized by the fact that the label of the experiments are unknown before experimentation. This is the case for all intended purposes of the preference learning setup.

The differentiating property between iterative set vs. active learning is thus, whether the label is known for the candidate set or not. While iterative/active set methods are certainly interesting; the main interest is the active learning setting since human labels are often difficult to obtain. Furthermore, the application is typically intended for novel situation where where no labels are available at all. In this situation a clear advantage of GPs over non-Bayesian models is the representation of both posterior and predictive uncertainly, which makes them highly suitable for introducing active learning/sequential design with the purpose to learn as efficient/fast as possible based on some notion of gain in information.

### Sequential Design / Active Learning

This section continues with a general overview of **active learning** or **sequential design** in the GP setting. This includes a general view of many of the stand-alone contribution in the applied GP literature. First, we differentiate between two aspects outlined in Tab. 3.3, namely computation and criterion, which together defines how the value of conducting an experiment is evaluated. An experiment is in the GP setting a number of inputs which is part

[6] Since the predictive variance does not depend on the actual response, $y$, as seen from Eq.3.3

of the experiment, i.e. one, two or more candidate input points collected in $\mathcal{E} = \{\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_E^*\}$.

The main setting considered here is the greedy one-step-ahead selection of experiments[7], thus the aim is to considering the next experiment only - not considering multiple steps ahead which refereed to as planning in Tab. 3.3

Many of the current methods in the GP literature [96][113][31] can be seen as reductions of the following general formulation

$$\underbrace{\underbrace{\frac{\partial}{\partial \mathcal{E}} \int \underbrace{p(y_k|\mathcal{E}_k, \mathcal{D})}_{prior\,predictive} \times \underbrace{U(\underbrace{p(\mathbf{f}|\mathcal{D})}_{prior} \underbrace{p(\mathbf{f}_{\mathcal{E}k}|\mathcal{E}_k, D), p(y_k|\mathcal{E}_k, D)}_{prior\,predictive} \underbrace{p(\mathbf{f}|y_k, \mathcal{E}_k, \mathcal{D})}_{posterior})}_{VOI} dy}_{EVIO}}_{G(E)VIO} \tag{3.21}$$

The G(E)VIO refers to the gradient of the VIO / EVIO function (as mentioned in [172]). The EVOI formulation is nothing more than a formulation of Lindley's expected information of an experiment (see e.g. [131][130, Sec. 12])[8]. The addition of the gradient version G(E)VOI is mainly computational, whereas the possible reduction to VOI can be both an explicit utility choice or computational issue.

Experiments are selected in a greedy fashion by either of the four options outlined

$$\arg\max_{\mathcal{E}} \text{VIO}(\mathcal{E}) \quad \arg_{\mathcal{E}} \text{GVIO}(\mathcal{E}) = 0 \Big|_{\mathcal{E}_0} \tag{3.22}$$

$$\arg\max_{\mathcal{E}} \text{EVIO}(\mathcal{E}) \quad \arg_{\mathcal{E}} \text{GEVIO}(\mathcal{E}) = 0 \Big|_{\mathcal{E}_0} \tag{3.23}$$

where $\mathcal{E}_0$ indicates the starting position of a gradient search. The function and distribution involved are:

- $U(\cdot)$ defines the actual criterion/utility based on all or a subset of the listed arguments in Eq. 3.22.

[7] This is strictly not a requirement since many of the methods considered can be generalized to a dynamic programming setting however of course at an increased cost. See e.g. [71][172] for examples for such an approach

[8] In the view of inference, not the decision theoretic view in which a decision loss is also defined [130, Sec. 4]

- $p(\mathbf{f}|\mathcal{D})$ is the current prior over the function (after seeing the data $\mathcal{D}$, but before seen a new candidate experiment).
- $p(\mathbf{f}_k|\mathcal{E}_k, \mathcal{D})$ given the current prior, this is the predictive distribution for the function values for the inputs in the candidate experiment. This is called the prior predictive distribution.
- $p(y_k|\mathcal{E}_k, \mathcal{D})$ based on the $p(\mathbf{f}_k|\mathcal{E}_k, \mathcal{D})$ and the likelihood function this is the predictive distribution for the observed variable $y$.
- $p(\mathbf{f}|y_k, \mathcal{E}_k, \mathcal{D})$ after seeing the response $y_k$ for the experiment $\mathcal{E}_k$, we can form the posterior, potentially needs to be approximated.

The criterion defining the utility function, $U(\cdot)$, is obviously a critical algorithm/model choice and often dependent on the particular use of the resulting model. If the aim is to learn a 'good' model for the entire (feasible) input space then, one should chose a criterion which ensures good generalization, often formulated in terms of the parameters, i.e. $f()$[9]. If the task, on the other hand, is to find input instance with the highest latent function value (or observed response), we should select a criterion which ensures that we find the maximum; not ensure that we learn the overall function. In the preference learning applications there is a need for both and in particular the combination of both (see for examples of specialized [44] utilities in the Bayesian linear model setting).

In terms of **generalization** the machine learning litterateurs has considered a large number of proposals for different models ranging from disagreement by query-by-committee [195], to minimum margin methods [218], and criterions based on statistical notions of uncertainty models such as [49][137]. The latter is a good starting point for the Gaussian process view point, which is based on the general notion of entropy [50],

**(E)VIO Entropy** : is a natural measure of information of an experiment which was proposed by Lindley in early work [129], where the change in entropy between prior and posterior over the parameters (here $f$) was considered. Thus, the criterion $U(\cdot)$ is evaluated as

$$U\left(p\left(\mathbf{f}|y_k, \mathcal{E}_k, \mathcal{D}\right), p\left(\mathbf{f}|\mathcal{D}\right)\right) \equiv S\left(\mathbf{f}|y_k, \mathcal{E}_k, \mathcal{D}\right) - S\left(\mathbf{f}|\mathcal{D}\right) \tag{3.24}$$

$$= \int p\left(\mathbf{f}|y_k, \mathcal{E}_k, \mathcal{D}\right) \log p\left(\mathbf{f}|y_k, \mathcal{E}_k, \mathcal{D}\right) d\mathbf{f}$$

$$- \int p\left(\mathbf{f}|\mathcal{D}\right) \log p\left(\mathbf{f}|\mathcal{D}\right) d\mathbf{f} \tag{3.25}$$

[9] The thesis will not consider utilities regarding the hyperparameters, $\boldsymbol{\theta}$, but [113] considers this aspects in a Normal regression case. Furthermore, we do not take into account the input density of $x$, $p(x)$

While seeming simple, the first factor in the first integral implies evaluating the posterior for each possible outcome $y$, but by taking the expectation EVIO this can be avoided as seen by Lindley [131], using information theoretic arguments, and the EVIO becomes,

$$EVIO\left(\mathcal{E}_k\right) \equiv \iint p\left(\mathbf{f}_k|\mathcal{E}_k,\mathcal{D}\right) p\left(y_k|\mathbf{f}_k,\mathcal{D}\right) \log p\left(y_k|\mathbf{f}_k,\mathcal{D}\right) dyd\mathbf{f} \qquad (3.26)$$
$$- \int p\left(y_k|\mathcal{E}_k,\mathcal{D}\right) \log p\left(y_k|\mathcal{E}_k,\mathcal{D}\right) dy$$

Thus, no posterior calculation is to be calculated for every possible outcome, however, the first integral still does not (often) result in closed form expressions. Numerical or analytical approximations is often required (currently the first option is a part of the setup). A specific approximation was proposed recently for probit based models [93]. For the pairwise case (not GPs) this utility was first considered in [75] .

An equivalent information theoretic view[10] was taken to yield the same result for the Normal regression case [113] showing general properties of this in standard regression. They also considering the implications imposed by the hyperparameters in sequential design.

**VOI** is often considered in the entropy case by only taking into a account the entropy of the predictive distribution over $f$, and the utility becomes

$$U\left(p\left(f_{\scriptscriptstyle k}|\mathcal{E}_k,\mathcal{D}\right)\right) \equiv S\left(f_{\scriptscriptstyle k}|\mathcal{E}_k,\mathcal{D}\right) \qquad (3.27)$$

which does not depend on the observation, $y_k$, and reduces to the VOI setting. For multiple inputs this requires evaluating the entropy of a multivariate Gaussian, which is available in closed form [50]. In the case of a single value likelihood this corresponds to the predictive variance of $f^*$.

A particular case is obtained for standard regression with a Normal likelihood, i.e., a single $f^*$,. Here the entropy is simply a monotone function of the predictive variance which does not depend on the observation, and the design can be made before observing any observations. This is also known as variance reduction or uncertainty sampling.

**Response entropy:** The approach considered so far was framed in terms of the parameter space, i.e., $f(\cdot)$. A different approach is based on the predictive aspect, thus considering the change in entropy between the predictive distribution of a particular response $y^*$ before and after conducting the experiment. However, the simplest version of this is the VOI setting in which the predictive entropy/variance can be used as the utility, $U \equiv S\left(y_z|\mathcal{E}_k,\mathcal{D}\right)$ (see e.g. [233] for a short discussion).

[10]The easiest way to arrive at Eq.3.26 is by using information theoretic arguments with conditional entropies and mutual information.

In the case of **optimization**; the field of design and analysis of computer experiments (DACE) and global optimization has in many years made use of Gaussian process models to perform global optimization [96][156][101]. The machine learning community has adopted many of these methods of which we will review a few, most relevant to the preference learning setup.

**(E)VIO Probability of Improvement (PoI)** is an old [116] yet intuitive measure [101], when attempting to optimize the latent function $f$. For the regression case, i.e. a single candidate point in the experiment, it is simply defined as $P\left(f\left(\mathbf{x}^*\right) \geq f^{\max}\right)$ with $\mathbf{x}^* \in \mathcal{E}$ and $f^{max}$ is the current maximum of the latent function. For regression this this amounts to evaluating a simple integral, resulting in a probit function. Seemingly this approach has not been investigated together with the relative observation methods, such as pairwise comparisons.

**(E)VIO Expected Improvement (EI)** is a criterion for selection new candidate points taking in to account the potential change for each candidate points [156][101][172]. Given the focus on the latent function is typically applied in the VIO setting (and often in normal regression), i.e., based on the predictive distribution for $f(\cdot)$. First consider the improvement, $I$, between the candidate point, $\mathbf{x}^*$ and the point with the current maximum latent value, $f^{max} = f(x^{max})$, thus the improvement is defined, $I(\mathbf{x}^*) = f(\mathbf{x}^*) - f(\mathbf{x}^{max})$. Only considering the positive improvement and evaluation the expectation yields,

$$\mathrm{EI} = \int_0^\infty I \times p(I) dI \tag{3.28}$$

which is defined as the expected improvement [102][96][156]. Since $f(\cdot)$ and $f(x^*)$ are jointly Gaussian, this results in a closed form expression for the utility, namely [96]

$$U \equiv \sigma_{EI} \cdot \mathcal{N}\left(\frac{\mu_{EI}}{\sigma_{EI}}\right) + \mu_{EI} \cdot \Phi\left(\frac{\mu_{EI}}{\sigma_{EI}}\right)$$

where $\mu_{EI} = \mu^* - \mu_{\max}$ and $\sigma_{EI}^2 = (\sigma^2)^* + \sigma_{\max}^2 - 2\sigma_{,\max}{}^*$ where $\mu^*$ is the predictive mean for the candidate points and the $\sigma$' are the variances and covariances between the currently best input and the candidate.

It is noted that the VIO depends both on the mean and the variance and gets bigger with both, i.e., an trade off between selection candidates with high mean value and candidates with high variance. This is often referred to as exploration and exploiting trade-off.

Contribution [D] applies this approach in an interactive system for a preference model with bounded absolute responses. Other applications in machine learning includes [79] for standard regression. [31] applies a variant

of the expected improvement in a EVIO setting for a pairwise likelihood which is seemingly computation infeasible for large problems.

For completeness, we notice that one could formulate the expected improvement on a continuous response variable, i.e., maximizing $y$ instead of $f$. In GPs where the latent function often close follows the response variables through the mean or mode, it is questionable if this is necessary and even feasible for most specialized likelihoods. Similar ideas has been proposed in for example [44, Sec 2.4] for standard Bayesian linear models.

**UCB-GP** : The UCB (Upper Confidence Bound) algorithm is a popular GP algorithm [205][79] for optimization tasks which effectively combines the predictive variance and means in order to find the maximum of the latent function. While often applied in a heuristic manner, recent and required theoretical work has shown regret bounds for the UCB-GP algorithm [205].

**Thompson GP** : Thompson Sampling is an old method not specific to Gaussian processes and can be applied in more versatile settings than optimization with GPs [215][214]. However, in the optimization setting for GP regression; the Thompson approach amounts to sampling a function for the full experiment set, $\mathcal{E}$. Hence,

$$\begin{aligned}
\hat{\mathbf{f}}^* &\sim p\left(\mathbf{f}^* | \mathcal{E}, \mathcal{X}, \mathcal{Y}, \theta\right) \\
\hat{\mathbf{f}}^* &= \left[\hat{f}\left(\mathbf{x}_1^*\right), \hat{f}\left(\mathbf{x}_2^*\right), ..., \hat{f}\left(\mathbf{x}_E^*\right)\right] \\
\mathbf{x}^* &= \arg\max_{\mathbf{x}^*} \hat{\mathbf{f}}^*
\end{aligned}$$

This has been considered in a few studies for standard regression [132], but not for the particular use with relative likelihoods, where the procedure must be altered.

The Thompson sampling approach introduces an implicit exploration elements, which is similar to **Criterion Weighted Sampling (CWS)**, which is a heuristic proposal for providing an (extra) exploration element to any of the criterions above. The principle simply draws a index of the candidate point $\mathbf{x}_i \in \mathcal{E}$ as follows,

$$i \sim Multinomial(\lambda)$$

where

$$\lambda\left(\mathbf{x}_i{}^*\right) = \frac{(E)VOI\left(\mathbf{x}_i{}^*\right)}{\sum_{i'} (E)VOI\left(\mathbf{x}_{i'}{}^*\right)}$$

The principle is like many of the other sequential methods, without any theoretical foundation, and still awaits a full empirical validation. It has so far been applied by co-authors in [170].

### 3.3.2 Induced Sparsity

Reducing the number of labeled inputs, $N$, by clever selection of the actual points as outlined in the previous section is one way of reduce the computational complexity. Another way is to use socalled inducing points proposed for the standard regression case in [199]. By effectively introducing a set of pseudo-inputs and from these predicting the actual function variables used in the likelihood function the complexity can be reduced to $\mathcal{O}(N \cdot L^2)$, where $L$ is the number of inducing points. The great advantage of this approach is that the pseudo points can freely be located as part of the model inference including ML-II / MAP-II approach. In high dimensions and relatively many pseudo points, $L$, this becomes a daunting task, however the increased predictive speed may justify the increased cost in approximating the posterior.

The original idea [199] has later been extended and covered in a vast body of literature [179][164][226][222] where the current general view of sparse GPs is the fully independet (training) condition FI(T)C and the partially independent (training) conditional, PI(T)C, approximations framed in the excellent review [179]. These generalizations can be realised in standard GP models by simply constructing effective covariance functions [222] which for FI(C)T is equivalent to the original pseudo-input formulation for standard regression and classification. The PI(T)C is slightly more involved approximation which does not assume all the conditional function values (random variables) to be independent, but typically divided into blocks of inputs which are when conditionally dependent on the block level. An overview of this particular approach is given in [222, p.23-24].

#### 3.3.2.1 ...for Pairwise Likelihoods

There are in essence (at least) three approaches to obtain sparse approximations to the pairwise model: the pseudo-input, the FITC and the PITC approach. Unlike the standard regression case none of these are necessarily equal. The main aspect to recall is that the likelihood depends on two function values and introducing the inducing points under the FITC approximation will render these conditionally independent. Considering the PITC approximation will allow for correlation inline with the original model, however, if all pairwise observation are made and strictly blocking into pairwise blocks then the PICT will after all combinations - or even just a reasonable amount - results in a dense posterior with a covariance matrix of size $N \times N$ as the original model.

With this in mind the first attempt to derive a sparse version of the model in

Sec. 3.2 starts with the original pseudo input formulation which is described in contribution [G]. Generally this follows the ideas in [199], i.e., given a set of pseudo-inputs $\bar{\mathbf{X}}$, the functional values $\bar{\mathbf{f}}$ comes from a GP like the real latent function values $\mathbf{f}$. Therefore, we can directly place a GP prior over $\bar{\mathbf{f}}$, i.e., $p\left(\bar{\mathbf{f}}|\bar{\mathbf{X}}\right) = \mathcal{N}\left(\bar{\mathbf{f}}|\mathbf{0}, \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}\right)$, where the matrix $\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}$ is the covariance matrix of the $L$ pseudo-inputs collected in the matrix $\bar{\mathbf{X}} = [\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_L]$.

The overall idea of the pseudo-input formalism is now to refine the likelihood such that the real $\mathbf{f}$ values which enter directly in the original, non-sparse likelihood function (through $\mathbf{f}_k$), exist only in the form of 'predictions', i.e. conditional distribution, from the the pseudo-inputs $\bar{\mathbf{f}}(\bar{\mathbf{X}})$. It can easily be shown that the sparse pairwise likelihood is of the similar form as the original, namely

$$p\left(y_k|\mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \bar{\mathbf{X}}, \bar{\mathbf{f}}, \boldsymbol{\theta}\right) = \Phi\left(y_k \frac{\mu_{u_k} - \mu_{v_k}}{\sigma_k^*}\right) \tag{3.29}$$

where $\mu_k = [\mu_{u_k}, \mu_{v_k}]^\top$, $\mu_{u_k} = \mathbf{k}_{u_k}^\top \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}}$, $\mu_{v_k} = \mathbf{k}_{v_k}^\top \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}}$ and

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{u_k u_k} & \sigma_{u_k v_k} \\ \sigma_{v_k u_k} & \sigma_{v_k v_k} \end{bmatrix} = \mathbf{K}_{\mathbf{x}_k \mathbf{x}_k} - \mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k}^\top \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k}$$

Furthermore, $(\sigma_k^*)^2 = 2\sigma^2 + \sigma_{u_k u_k} + \sigma_{v_k v_k} - \sigma_{u_k v_k} - \sigma_{v_k u_k}$, which all together results in the pseudo-input likelihood

$$p\left(y_k|\mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \bar{\mathbf{X}}, \bar{\mathbf{f}}, \boldsymbol{\theta}\right) = \Phi\left(z_k\right), \tag{3.30}$$

where $z_k = y_k \left(\mathbf{k}_u - \mathbf{k}_v\right) \mathbf{K}_{\bar{\mathbf{X}},\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}} / \sigma_k^*$. Thus, the likelihood is still a cumulative Gaussian, however, with elements 'predicted' from the pseudo inputs via the covariance between pseudo inputs $\bar{\mathbf{X}}$ and real inputs, $\mathbf{k}_u$ and $\mathbf{k}_v$ - and the $L \times$ Ł covariance matrix between pseudo-inputs, $\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}$. This results in a scaling of $\mathcal{O}(N \cdot L^2)$ compared to the $\mathcal{O}(N^3)$ scaling of a standard GP.

The primary issue in this type of sparsity is the inference of the $L$ pseudo inputs. Like traditional models with pseudo inputs, the setup considers only ML/MAP-II type inference, however considering the number of parameters in say 5 dimension and 10 pseudo inputs, the solution is very likely to be prone to local minimum which is left for further investigation. Details regarding the ML/MAP-II inference can be found in contribution [G].

## 3.4 Evaluation Methods

Evaluating the models outlined above is an important part of the modelling and system aspects. This section provides a overview of the methods and means considered in the contributions divided by the response types defined in Sec. 3.2.

**Absolute** For the models presented in Sec. 3.2 relying on absolute ratings the evaluation is similar to standard Gaussian process evaluation methods [182]. The recent review [224] provides a excellent and comprehensive overview of evaluation methods, however, we will simply mention a few standard metrics commonly applied in contributions,

Predictive log-likelihood: The predictive distribution is the basis for many Bayesian evaluation metrics, i.e. a prediction $y^*$, may be obtained by considering the mean or the mode of the predictive distribution $p(y^*|\mathcal{D}, x^*)$. However, the most commonly used metric is the predictive log-likelihood, i.e. $\log p(y^*|\mathcal{D}, x^*)$.

Mean Square Error (MSE) is the natural error metric in least squares regression due to the normally distributed noise assumption. Based on a point prediction, $y^*$ , (e.g. the mode or mean of the predictive distribution) it is defined for all test input $K$, as usual, i.e, $MSE = \frac{1}{K}\sum_{k}^{K}(y - y_k^*)^2$ whereas Mean Absolute Error (MAE) is typically applied, when the noise is not assumed to be Normally distributed, for example for bounded regression models outlined above. It is defined by $MAE = \frac{1}{K}\sum_{k}^{K}|y - y_k^*|$

**Relative** Evaluating rankings is a general topic in information retrieval (see e.g. [144] for retrieval), where search results are subject to many evaluation studies. In the current setting, which effectively spans both such information retrieval settings and smaller elicitation problems, one can generally differentiate between two settings, namely

**Ranking** In the ranking scenario, we are generally interested in ordering a fixed set of objects, not predicting for new objects. The evaluation metric is whether the ordering is correct or not, not how the model predicts to new unseen objects. In this setting, cross-validation is performed by for example holding out a single pairwise observation between two objects and training on the remanding observations (comparisons). A test is then performed to check if the heldout comparison is explained by the ordering. This means that the held out aspect is the relation, not the objects per say.

**Predictive** The main interest in the current report relates to the predictive setting, in which predictions are made for left out objects. Creating proper training, test and validation splits is an added complication in evaluating predictive ranking scenarios since one must hold-out both the object and the relations to other objects in order to obtain a valid cross-validation.

Choosing performance metrics for ranking of objects is a ongoing research field in information retrieval [144], and also heavily relates to the setting of preference learning. Common evaluation metrics include,

**Predictive log-likelihood:** The predictive log-likelihood is again a sensible measure, however, not directly applied in the contributions on relative but exploited indirectly in action learning settings.

**Error rate**
The simplest and most often used metric in the contributions is the error rate of for example pairwise relations. These predictions are given by the predictive distribution or can in most cases be obtained directly from f. Considering $M$ observations, the error rate is simply defined by $ER = \frac{M_{incorrect}}{M}$. This gives a metric similar to the error rate in standard classification

**Kendalls tau**
Evaluating the correlation between two different orderings of $M$ objects can be performed using Kendalls tau defined by
$\mathrm{KT} = 2\frac{M_{correct} - M_{incorrect}}{M(M-1)}$, with $\mathrm{KT} \in [-1; 1]$. For a perfect alignment between rankings $\mathrm{KT} = 1$, and for a perfectly reverse ranking $\mathrm{KT} = -1$. If the two rankings are independent $\mathrm{KT} = 0$, for example if one of the rankings is permuted randomly.

**Precision/Recall and DCG/nDCG**
A particular application of the ranking scenario is the retrieval of objects by their ordering. For this purpose metrics such as Recall and Precision are typically applied. Other similar aspects are the Discounted Cumulative Gain (DCG) and normalized DCG which weights the relevance by its position in the ranking. We refer to [144] for an introduction to this aspects of ranking and retrieval.

## 3.5 Alternatives

Preference learning and ranking with Gaussian processes is not by any means the standard approach to modelling ratings and ranking - on the contrary. A number of other approaches has been proposed for both ranking and specialized regression. The following provides a short overview of the most common approaches with focus on the machine learning literature.

**Generalized Linear Models (GLM)** The framework of GLM is very well-described in standard statistic literature (see e.g. [145]) with the comfort of having standard hypothesis tests and diagnostic tools. Bounded regression GLM models has for example been proposed in [66] similar to the absolute bounded GP model proposed here. Pairwise models has been considered in [225] simply starting from logistic regression based on the difference between two vectorial inputs i.e. the latent function is given as standard regression on the difference $f_k = \mathbf{w}^\top \times (\mathbf{x}_{u_k} - \mathbf{x}_{v_k})$, which is used as a baseline in contribution [E].

Non-linear versions of this may be constructed easily by the kernel trick [26]. For example can a (presumably novel) version of the pairwise GLM model be constructed resulting in a effective kernel similar to the recently proposed Pairwise Judgment Kernel (PJK) [93]. This kernel can then be applied in for example classic kernel logistic regression without any further complications than calculating a special kernel between two sets of pairwise comparisons, i.e. $k\left(\{\mathbf{x}_u, \mathbf{x}_v\}, \{\mathbf{x}_u, \mathbf{x}_v\}'\right)$.

**Neural Networks** Neural networks (NN) [25] for pairwise ranking has been proposed in for example [37], and extended for listwise ranking using the Plackett-Luce model as the basis in [40] and [236].

**Support Vector Machines** Application of SVM for ranking (so-called SVM-rank) has been considered in [241] and [100] in general framework of structure output prediction with SVM. This method is for example applied [235] for metric learning in music.

## 3.6 Summary & Perspectives

This chapter provided an overview of the general setup for ranking and ratings with Gaussian processes, supporting different experimental paradigms, such as pairwise comparisons and bounded ratings. The focus was especially on the

application in real-world settings requiring sequential design, efficient and robust inference methods with application in real audio, music, and other systems.

The framework can at present be applied in many settings, and further development, beyond the scope of this thesis, will include further investigation of sequential designs for Gaussian process based models in real-world scenarios. The many sequential design principles outlined are in many regards heuristics which calls for further theoretical justification in line with recent work on the UCB-GP algorithms [205].

Further development is required for effective inference in the BTL and Plackett-Luce model, in particular extending the current Laplace approximations to other inference methods, like expectation propagation or variational Bayes. A further modelling aspect is the relation between the bounded regression models proposed here with the general framework of Copular processes [232], which may generalize the ideas presented here and in contribution [J].

In order for the GP models to become standard tools not only in machine learning community, but also in established statistical domains such as sensormetric and general statistical hypothesis testing, there is a need to develop statistical test and inference methods ensuring robust results and interpretability. Some progress has been made in this regard such as [59] proposing a ANOVA style analysis based on specialized covariance function. However, the Gaussian process approach still lacks robust estimation and interpretation of for example explanatory variables.

An aspect not considered in dept is the multi-task or collaborative filtering setting in which multiple users collaborate to lean their individual preferences or other aspects (also known as transfer learning). Recent advances (see e.g. [93]) has proved that preference learning with GP's can be scaled to such settings by employing similar ideas as outlined in this section. While truly large scale systems will probably refrain from GP based models, collaborative systems based on GPs may very well find its way into smaller domains which can surely benefit from collaboration such as optimization of audio reproduction systems [I] and hearing aid personalization [170][23]. This, however requires further research into the current multi-task and hierarchical Bayesian models [93][24] to quantify the behavior and properties as initiated in [93][24].

Chapter 4

# Predictability of User Context

Knowing the current and future context of users is an important element towards truly personalized systems, where user's information needs changes according to for example location and social setting. With modern portable devices, such as smartphones, it has become feasible to monitor and record at least some aspect of human context such as human location and social interactions, which makes it possible to characterize humans (or the users from the system perspective), based on aspects of location and proximity to other people. A more subtle aspect is the predictability of the context such as location, i.e. can the location be predicted based on the past.

In this chapter, we consider the particular task of predicting the next context state $X_{n+1}$, such as location, conditioned on previously observed states resulting in one-step-ahead forecasting of discrete time series. A multitude of machine learning models/algorithms can be considered for this task ranging from N-order Markov models [50], Hidden Markov Models [180, 26] or neural networks [25] and so on. Such machine learning algorithms, both unsupervised clustering or supervised classification, can in essence be seen as a way to reduce or compresses the often noisy and high dimensional data into something less complex with the aim to provide a representation which has a higher interpretability or usability. The simplest example is possibly binary classification where data living in a potentially high dimensional input space is grouped by a classifier into two classes. By assigning a simple label to a complex high dimensional represen-

tation of an object, we in effect provide a compression of the raw information. The predictability can then be determined by the number of correctly classified assignments.

Instead of having to decide on a model, estimating parameters and performing the predictions, this chapter addresses the problem from another angle. Instead of actually performing the compression as outlined above, we apply the tools usually considered when describing more general aspects of compression, namely information theory. Based on these, we estimate bounds for the predictability, based - not on the number of correctly classified predictions - but on a bound derived from information theoretic measures. This results in bounds for the context predictability, e.g. location, and provides a general characteristic of human behavior and of a given user. The methods are largely variations of the approach suggested in [201] with some refinements and extensions.

**Outline:** this chapter continuous in Sec. 4.1 with a basic overview of information and randomness from an information theoretic view. Sec. 4.1 continuous to describe a relatively recent view on predictability based on measures of information. This results in so-called predictability bounds and the chapter provides a simple proof of these bounds in Sec 4.2.

## 4.1 Basic Measures of Information

We consider the problem of quantifying the predictability obtainable in a discrete time process, $\mathbf{X} = (X_1, X_2, .., X_n)$, where $n$ is the time index and $X$ is the state variable with $M$ discrete states. This is to a large degree motivated by previous work on predictability, complexity and learning see e.g. [**?**]),

Here predictability is defined as the probability of an arbitrary algorithm correctly predicting the next state. Hence, given the history, the main distribution of interest is $P(X_{n+1}|X_1, X_2, .., X_n)$. In the case where we have no information regarding the history, the distribution naturally reduces to $P(X)$. When $P(X)$ is uniform, i.e., each of the $M$ states have the same probability of occurring, the Shannon entropy (in bits) is defined as $S^{rand}(X) = \log_2 M$.

In the case where the distribution of $X$ is non-uniform, the entropy is given as

$$S^{unc}(X) = -\sum_{n=1}^{N} p(x_n) \log(p(x_n)) \tag{4.1}$$

with $p(x) = P(X = x)$. This in turn represents the information available when no history is available, hence, the acronym unc (uncorrelated).

The entropy rate of the process, or the average number of bits needed to encode the states in the process, can be estimated taking into account the complete history. This is done by defining the stationary stochastic process $\mathbf{X} = \{X_1, X_2, ..., X_N\}$, which have the entropy rate [50] defined as

$$S(\mathbf{X}) = \lim_{N\to\infty} \tfrac{1}{N} S(X_1, X_1, ..., X_N) \tag{4.2}$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} S(X_n|H_{n-1}) \text{ Chain rule for entropy)} \tag{4.3}$$

where the variable $H_{n-1}$ at time step $n$ is $H_{n-1} = \{X_1, X_2, ..., X_{n-1}\}$, i.e. the history. Generally, we have that $0 \leq S^{true} \leq S^{unc} \leq S^{rand} < \infty$, where $S^{true}$ denotes the true entropy rate of the process.

**Predictive Information** is an interesting measure which can immediately be derived from the already measures namely the so-called predictive information as defined in [21]. It is defined as

$$I_{pred} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{n} S(X_n) - S(X_n|h_{n-1}) \tag{4.4}$$

$$= S^{unc}(X) - S(\mathbf{X}) \tag{4.5}$$

In effect $I_{pred}$ represents the mutual information between the distributions corresponding to knowing and not knowing the past history. Hence, it quantifies the fundamental information gain in knowing the (complete) past when predicting the future (one step ahead).

### 4.1.1 Entropy Rate Estimation

A challenge in using these information theoretic measures based on real and unknown processes is the estimation of the entropy rate, $S(\mathbf{X})$. A number of ideas has emerged based on compression techniques such as Lempel-Ziv (LZ) (including string matching methods) and Context Weighted Trees for binary processes. [70] provides a general review of such methods. An appealing aspect of these non-parametric methods is that it avoids directly limiting the model complexity as would be necessary if we applied parametric or semi-parametric models. One of the simplest entropy rate estimators are LZ based estimators as described in [111, 70] and also applied in [202]. The entropy rate estimate for a time series/process, $\mathbf{X}$, of length $N$ is given by

$$S(\mathbf{X})^{est} = \left[ \frac{1}{N} \sum_{n=1}^{N} \frac{L_n}{\log_2 n} \right]^{-1} \tag{4.6}$$

where $L_n$ is the length of the shortest substring starting at time step $n$, which has not been seen in the past. The consistency under stationary assumptions is established in [111].

## 4.2 Bounds on Predictability

The entropy rate itself is a difficult measure to interpret and understand from an predictive and algorithmic point of view. In order to provide a measure directly expressing the predictability, we start from Fano's inequality [64][50] relating the conditional entropy of a variable with the probability of seeing an error, $P^e$, or more relevant the probability of a success, $P^s$. Thus, $P^s$ expresses the probability that the prediction $\hat{X}_n$ equals the correct state $X_n$ as given by, $P^s = P(X_n = \hat{X}_n) = 1 - P^e$, when prediction $X_n$ based on on another variable, namely the history (variable) $H_{n-1}$ with outcomes $h_{n-1} \in \mathcal{H}$ where $\mathcal{H}$ denotes all possible past/historic outcomes. Fano's inequality now reads [50],

$$
\begin{aligned}
S\left(X_n|H_{n-1}\right) &= \sum_{h_{n-1}\in\mathcal{H}} P(H_{n-1}=h_{n-1}) S\left(X_n|H_{n-1}=h_{n-1}\right) && (4.7)\\
&\leq H_b\left(e\right) + P_n^e \log_2\left(N-1\right) && (4.8)\\
&= H_b\left(s\right) + \left(1-P_n^s\right)\log_2\left(N-1\right) && (4.9)\\
&= -P_n^s \log_2\left(P_n^s\right) - \left(1-P_n^s\right)\log_2\left(1-P_n^s\right) + \left(1-P_n^s\right)\log_2\left(N-1\right) && (4.10)\\
&= S_{FANO-S}\left(P_n^s\right), && (4.11)
\end{aligned}
$$

where the binary entropy function with a positive outcome is defined as

$$
S_b\left(P_n^s\right) = -\left(1-P_n^s\right)\log_2\left(1-P_n^s\right) - P_n^s \log_2\left(P_n^s\right). \tag{4.12}
$$

In [202] this positive version of Fano's inequality is proved from scratch.

Assuming stationarity, the definition of entropy rate is given as, $S\left(\mathbf{X}\right) = \lim_{N\to\infty} \frac{1}{N} S\left(X_1, X_2, ..., X_N\right)$,

thus

$$
\begin{aligned}
S(\mathbf{X}) &= \lim_{N\to\infty} \tfrac{1}{N} S(X_1, X_2, ..., X_N) \quad \text{Definition of entropy rate} && (4.13)\\
&= \lim_{N\to\infty} \tfrac{1}{N} \sum_{n=0}^{n} S(X_n|H_{n-1}) \quad \text{Chain rule for entropy} && (4.14)\\
&\leq \lim_{N\to\infty} \tfrac{1}{N} \sum_{n=0}^{n} S_{FANO}(P_n^s) \quad \text{Fano} && (4.15)\\
&\leq S_{FANO-S}\left(\lim_{n\to\infty} \tfrac{1}{N} \sum_{n=0}^{N} P_n^s\right) \quad \text{Concave} && (4.16)\\
&= S_{FANO-S}\left(\langle P_n^s \rangle_n\right) && (4.17)\\
&= S_{FANO-S}(P^s). && (4.18)
\end{aligned}
$$

Thus, given the entropy rate of a process, $S(\mathbf{X})$, the one-step-ahead predictability can be bounded from above by, $P^s$, solving the following equality $S(\mathbf{X}) = S_{FANO-S}(P^s)$ for $P^s$. Given the monotone behavior of $S_{FANO-S}$ this provides a unique solution, and can easily be found by standard numerical methods.

### 4.2.1 Upper Bound

Given the estimated entropy rate of the observed process, $S(\mathbf{X})^{est}$, we may directly find the upper bound on the one-step ahead predictability, $P^{s,max}$, solving Eq.4.13. This upper bound is the main focus of [201][202] and contribution [A] and [B]. The interpretation of the this predictability bound should be performed with the entropy estimation in mind, since the provided estimates defines the quality/accuracy of the bound.

### 4.2.2 Lower Bound

The bound proposed by [201] is the upper bound on the predictability, however, it is equally important to consider a lower bound in order to for example evaluate the triviality of the problem, i.e., the upper bound may be close to one for trivial/constant processes. In order to obtain such a lower bound, it a natural choice is to consider the simplest model adhering to the one-step-ahead setting, namely a first order Markov process. Thus, the probability of the next state being $r$ is given by a transition probability $P_{rs} = P(X_{n+1} = r|X_n = s)$

$$
P(X_{n+1} = r) = \sum_{r\in R} P_{rs} P(X_n = s)
$$

The entropy rate of this process is given in closed form as $S(\mathbf{X}) = -\sum_n \sum_j \mu_i P_{ij} \log P_{ij}$, where $\mu_j$ is the stationary distribution of the process which can be found by solving the following equations, $\mu_j = \sum \mu_i P_{ij} \ \forall \ j$.

Thus, given the general bound in Eq. 4.13-4.18, it is possible to find the maximum predictability given this simple model. However, this is still a bound and a more direct measure is the actual predictability of this minimally complex model, i.e., the actual percentage of correctly classified states. This was applied in [A] where classification is based on standard maximum likelihood, i.e., $\arg\max_r P_{rs}(X_{n+1} = r | X_n = s)$.

With the lower bound in place, it is possible to compare the lower and upper bound for the given process, $\mathbf{X}$. In combination with the measure of predictive information in Sec. 4.1, this indicates the real gain in applying a more advanced predictive models compared to the simpler one and effectively determines the operating range of any sensible algorithms.

## 4.3 Summary

The information theoretical view on predictability provides a relatively simple and easy way to obtain bounds on the one-step-ahead predictability of discrete time series. These bounds can then be applied in engineering systems for e.g. resource allocation or in information processing systems as outlined in Chapter 1 for providing context aware systems or for characterizing users by their predictability.

Chapter 5

# Summary & Conclusion

The overall aim of the thesis was to investigate elements of the general system described in Chapter 1. The focus has been the on integration of bottom-up and top-down information eliciting and modelling of perceptual and cognitive aspects and finally quantifying the predictability of context. The thesis first of all contributes with a holistic and general system perspective towards the integration of top-down and bottom-up for audio and music organization presented in 1. The three focus areas, outlined in the summary report - and considered in the contributions - each contribute with new methods and/or results relevant to the field of audio and music organization. These focus areas (and related contributions) are shortly summarized and concluded individually.

**Computational representation of music**
The first aspect under investigation was the integration of top-down and bottom-up views of music. To this end a general review of bottom-up audio features was presented in the summary report (Sec.2.2) focusing on low-level features and how they can be representation on a song level for use in machine learning algorithms. The summary report further outlined how these two views can be integrated in a single model based on probabilistic topic models. This was described in some detail (Sec.2.4) providing a natural route from the well known multi-modal pLSA to the multi-modal LDA model, which is evaluated in H.

In conclusion, it was found that the bottom-up representation, based on audio

feature and lyrics, was more aligned with genre based tags than personal or emotional tags. It was found that the joint representation including tags, lyrics and audio performed best in a genre and style classification task, as expected. The combination of lyrics and audio improved the classification performance over individual audio and lyric representation.

**Preference learning**
The primary focus of the thesis has been the realization of a general Bayesian and modular setup for learning and modelling preference and other cognitive aspects (such as the expressed emotion music), which supports specialized experimental paradigms and observation types, ranging from special bounded regression and ranking of objects.

The summary report provided a general overview of the realized setup and placed the individual contribution in the general context of learning and modelling preference and other cognitive aspects.

- Contribution [C] proposed and realised a novel likelihood model for letting the user convey a degree of preference for paired comparisons. Based on simulations on a toy problem it was concluded that learning with such as response is faster than a standard binary response type also under adverse noisy conditions.

- Contribution [G] proposed and derived a sparse likelihood model which allows scaling the pairwise likelihood to larger problems. Based on both toy and a relatively simple real-wold data, it was concluded that the sparse model is an effective tool in pairwise learning by comparing it to a standard GP.

- Contribution [D] proposed to use pairwise comparisons in modelling and predicting personal music preference, modelling this by audio feature and the probability product kernel. The approach was evaluated on a small three-genre dataset, and it was concluded that the model was able to predict the preference, but that a larger dataset is required to provide more support for this claim.

- Contribution [F][E] considered the modelling and representation of expressed emotion in music. The contributions demonstrated how to elicit and model such complex cognitive task in a novel and robust way using the proposed setup. The contributions concluded that the resulting ranking is inline with common understanding of expressed emotion, and that the pairwise Gaussian process approach performed better or equal to generalized linear models.

- Contribution [I] applies the setup for interactive elicitation of preference used in optimizing an audio reproduction system with many parameter settings. It is concluded that the active learning approach based on Expected Improvement (EI) is significantly more efficient than a random design, and a feasible approach forward.

Based on the individual investigations and findings it can be concluded that the setup provides a feasible and modular approach for elicitation and modelling in many different and real-world applications. Finally, it is concluded that the setup provides a viable foundation for further investigation and applications in for example music and audio applications.

**Predictability**
The thesis finally investigated the predictability of human location as represented by location using mobile phone location as a proxy. This was quantified by bounding the predictability using information theoretic methods as described in the summary report Sec. 4. Contribution [A][B] applied the methods on a dataset obtained by sensing location by WLAN and GSM associations for fourteen users as a proxy for human location location. The contributions presented results on the predictability on many different time scales and for different sensors. The results in [A][B] indicates that we as humans are highly predictable in line with previous results [201], but also on much lower time scale than previously considered.

Appendix A

# Estimating Human Predictability from Mobile Sensor Data

Bjørn Sand Jensen, Jakob Eg Larsen, Kristian Jensen, Jan Larsen and Lars Kai Hansen, Estimating Human Predictability from Mobile Sensor Data, IEEE International Workshop on Machine Learning for Signal Processing 2010.

# ESTIMATING HUMAN PREDICTABILITY FROM MOBILE SENSOR DATA

*Bjørn Sand Jensen, Jakob Eg Larsen, Kristian Jensen, Jan Larsen and Lars Kai Hansen*

Technical University of Denmark,
Department of Informatics and Mathematical Modeling,
2800 Lyngby, Denmark

## ABSTRACT

Quantification of human behavior is of prime interest in many applications ranging from behavioral science to practical applications like GSM resource planning and context-aware services. As proxies for humans, we apply multiple mobile phone sensors all conveying information about human behavior. Using a recent, information theoretic approach it is demonstrated that the trajectories of individual sensors are highly predictable given complete knowledge of the infinite past. We suggest using a new approach to time scale selection which demonstrates that participants have even higher predictability of non-trivial behavior on smaller timer scale than previously considered.

## 1. INTRODUCTION

How predictable are human whereabouts? Very. At least according to a recent study by Song et al. [1][2] based on the predictability of location trajectories. When including the full history of the participants, the upper bound for prediction the next location lie in the 93 % range based on 45.000 participants. Song et al. use an information theoretic approach and by the use of Fano's inequality they show [2] that the entropy rate transforms into an upper bound of the predictability.

The study in [1] is, however, not the first study on human predictability using mobile phones as sensors and the area of mobile phone computing have received considerable attention recently (see e.g. [3]). Of interest to the present study is Eagle et al. [4], who have made extensive studies based on the Reality Mining dataset. Using a Hidden Markov model and the time of day they show prediction accuracies (*home, work, elsewhere*) in the range of 95 %. Using principle component analysis they furthermore analyze the temporal patters extracted as the eigenvectors showing clear temporal patters in daily life. Similar results have been found by Farrahi et al. [5] by analyzing the same dataset using unsupervised methods such as LDA [5] also with the purpose to extract patterns in everyday life.

While Eagle and Farrahi et al. use explicit modeling of the data and extract concrete patters, Song et al. provide a estimation of the information content using non-parametric methods altogether resulting in the upper and lower predictability bounds of the participant behavior. We refer to Bialek et al. [6] for further discussion on information complexity in parametric and non-parametric models.

In this paper we apply a similar information theoretic approach as in [1] to the problem of human predictability. In comparison, we analyze a smaller but more detailed mobile phone based dataset. This means fewer participants, similar time span, but considerably more mobile phone sensors and in particular higher temporal resolution.

Our goal is twofold: The possibility to analyze multiple sensors is quite unique and will provide valuable insight into the predictability of each unique sensor. The combination of sensors will furthermore convey considerably more information about location than e.g. a GSM cell alone [1]. We show that the location and proximity based sensors all have relatively high predictability for the entire population. Whereas Song et al. provide important results on location prediction, the multiple sensors in our scenario can potentially provide a much richer description of context in general. If the extended set of sensors share the predictability of GSM cell location they may contribute additional useful information.

Secondly, we present extended results of predictability on different time scales which indicate that it is possible to predict non-trivial behavior on much smaller scales than the hour based prediction considered in [1]. In addition, we suggest applying mutual information - or prediction information [6] - as a method to easily estimate the optimal scale in a non-parametric fashion.

## 2. MOBILE DATA

We start out describing the experimental setup underlying the so-called *lifelog* dataset obtained and used in the further analysis.

In the experiment standard Nokia N95 8GB mobile phones were used to collect data from embedded sensors including accelerometer, GSM, GPS, Bluetooth, WLAN, and phone activity (calls and messages) as shown in Table 1.

| Sensor | Sampling | Data collected |
|---|---|---|
| Accelerometer | 30/minute | 3D accelerometer values |
| GSM | 1/minute | CellID of base station |
| GPS | 2–3/hour | Coordinates |
| Bluetooth | 20–40/hour | MAC, name, and type |
| WLAN | 1/minute | MAC, SSID, and RX level |
| Phone activity | Event | Phone no. and direction |
| Annotations | Event | Free text |

**Table 1**. Embedded mobile phone sensors and sample rate (non-uniform) included in the experiment.

The mobile phones were running the standard Symbian S60 operating system with standard applications installed, and they were equipped with the Mobile Context Toolbox software [7], running silently in the background not to infer with user behavior.

All sensor recordings were time stamped using the mobile phone embedded timer. The accelerometer values were sampled every 2 seconds. GSM cellular information acquired included the Cell ID with country code, operator code, and location area code. The phone software API in the present system only allows reading the CellID of the GSM base transceiver station to which the phone is currently connected (not the ones visible). GPS was only sampled 2–3 times an hour due to the fact that the sensor is by far the most expensive sensor in terms of energy consumption. Bluetooth scanning was done approximately every 1.5-3 minutes. The sampling rate varies as the Bluetooth discovery time increases with the number of Bluetooth devices available within discovery range. WLAN scanning was done approximately once a minute, recording MAC address, SSID, and RX level of discovered Access Points. Finally phone activity (SMS, MMS, and calls) were recorded whenever it occurred (phone number and direction). In addition, our software application allowed the participant to manually label his/her present location and activity using a text string.

Our experiment included continuous use by 14 participants, each equipped with a mobile phone with the software pre-installed. The participants used the mobile phone as their regular mobile phone for a period of five weeks or more, as they were given instructions to insert their own *SIM card* into the phone. They were furthermore instructed to make and receive calls, send messages, as they would usually do, and generally use the phone as they would use their own phone. Therefore, no particular instructions were given as to carry the phone as we wanted to establish data from regular usage of the mobile phone, and thereby acquire real life data. This means that the participants would not necessarily carry the phone on the body all the time, such as carrying the mobile phone in a pocket. The sampling of sensor on the phone, as shown in Table 1, was based on optimizing the resource consumption, so that the participants should only need to recharge the phone once a day (typically during the night). Yet, this allows for considerably finer time scale than e.g. considered in [1]. Situations where no data is recorded may occur due to the phone running out of battery or actively being switched off. In addition, sensors can fail individually and not return any data, which corresponds to a unknown/no-connection state for that particular sensor.

The experiment started on October 28, 2008 and ended January 7, 2009 and the participants were students and staff from Technical University of Denmark volunteering to be part of the experiment and consented to use the data for research purposes. Thus, the participants were mainly situated in the Copenhagen area, Denmark. The participants took part in the experiment between 31 and 71 days, resulting in approximately 675 days worth of mobile data recordings (approximately 20 mill. data points). The average duration in which data was recorded was 48.2 days. It is worth noticing the number of unique Bluetooth devices, unique GSM cells, and unique WLAN access points discovered during the experiment by all participants: 20408, 2837, and 28110 respectively. In total 9479 readings of GPS position were recorded with great variability among participants, due to the nature of the GPS technology. A total of 6538 calls and messages took place during the experiment. Approximately half of these are outgoing which corresponds to the data available in [1]. The density of our calls is, however, not high enough to reliably localize the participants on e.g. an hourly basis. Instead we use the much more detailed WLAN, GSM, Bluetooth and acceleration data to provide insight into human predictability.

## 3. METHODS

The dataset in Section 2 allows for a large number of different models and analysis methods. In this study we will, as mentioned, apply an information theoretic approach, although before describing the details involved in this we consider the preprocessing required for the final analysis.

A fundamental issue in obtaining discrete times series is the number of quantization levels and sample rate of the true process [6]. The present dataset provides a number of challenges in this regard. The scan cycles are non-uniformly sampled and the scan cycles are of different lengths for each sensor. We therefore derive a commonly aligned time series for each sensor by construction non-overlapping frames of a given window length and assigning the original samples falling within the frame to it. If multiple samples are available for each frame they are merged, which is logical for the indicator type of sensors (WLAN, GSM, BT). We can think of this as combining states in a Markov model, effectively altering the state transitions. A similar idea is used for the acceleration sensor where the feature is calculated

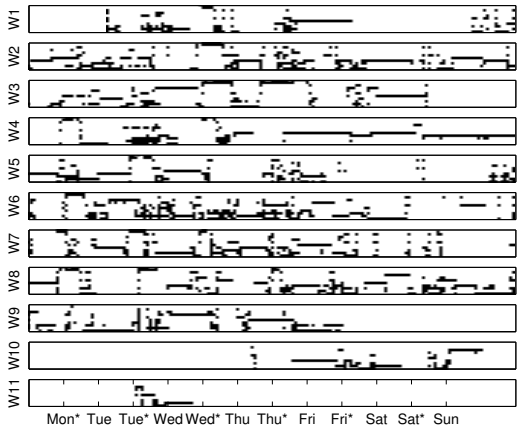



**Fig. 1**. Participant 1's GSM data for 11 weeks, with * denoting noon. The figure shows the top 20 most visited GSM cell towers as rows in each week. The time series is considered in a vector space representation, hence each unique column corresponds to a state.

as the average power within a window represented as a discrete levels, i.e. $X^{ACC} \in \{\text{off}, 1, 2, 3\}$. For example, using the WLAN and integrating all the nets seen into a state effectively means that the predictive variable becomes the WLAN state and not the location as such. If a location is required we could lookup the WLANs positions and generate a location variable from that. The alternative choice of directly working on a location variable generated from the WLANs is not considered here.

The proposed representation is a very detailed description of the participant state. On the contrary the alternative, as used in [1] [2], suggest using only the location (GSM cell) seen the most within a time window as the state. Both methods involve a quite complex quantization of the original data into windows. To evaluate the consequence of this on predictability we consider the change in predictability bounds as the window length is increased from one minute to one hour based on the GSM series.

### 3.1. Information Theoretic Measures

We consider the task of quantifying the predictability possible in our discrete time series, $\mathbf{X} = (X_1, X_2, .., X_n)$, where $n$ is the time index. This is largely inspired by previous work on predictability, complexity and learning (see e.g. [6]) on dynamical systems, and the recent development on a similar dataset as ours [1]. Here predictability quantifies the probability of an algorithm correctly predicting the next state. Thus, given the history the fundamental distribution of interest is $P(X_{n+1}|X_1, X_2, .., X_n)$. In the case where we have no information regarding the past the distribution reduces to $P(X)$. When $P(X)$ is uniform, i.e. each of $N$ states have the same probability, the Shannon entropy (in bits) is defined as $H^{rand}(X) = \log_2 N$.

If, however, the distribution of $X$ is non-uniform, the entropy is given as $H^{unc}(X) = -\sum_{i \in I} p(x_i) \log(p(x_i))$, with $p(x) = P(X = x)$. We use a maximum likelihood plug-in estimate of $p(x)$. This in turn represents the information when no history is available, hence the acronym unc (uncorrelated).

The true entropy of the participant, or the average number of bits needed to encode the symbols in the sequence, can be estimated taking into account the infinite history. Formally this the done by defining the stationary stochastic process $\mathbf{X} = \{X_i\}$ which have the entropy rate given by $H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H\left(X_i|h^{i-1}\right)$, where the history $h^i$ at time step $i$ is $h^i = \{X_1, X_2, ..., X_{i-1}\}$. In our case it is noted that $0 \leq H \leq H^{unc} \leq H^{rand} < \infty$.

An appealing aspect of these non-parametric methods is that we avoid directly limiting model complexity as in any parametric or semi-parametric models. However, a major challenge in using these information measures based on real and unknown processes is the estimation of the entropy rate. A number of ideas have emerged based on compression techniques such as Lempel-Ziv (LZ) (including string matching methods) and Context Weighted Trees (we refer to [8] for an overview). In this study we estimate the entropy rate using the efficient and fast converging LZ based estimator as described in [9][8] and also applied in [2]. The LZ based estimate for a time series of length $n$ is given by $H^{est} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{L_i}{\log_2 n}\right]^{-1}$, where $L_i$ is the shortest substring starting at time step i, which has not been seen in the past. The consistency under stationary assumptions is proved in [9] where it is applied to English text.

Given estimates of the entropy and the entropy rate we consider a related quantity, namely mutual information - or predictive information [6] - defined using the information measures above.

$$I_{pred} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) - H\left(X_i|h^{i-1}\right) \quad (1)$$

$$= H^{unc}(X) - H^{est}(\mathbf{X}) \quad (2)$$

$I_{pred}$ represents the mutual information between the distributions corresponding to knowing and not knowing the past history. That is, it quantifies the fundamental information gain in knowing the (complete) past when prediction the future and we propose it as a way to evaluate optimal quantization and time scale selection. We illustrate the behavior of the measure on a time scale selection problem.

### 3.2. Predictability Bounds

We consider the probability, $\Pi$, that an arbitrary algorithm is able to predict the next state correctly.

Based on the entropy rate and Fano's inequality Song et al. derive a bound so $\Pi \leq \Pi^{\max}(H(X), N)$ with $\Pi^{max}$ given by the relation $H(\Pi^{\max}) = -\Pi^{\max}\log_2(\Pi^{\max}) - (1-\Pi^{\max})\log_2(1-\Pi^{\max}) + (1-\Pi^{\max})\log_2(N-1)$. This non-linear relation between $\Pi^{max}$ and the estimate of $H(X)$ is easily solved using standard methods (see [2] for a full derivation). .

By applying this approach we obtain three upper bounds based on the entropy estimates previously mentioned. The first, $\Pi^{rand}$, provides an upper bound on a random predictor. The second upper bound is $\Pi^{unc}$ which bounds the performance obtainable with a predictor utilizing the state distribution. Finally, the most interesting bound, $\Pi^{max}$, provides a upper limit for the performance of any predictor utilizing the infinite past.

Equally interesting as the upper bound is a lower bound on the predictability, i.e. the worst we can expect from any predictor. Song et. al. shows how a simple lower bound can be constructed based on the so-called regularity. The regularity is in essence a simple predictor based on the most likely situation given the time of day. This is an intuitive measure for some time periods such as daily patters, e.g. utilizing where a person is most likely to be each morning at 6.00. However, if the time scale is in the minute range it does not necessarily make sense to consider the regularity.

We propose instead to use a predictor using the immediate past as the representative of the lower predictability bound. Hence, we select a first order Markov model with the transition probabilities estimated from our finite process.

For generalization purposes we use a resampling scheme in which the entropy and the next state distribution is estimated based on 2/3 of the data and tested on the remaining 1/3. This is performed for nine repetitions in a compromise between accuracy of the generalization estimate and the required samples in order for the entropy estimator to converge. The resampling further allows us to verify that the LZ entropy estimate converges to reasonable similar values for separate temporal sections of the participants life. Any variation over sections will result in a greater variance of the predictability.

### 3.3. Missing data

Missing data constitutes a major issue in applying mobile phones as proxies for human behavior. Our dataset contains two types of missing data: A well-defined off-state and a state where an individual sensor is off/disconnected. The current dataset does not provide any option to differentiate the latter case from not being in range of e.g. any Bluetooth devices, and will consequently be treated as the same state.

We can either predict the true off-state or simply ignore it. Based on initial studies it was found that the convergence of the entropy estimate was more consistent when not including the off-state and instead representing off-states by a single terminal symbol in the LZ based estimation.

An alternative suggested by Song et al. [1] provide an ad-hoc method for extrapolating the entropy corresponding to a missing data fraction, $q$, of zero. They show empirically that the ratio $\ln(H^{est}/H^{unc})$ has a linear relationship across a wide range of missing data fractions. Unfortunately we are not able to obtain such a linear relationship in our data set except for four participants. When extrapolated, these participants show precision within $\leq$10 % for $q = [0; 0.5]$, however, the majority, of participants (ten) deviates with $\gg$10%, which is not found acceptable for our study.

## 4. EXPERIMENTAL RESULTS

We apply the methods described in Section 3 and illustrate the potential predictability by analyzing the GSM, WLAN, BT and acceleration data, described in Section 2. From the results of the 14 participants we estimate the probability density function using a standard kernel density estimator with fixed width for each density across all experiments.

### 4.1. Sensors

Initially we consider the estimated predictability bounds comparing individual sensors at the same time scale, specifically 15 minutes. Fig. 2 shows the predictability bounds of the GSM, WLAN, BT and activity variables. By examining the difference between $\Pi^{\max}$ and $\Pi^{\text{unc}}$ we find that there is considerable gain in knowing the past. From the difference between the Markov model $\Pi^{\text{Markov}}$ and the upper bound, $\Pi^{\text{Max}}$, we see that knowing more than just the previous state seems on average to provide considerable benefit.

In the uncorrelated model participants show considerable differences in the entropy, $H^{unc}$, for all sensors as typically expected in a real population. However, conditioned on the past, the variability is typically lower (except for WLAN), indicating that on average, participants with high entropy have a relative predictable trajectory. This corresponds well with the results observed in [1] for GSM based localization. The Bluetooth data does, as mentioned, contain a considerable fraction of unknown states (not off-states), which will tend to underestimate the true entropy. This means that the bounds are most likely overestimated for Bluetooth.

The GSM, WLAN and Bluetooth sensors are inherently location or proximity oriented sensors, and the estimated distributions all have modes in the high range above 80%, but with noticeable difference in variability. Whereas GSM

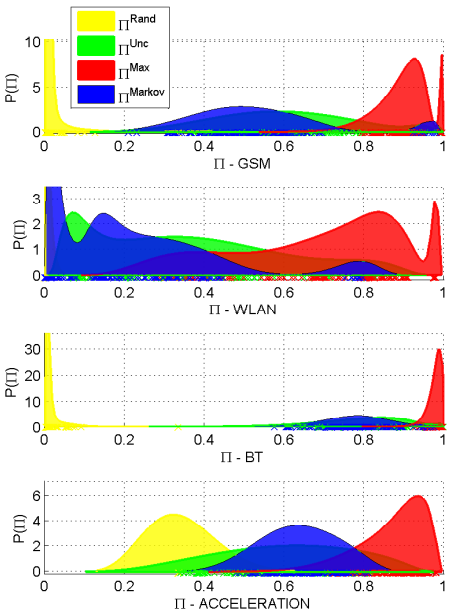



**Fig. 2**. GSM, WLAN and Acceleration. Crosses indicate individual participant estimates.

provides small variability among participants, WLAN seems to provide a much larger difference among participants. This is not surprising since WLAN is considerably more noisy than GSM and captures a more local and detailed state than GSM. Thus, some participants tend to have a relatively low predictability while others are just as predictable as using GSM. The GSM complexity is more similar between mobile phone users. Thus, the WLAN sensor will probably provide the more interesting data, but it is the harder sensor to predict. At least when using the very detailed representation.

In order to do a preliminary analysis of the information shared between the sensors, we consider the averaged normalized mutual information (not conditioned on the past) in Table 2. We generally see that the location/proximity driven sensors show some redundancy as expected, while the activity seems to provide a different element of human behavior.

### 4.2. Time Scale

Fig. 4 shows the effect of varying the window length from one hour to one minutes for the GSM sensor given the state representation described in 3. In doing so, we note that the number of states reduces (see $\Pi^{rand}$). From Fig. 4 we find that the increase in window length towards 3600 seconds tends to remove the correlation effectively rendering the process more random seen by the decrease in estimated predictability. It should be noted that the behavior of these time selection graphs are highly influenced by the representation applied, i.e. the way the different series are constructed (see Section 3).

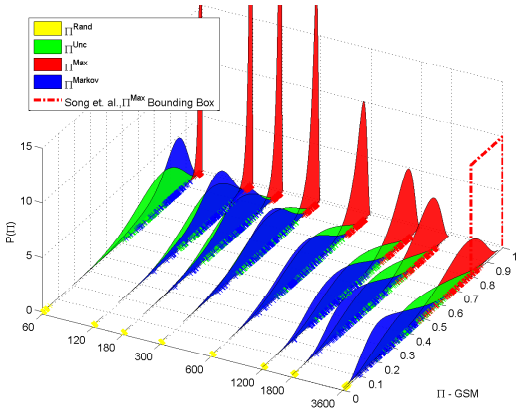



**Fig. 3**. GSM predictability vs. window length in sec. (log scale). The density estimate of $\Pi^{rand}$ is removed for clarity. Participant 3 is furthermore removed for clarity due to his/her outlier nature.

An important point to emphasize, is the fact that the increased predictability is not only an effect of a trivial behavior (i.e. constant). This would imply that the markov model would perform nearly as good as the upper bound suggests. Hence, the increase in predictability can actually be used to predict a non-trivial behavior, and improve for example context-aware applications.

In examining the predictive information Fig. 2 we find an optimal window of approx. 4-5 minutes. Hence, this is the point where we obtain the highest gain in knowing the infinite past when predicting the future. We note that the simpler Markov model actually continues to improve as the window length decreases. This indicates that the process becomes more and more trivial (although not completely) and the gain decreases at the very lowest window length.

| | GSM | WLAN | BT | ACC |
|---|---|---|---|---|
| GSM | 1.00 | 0.45 (0.17) | 0.42 (0.14) | 0.07 (0.02) |
| WLAN | | 1.00 | 0.55 (0.17) | 0.10 (0.03) |
| BT | | | 1.0 | 0.08 (0.05) |
| ACC | | | | 1.00 |

**Table 2**. Average, normalized mutual information and standard deviation across participants. Normalization is performed as $I(X;Y) / \max\{H(X), H(Y)\}$.

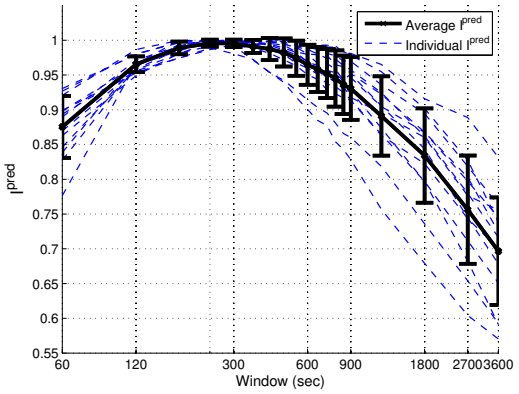



**Fig. 4**. Predictive Information (normalized) vs. window length (log scale). Dashed blue graphs: individual participants. Black graphs: participant mean and std. deviation.

Therefore, we aim for the time scale with some non-trivial, but predictable behavior.

For direct comparison with the results in [1] we refer to the one hour window in Fig. 4. We find that the estimated bound and distribution is in the same range. However, the datasets are inherently different and our results suggest a slightly lower predictability with higher variability between users at this time scale. Overall, though, our results at other times scales supports the conclusion that human trajectories on GSM cell scale are indeed very predictable despite the apparent difference in the number of visited states.

## 5. DISCUSSION

A major issue in the method proposed by Song et al. [2] is the quality (bias/variance) and convergence of the entropy estimator, which is not addressed in [1]. Secondly, the assumed stationarity does not per default apply to all participants, hence quantifying the effect on the entropy estimate is of vital importance. In order to provide reliable results, we have verified a reasonable convergence on the individual subsections, however, the reported upper bound should only be considered a rough estimate. Future work will attempt to improve the apparent issues using the entropy estimate.

An obvious extension of the current analysis is evaluation of sensor combinations. However, computing e.g. the mutual information conditioned on the past is rendered difficult by the present lack of statistics for the LZ based entropy estimator.

## 6. CONCLUSION

In this study we presented a new dataset offering vast possibilities of modeling human behavior. In this initial analysis we adopted an implicit modeling approach based on information theoretic methods to provide bounds for the performance potentially obtainable using explicit modeling.

We presented novel results on the predictability of multiple mobile phone sensors, showing that the findings in [1] generalizes to many more location based sensors. This outlines interesting potential for future context-aware mobile services and applications.

Finally, we showed that the prediction of human mobility is not limited to the one hour time scale previously studied. In particular, we find that predictability seems to further improve at time scales down to about 4-5 minutes.

## 7. REFERENCES

[1] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[2] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási, "Limits of predictability in human mobility - supplementary material," 2010.

[3] Roberta Kwok, "Personal technology: Phoning in data.," *Nature*, vol. 458, no. 7241, pp. 959–961, 2009.

[4] Nathan Eagle and Alex (Sandy) Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[5] Katayoun Farrahi and Daniel Gatica-Perez, "Learning and predicting multimodal daily life patterns from cell phones," in *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 277–280.

[6] William Bialek, Ilya Nemenman, and Naftali Tishby, "Predictability, complexity, and learning," *Neural Comput.*, vol. 13, no. 11, pp. 2409–2463, 2001.

[7] Jakob Eg Larsen and Kristian Jensen, "Mobile context toolbox - an extensible context framework for s60 mobile phones," in *Proceedings of the 4th European Conference on Smart Sensing and Context (EuroSSC)*, 2009.

[8] Yun Gao, Ioannis Kontoyiannis, and Elie Bienenstock, "Estimating the entropy of binary time series: Methodology, some theory and a simulation study," *Entropy*, vol. 10, no. 2, pp. 71–99, 2008.

[9] I. Kontoyiannis, P.H. Algoet, Yu M. Suhov, and A.J. Wyner, "Nonparametric entropy estimation for stationary process and random fields, with applications to english text," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1319–1327, 1998.

Appendix B

# Predictability of Mobile Phone Associations

Bjørn Sand Jensen, Jan Larsen, Kristian Jensen, Jakob Eg Larsen, and Lars Kai Hansen, Predictability of Mobile Phone Associations, European Conference on Machine Learning : Mining Ubiquitous and Social Environments Workshop, 2010.

# PREDICTABILITY OF MOBILE PHONE ASSOCIATIONS

Bjørn Sand Jensen, Jan Larsen, Kristian Jensen,
Jakob Eg Larsen, and Lars Kai Hansen

Technical University of Denmark
{bjje,jl,krije,jel,lkh}@imm.dtu.dk

**Abstract.** Prediction and understanding of human behavior is of high importance in many modern applications and research areas ranging from context-aware services, wireless resource allocation to social sciences. In this study we collect a novel dataset using standard mobile phones and analyze how the predictability of mobile sensors, acting as proxies for humans, change with time scale and sensor type such as GSM and WLAN. Applying recent information theoretic methods, it is demonstrated that an upper bound on predictability is relatively high for all sensors given the complete history (typically above 90%). The relation between time scale and the predictability bound is examined for GSM and WLAN sensors, and both are found to have predictable and non-trivial behavior even on quite short time scales. The analysis provides valuable insight into aspects such as time scale and spatial quantization, state representation, and general behavior. This is of vital interest in the development of context-aware services which rely on forecasting based on mobile phone sensors.

## 1 Introduction

The wide acceptance of sensor rich mobile phones and related applications enables deep studies of human behavior. According to a recent study by Song et al. [11, 12] based on mobile phone location trajectories, individual human mobility patterns are highly predictable. When including the complete history of the participants they derived an upper bound on prediction of the next location of 93% in a large cohort of 45,000 users. The upper bound is based on information theory. Using Fanos inequality it was shown in [12] that the entropy rate transforms into an upper bound of the predictability.

Interest in understanding human behavior using mobile technology is increasing, see e.g., the recent review by Kwok [9]. The work of Eagle et al. [4] on the Reality Mining dataset, marks an early and important contribution. Using a Hidden Markov model and the time of day, they demonstrate explicit prediction accuracies (home, work, elsewhere) in the order of 95%. Furthermore, they use principle component analysis (PCA) to visualize temporal patterns in daily life. The stability of these temporal patterns was confirmed by Farrahi et al. [5] in the same dataset using unsupervised topic models.

While Eagle et al. focus on finding statistical regularities in behaviors at the group level using parametric models Song et al. [11] are interested in individual predictability using non-parametric methods and argue that inter-participant variability is significant and in fact power-law distributed. For a further discussion on parametric and non-parametric models, and the relation to information theory, we refer to Bialek et al. [3].

We follow the implicit modeling approach by Song et al., i.e., using bounds rather than explicit predictors to discuss human behavior in a novel mobile phone data set that complements the analysis of Song et al. Our data set has significantly higher temporal resolution and involves more sensors, however, in a much smaller cohort ($N = 14$).

The opportunity to analyze multiple sensors is quite unique and produces new insight both on the predictability of each sensor and on sensor dependencies. We show that all the location and proximity based sensors have a relatively high predictability bounds for the entire population. Whereas Song et al. [11] provide important results on location prediction, the multiple sensors applied in our study can potentially provide a richer description of context beyond location. And as the extended set of sensors enjoys similar high predictability rates, it may contribute additional useful information on human behavior and support more general context dependent services.

Furthermore, we are interested in the predictability on different time scales, and to probe whether it is possible to predict non-trivial behaviors on smaller scales than the one hour time scale considered in [11]. Finally, we suggest applying mutual information - or prediction information [3] - as a method to easily estimate an optimal time scale in a non-parametric fashion. The information theoretic approach is based on upper and lower bounds on predictability. In this paper we use the upper bound proposed by Song et al., and derive and analyze a tighter lower bound based on a first Markov model [7], rather than the zero order model of [12].

An early version of this work was presented to the machine learning community in [7]. The present paper extends this work with; 1) an extended description of the experimental setup; 2) comparison between WLAN and GSM, in particular in relation to optimal time scale; 3) extended discussion and interpretation.

The paper first gives a presentation of the experimental setup and the acquired mobile phone dataset. In Section 3 we present the information theoretic tools necessary to follow the analysis of the dataset. The result of the study is presented and discussed in Section 4, followed by the conclusion in Section 5.

## 2 Experimental Setup

Within the last decade there has been a number of studies of real-world dataset or *lifelogs* reflecting human life [1, 2]. In this section we present a software platform for obtaining such *lifelogs* using standard off-the-shelf mobile phones functioning as individual wearable sensor platforms.

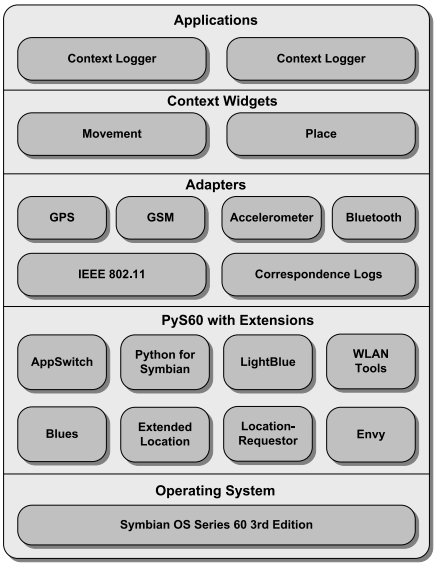



**Fig. 1.** Mobile Context Toolbox architecture [10]. The bottom two layers provide low-level access to the embedded sensors, whereas the adapters and context widget layers provide high-level Python interfaces to sensors and inferred information for applications.

### 2.1 Mobile Context Toolbox

Since utilizing multiple sensor inputs on mobile devices can be a complex task we have used our Mobile Context Toolbox [10], which provides an open extensible framework for the Nokia S60 platform running the Symbian mobile Operating System present in mobiles phones such as Nokia N95. The framework provides access to multiple embedded mobile phone sensors, including accelerometer, microphone, camera, etc., as well as networking components such as phone application data (calendar, address book, phone log, etc.), and phone state (profile, charge level, etc.).

In principle, mobile devices can acquire information from the surrounding environment as well as from online sources, but in the present study we focus on information that can be acquired through the large variety of sensors embedded in the device. The framework has multiple layers (as depicted in Fig. 1) on top of the Nokia S60 platform. The framework uses Python for S60 (PyS60) with a set of extensions for accessing low-level sensors and application data. The adapters layer provides interfaces for the low-level sensors, whereas the context widgets uses one or more adapters to infer higher-level contextual information. Finally,

the application layer utilize contextual information inferred from context widgets and/or directly from context adapters. Further details and explanation of the Mobile Context Toolbox is provided in [10].

### 2.2 Logging Data From Embedded Sensors

For the purpose of using mobile phones as an instrument for gathering *lifelog* information we have built a *Context Logger* application, which subscribes to all sensors through the relevant system components and then continuously records all events received from the multiple adapters and widgets. In effect all accessible sensor data is recorded, as shown in Table 1.

All sensor recordings are time stamped using the embedded mobile phone timer. The accelerometer was sampled every 2 seconds. Samples are read-out every 30s, in batches of 15 samples. The purpose of reading in bursts is to enable reading short bursts with higher sampling rate. GSM cellular information acquired included the Cell ID with country code, operator code, and location area code. In the present system the phone software API only allow reading the CellID of the GSM base transceiver station to which the phone is currently connected (not the ones visible). Since the GPS sensor is the most energy expensive sensor, it was only sampled 2–3 times an hour to reduce the energy consumption. Bluetooth scans was performed approximately every 1.5-3 minutes. The sampling rate varies as the Bluetooth discovery time increases with the number of Bluetooth devices available within discovery range. A discovery of a Bluetooth device will always produce the unique MAC address of the device, however, the lookup of the Bluetooth "friendly name" and device type might fail as more time is required to obtain this information. WLAN scanning is performed approximately once a minute, recording MAC address, SSID, and RX level (power ratio in decibels – a measurement of the link quality) of discovered Access Points. Finally, phone activity (SMS, MMS, and calls) were recorded whenever it occurred (phone number and direction).

In addition to the above mentioned sensor data, the *Context Logger* application allows a user to manually label his/her present location and activity. The label is a text string which can be entered by the user on the mobile phone, such as, *home*, *running*, and *having dinner*. Entered labels will be stored and subsequently shown on a list to pick from in order to avoid re-typing location and activity labels. Users can manually choose to label location and activity by

| Sensor | Sampling | Data |
|---|---|---|
| Accelerometer | 30/minute | 3D Accelerometer values |
| GSM | 1/minute | CellID of GSM base transceiver station |
| GPS | 2–3/hour | Longitude, Latitude, and Altitude |
| Bluetooth | 20–40/hour | Bluetooth MAC, friendly name, and device type |
| WLAN | 1/minute | Access Point MAC address, SSID, and RX level |
| Phone activity | Event | Phone number and direction of call or message |

**Table 1.** List of embedded mobile phone sensors used for collecting data

selecting a menu item in the *Context Logger* application, however, this mechanism is further enhanced with the ability to automatically prompt for labels. Thus, a user can choose to enter a label at any point in time, but the application will also prompt the user to label a location and activity 2–3 times a day in order to receive feedback. The data location and activity data recorded on the mobile phone is a key-value pair along with the time stamp. There are no pre-defined location and activity labels defined in the application and the labeling is completely free form with the users determining, how they want to write their text labels.

Situations with missing data may occur due to the phone running out of battery, being switched off, or simply not able to acquire data through one or more sensors (for instance no GSM reception).

### 2.3 Data Collection

An initial deployment of the system included continuous use by 14 participants, each equipped with a standard mobile phone (Nokia N95) which had the Mobile Context Toolbox pre-installed along with the *Context Logger* application that would continuously record the data acquired from all sensors currently supported by our framework. The participants were using the mobile phone as their regular mobile phone for a period of five weeks or more, as they were given instructions to insert their own simcard into the phone. Furthermore, they were instructed to make and receive calls, send messages, etc., as they would usually do, and generally use the phone as they would use their own phone. Therefore, no particular instructions were given, since we wanted to establish data from regular usage of the mobile phone, and thereby acquire real-life data. This means that the participants would not necessarily carry the phone on the body all the time such as carrying the mobile phone in a pocket.

As the survey is completely dependent on the cooperation of the participants and due to increased use of sensors, the lack of battery time was considered a risk in terms of participants leaving the survey. Thus, the sensor configuration of sampling on the phone was based on optimizing the resource consumption, so that the participants should only need to recharge the phone once a day (typically during the night).

The experiment started on October 28, 2008 and ended January 7, 2009 and the participants were students and staff members from The Technical University of Denmark volunteering to be part of the experiment. Thus, the participants were mainly situated in the greater Copenhagen area, Denmark. The 14 participants took part in the experiment between 31 to 71 days, resulting in approximately 472 days of data covering data collection periods totalling 676 days. The average duration was 48.2 days. An overview of the collected data is provided in Table 2.

During the experiment a total of approximately 20 million data points were collected with the accelerometer contributing the most with 14.5 million data points. A summary of recordings from Bluetooth scans, WLAN scans, GPS readings, and GSM readings can be seen in Table 2. It is worth noticing the

| Part. | Accel | BT. | BT.* | GPS | GSM | GSM* | Ann. | PA. | WLAN | WLAN* | Days |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1474480 | 54349 | 2846 | 516 | 69458 | 529 | 533 | 544 | 224101 | 6387 | 71 |
| 2 | 2045773 | 38028 | 2478 | 1514 | 75669 | 603 | 596 | 1062 | 364272 | 6040 | 66 |
| 3 | 318597 | 27329 | 790 | 12 | 37217 | 98 | 222 | 21 | 125600 | 630 | 31 |
| 4 | 875287 | 7880 | 743 | 2 | 17750 | 228 | 134 | 620 | 186421 | 2394 | 52 |
| 5 | 1117147 | 13575 | 2373 | 4058 | 56206 | 227 | 386 | 277 | 251016 | 2347 | 48 |
| 6 | 711490 | 23702 | 1141 | 95 | 51702 | 235 | 82 | 839 | 92396 | 2119 | 50 |
| 7 | 1184457 | 13327 | 1765 | 3 | 45826 | 955 | 272 | 581 | 139466 | 4017 | 46 |
| 8 | 700258 | 42346 | 2080 | 614 | 74250 | 172 | 212 | 74 | 154108 | 3359 | 41 |
| 9 | 1101926 | 42346 | 1050 | 119 | 37393 | 100 | 104 | 497 | 104576 | 1804 | 38 |
| 10 | 1103086 | 21676 | 2104 | 419 | 63937 | 419 | 414 | 116 | 192338 | 2650 | 48 |
| 11 | 1122315 | 12492 | 655 | 929 | 46158 | 929 | 163 | 121 | 295716 | 2286 | 47 |
| 12 | 796452 | 30610 | 2317 | 40 | 51548 | 40 | 143 | 151 | 97769 | 2403 | 50 |
| 13 | 1024276 | 27550 | 1741 | 1114 | 49349 | 1114 | 137 | 949 | 171951 | 5463 | 51 |
| 14 | 971558 | 21502 | 1303 | 44 | 40017 | 44 | 149 | 686 | 118687 | 1263 | 36 |
| Total | 14547102 | 350879 | 20408 | 9479 | 716480 | 2837 | 3547 | 6538 | 2518417 | 28110 | 48.2 |

**Table 2.** Overview of collected data for each participant in the experiment: Participant, Accelerometer, Bluetooth, Unique Bluetooth devices, GPS, GSM, Unique GSM cells, Annotations, Phone Activity, WLAN Access Points, Unique WLAN Access Points, Duration in days.

number of unique Bluetooth devices, unique GSM cells, and unique WLAN access points discovered accumulatively during the experiment by all participants: 20408, 2837, and 28110, respectively. In total 9479 readings of GPS position were recorded, but the recordings varies a lot among the participants due to the nature of the GPS technology. As a GPS position typically can not be obtained indoor only a subset of users have sufficient recordings of GPS position. For instance, if they typically place the mobile phone near a window when indoor where a GPS position can be obtained. A total of 6538 calls and messages took place during the experiment.

The participants provided 3547 annotations of locations and activities in total. On average the participants provided 253 annotations during their participation, with an overall average of 5.3 labels provided per user per day. The most active participants provided 8-9 labels per day on average, whereas the least active participants provided 2 labels per day on average.

## 3 Methods

In this study we will apply an information theoretic approach in the analysis of the dataset obtained and described in Section 2. Although before describing the details involved in this, we consider the preprocessing required for the final analysis.

A general issue in obtaining discrete times series is the number of quantization levels and sample rate of the true process [3]. The scan cycles used to obtain the

present dataset are non-uniformly sampled and the scan cycles are of different length for each sensor. We therefore construct a commonly aligned time series for each sensor by creating non-overlapping frames of a given window length. The original samples falling within the frame is then assigned to it. If multiple samples falls within a frame they are merged, which is reasonable for the indicator type of sensors (WLAN, GSM, BT). This is similar to combining states in a Markov model effectively altering the state transitions. In case of the acceleration sensor, the feature is calculated as the average power within a window represented as a discrete levels[1], i.e., $X^{ACC} \in \{\text{off}, 1, 2, 3\}$. Considering the WLAN sensor and integrating all the networks seen into a state effectively means that the predictive variable becomes the WLAN state and not the location as such. If a specific location is needed a lookup to a database could return the position of the WLAN access point and generate a location variable from that. An alternative would be to directly work on a location variable generated from the WLAN access point, but this is not considered here.

The proposed representation constitutes a very detailed description of the participant state. An alternative suggested in [11, 12], represents the state as the most visited GSM cell location within a time window. Both approaches involve a relatively complex temporal quantization and resampling of the original data. To evaluate the consequence temporal scale, we consider the change in predictability bounds as the window length is decreased from one hour to one minute.

### 3.1 Information Theoretic Measures

We consider the problem of quantifying the predictability obtainable in a discrete process, $\mathbf{X} = (X_1, X_2, .., X_i)$, where $i$ is the time index and $X$ is the state variable. This is to a large degree motivated by previous work on predictability, complexity and learning (see e.g. [3]), and recent development on a similar dataset [11]. Here predictability is defined as the probability of an arbitrary algorithm correctly predicting the next state. Hence, given the history the basic distribution of interest is $P(X_{i+1}|X_1, X_2, .., X_i)$. In the case where we have no information regarding the history, the distribution naturally reduces to $P(X)$. When $P(X)$ is uniform, i.e., each of $M$ states have the same probability of occurring, the Shannon entropy (in bits) is defined as $H^{rand}(X) = \log_2 M$.

In the case where the distribution of $X$ is non-uniform, the entropy is given as

$$H^{unc}(X) = -\sum\nolimits_{i \in I} p(x_i) \log(p(x_i)) \tag{1}$$

with $p(x) = P(X = x)$. This in turn represents the information when no history is available, hence, the acronym unc (uncorrelated).

The entropy rate of the participants trajectory, or the average number of bits needed to encode the states in the sequence, can be estimated taking into

[1] An equiprobable quantization is used, i.e., each level has the same frequency of occurrence within the entire dataset.

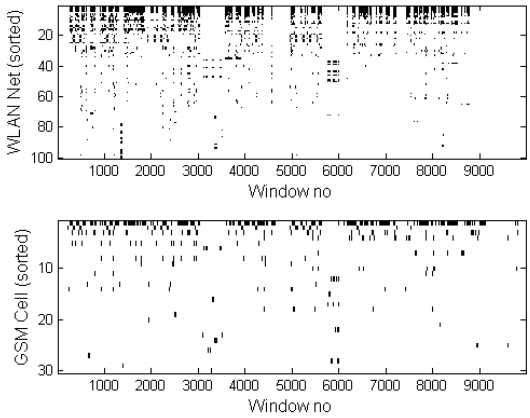



**Fig. 2.** Participant 2. Time series for WLAN (top 100) and GSM data (top 30). The time series are considered in a vector space representation, hence, each column is a state vector.

account the complete history. This is done by defining the stationary stochastic process $\mathbf{X} = \{X_i\}$ which have the entropy rate defined as

$$H(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} \sum\nolimits_{i=1}^{n} H\left(X_i | h^{i-1}\right), \tag{2}$$

where the history $h^i$ at time step $i$ is $h^i = \{X_1, X_2, ..., X_{i-1}\}$. It is noted that $0 \leq H \leq H^{unc} \leq H^{rand} < \infty$.

A challenge in using these information measures based on real and unknown processes is the estimation of the entropy rate. A number of ideas have emerged based on compression techniques such as Lempel-Ziv (LZ) (including string matching methods) and Context Weighted Trees for binary processes. For a general overview, we refer to [6]. An appealing aspect of these non-parametric methods is that we avoid directly limiting model complexity as would be necessary if we applied parametric or semi-parametric models. In this study we estimate the entropy rate using a LZ based estimator as described in [8, 6] and also applied in [12]. The entropy rate estimate for a time series of length $n$ is given by

$$H^{est} = \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} \frac{L_i}{\log_2 n}\right]^{-1} \tag{3}$$

where $L_i$ is the shortest substring starting at time step $i$, which has not been seen in the past. The consistency under stationary assumptions is proved in [8] where the method is applied to the analysis of English text.

Given estimates of the entropy and the entropy rate we consider a related quantity, namely mutual information - or predictive information [3]. This measure is already available given the information measures above as

$$I_{pred} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) - H\left(X_i|h^{i-1}\right) \tag{4}$$

$$= H^{unc}(X) - H^{est}(\mathbf{X}) \tag{5}$$

In effect $I_{pred}$ represents the mutual information between the distributions corresponding to knowing and not knowing the past history. Hence, it quantifies the fundamental information gain in knowing the (complete) past when prediction the future, and we propose it as a easy way to evaluate quantization and time scale effects. We illustrate the behavior of the measure on a time scale selection problem in Section 4.

### 3.2 Predictability

In order to construct bounds on the predictability we consider the probability, $\Pi$, that an arbitrary algorithm is able to predict the next state correctly.

Based on the entropy rate and Fano's inequality Song et al. derives a bound so $\Pi \leq \Pi^{\max}(H(X), M)$ with $\Pi^{max}$ given by the relation [12]

$$H(X) = H(\Pi^{\max}) + (1 - \Pi^{\max}) \log_2 (M - 1) \tag{6}$$

with the function, $H(\Pi^{\max})$, given by

$$H(\Pi^{\max}) = -\Pi^{\max} \log_2 (\Pi^{\max}) - (1 - \Pi^{\max}) \log_2 (1 - \Pi^{\max}) \tag{7}$$

This non-linear relation between $\Pi^{max}$ and the estimate of $H(X)$ is easily solved using standard methods (for a full derivation see [12]).

Adopting this approach we obtain three upper bounds based on the entropy estimates previously mentioned. The first, $\Pi^{rand}$, provides an upper bound on a random predictor. The second upper bound is $\Pi^{unc}$ which bounds the performance obtainable with a predictor utilizing the observed state distribution. Finally, the most interesting bound, $\Pi^{max}$, provides a upper limit for the performance of any algorithm utilizing the complete past.

The upper bound is of course interesting in understanding the potential predictability, although we find a lower bound equally important in the analysis. Song et. al. [11] show how a simple lower bound can be constructed based on the so-called regularity. The regularity is in essence a zero order Markov model based on the most likely state at any given time of the day, i.e., using only the time of occurrence and no sequence information. This is an intuitive measure for some time periods such as daily patterns, e.g., utilizing where a person is most likely to be each morning at 7.00. However, if the time scale is in the minute range it does not necessarily make sense to consider the regularity.

Instead we propose to use a predictor using the immediate past as the representative of the lower predictability bound. For this purpose we use a first order

Markov model with the transition probabilities estimated from the finite process. Thus, the next state prediction is based on the distribution $P(X_{i+1}|X_1, X_1, ..., X_i) = P(X_{i+1}|X_i)$.

To avoid overfitting which tends to render the bounds overly optimistic, we use a resampling scheme in which the entropy and the next state distribution is estimated based on 2/3 of the data and tested on the remaining 1/3. This is performed for nine distinct subsections in a compromise between accuracy of the estimate and the needed samples for the entropy estimator to converge. The resampling further allows us to verify that the LZ entropy estimate converges to reasonable similar values for separate temporal sections of the participants life. Any variation across subsections will result in a greater variance of the estimated bound.

## 4 Results

In order to provide insight into the predictability of mobile phones sensors and thereby indirectly insight into human behavior, we apply the information theoretic methods described in Section 2. The density estimates of the bounds are all made using a standard kernel based density estimator (the bandwidth is hand-tuned for visualization).

### 4.1 Individual Sensors

One of the goals in this study is to analyse the potential predictability of different sensors and to that end we provide the predictability bounds for the four prominent sensors in the dataset, specifically GSM, WLAN, Bluetooth and acceleration at 15 min. window length. Fig. 3 shows the predictability bounds for the GSM, WLAN, Bluetooth and acceleration/activity sensors. By examining the difference between $\Pi^{\max}$ and $\Pi^{\text{unc}}$ we find that there is considerable gain in knowing the past. From the difference between the Markov model $\Pi^{\text{Markov}}$ and the upper bound, $\Pi^{\text{Max}}$, we see that knowing more than just the previous state seems to provide considerable benefit.

In the uncorrelated case, $H^{unc}$, participants show considerable differences in the entropy for all sensors as typically expected in a real population. However, conditioned on the past, the variability is typically lower (except for WLAN) indicating that participants with high entropy have a relative predictable trajectory. This corresponds well with the results observed in [11] for GSM based localization. The Bluetooth data does, as mentioned, contain a considerable fraction of unknown states (not off-states), which will tend to underestimate the true entropy. This means that the bounds are most likely overestimated for Bluetooth.

The GSM, WLAN and Bluetooth sensors are inherently location or proximity oriented sensors, and the estimated distributions all have modes in the high range above 80%, but with noticeable difference in variability. Whereas GSM provides small variability among participants, WLAN seems to provide a much larger difference among participants. This is not surprising since WLAN is considerably

more noisy than GSM and captures a more local and detailed state than GSM. Thus, some participants tend to have a relatively low predictability while others are just as predictable as using GSM. The GSM complexity is on the other hand more similar between mobile phone users. In effect WLAN is most likely the more interesting sensor, but harder to predict, at least using the very detailed representation.

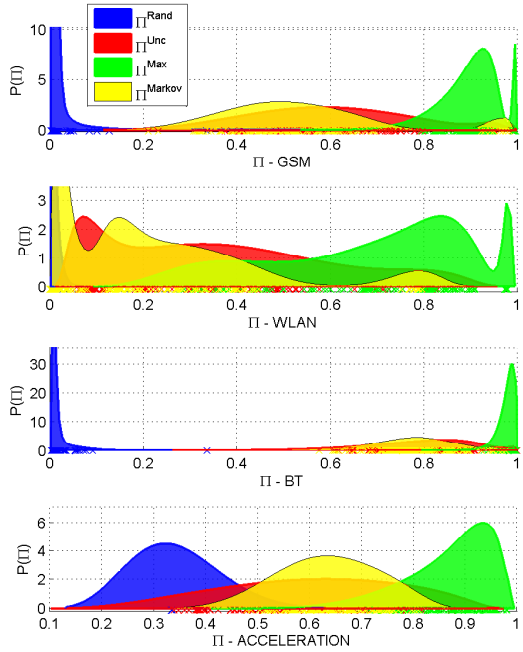



**Fig. 3.** Detailed representation: Predictability of GSM, WLAN, Bluetooth and Acceleration sensors. Crosses indicate individual participant estimates. The mode at 0.99 and 0.98 in the GSM and WLAN densities are due to participant 3 which is left out in the further analysis for clarity.

### 4.2 Time Scale

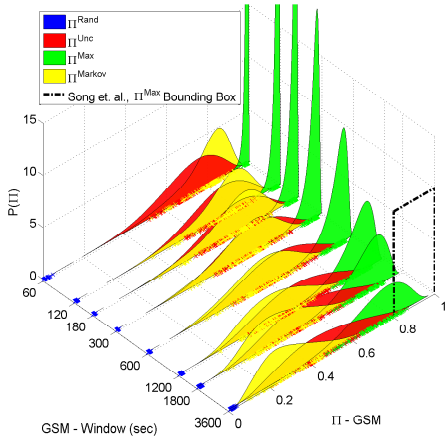
A primary goal in this study is the analysis of the time scales involved in the prediction of location based sensors with the aim to provide support for context-based services and general understanding of human mobility. The focus in this part is thus focused on the GMS and WLAN sensors and the predictability on a wide range of window lengths - and the examination of the optimal scale suggested by the predictive information.

Figure 4 shows how the predictability bounds changes with the window length. Noticeable are the GSM results in Figure 4(a) which are directly comparable with the original results in [11]. The bounding box indicates that the predictability is in the same range, although smaller in our case, possibly due to the more detailed representation utilized here. However, we obtain a similar upper bound at approximately 10 min. scale. This trend towards a high upper bound continues as the scale progresses downwards to 60 seconds.

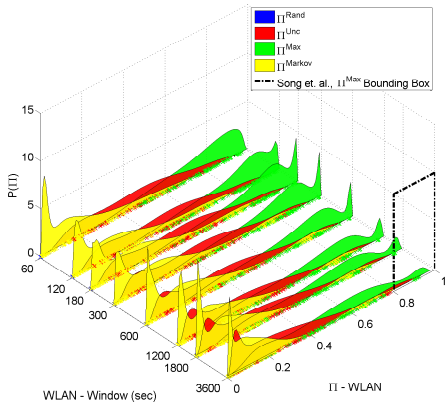
In general there are various fundamental ways why this might happen. First of all, we may simply oversample a constant process leaving the resulting times series highly trivial to predict. The second reason is that the fundamental dependencies are removed when aggregating the cell at long scales and the shorter time scale provides the best representation. A simple way to examine the first options is to look at the predictability suggested by the first order Markov model and determine how far it is from reaching the upper bound. We notice that the despite $\Pi^{Markov}$ approaching $\Pi^{Max}$, there seems to be some non-trivial behavior which is not predicted by the first order model. Not surprisingly this indicates that the first order Markov model is too simple. However, the bound provides a very convenient indication of what a more complex model is able to obtain.

Whereas GSM provides a rather rough indication of mobility and specific location, WLAN cells have quite high location resolution. Examining the time scale for WLAN reveals that the complexity of the problem is very high compared to the GMS case as seen on the pure results obtained by the Markov model. Despite this, we notice that the upper bound is still quite high. This suggests that there is an large unexploited potential in applying a more complex model than for example a first order Markov model. As with the GSM sensor we find that the shorter time scales provides a higher upper bound, and noticeably that the variability among the participants are lower, in effect offering better generalization of the predictability across multiple users.

As we have hinted, the optimal scale time scale for predictability for both GSM and WLAN at small time scales, and to examine the precise scale at which the past offers the most information in predicting the future we consider the predictive information. This is depicted in Figure 5 showing how the predictive information depends on the time scale. We find that the optimal scale is in the 3-4 minute range for both types of sensors. This is our main result and supplements the results in Song et al. [11] who focused on longer time scales (60 minutes). The high predictability at short time scales is of great interest for applications and is "good news" for pro-active services based on predicting human needs and behavior. Furthermore, the fact that the two distinct sensors operating on



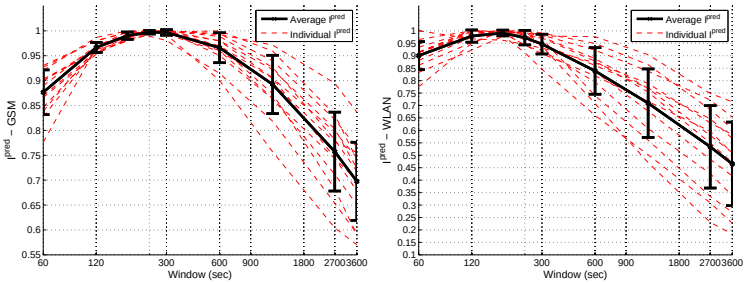

(a) GSM




(b) WLAN

**Fig. 4.** Predictability vs. window length in sec. (log scale). Notice that participant 3 has been removed from the density estimate due to his/her outlier nature as noticed in Figure 3

different spatial resolution yet still suggest the same optimal temporal scale, indicates that there exists fundamental information at this scale where both the upper bound on GSM and WLAN predictability are quite high. Yet, the exact information available at this scale and implications of this is to be examined in




(a) GSM (b) WLAN

**Fig. 5.** Predictive Information (normalized) vs. window length (log scale). Participant 3 is left out.

a future analysis, for example using explicit modeling paradigms such as (multiway) factor analysis and advanced dynamical models.

## 5 Conclusion

In this paper we described an experimental setup for obtaining so-called *lifelog* data using embedded mobile phones. The resulting dataset offers many possibilities for investigating interesting elements of human behavior.

In the analysis we adopted an implicit modeling approach based on recent information theoretic methods to provide bounds for the prediction one could hope to obtain using explicit modeling. We presented results on the predictability of multiple mobile phone sensors showing that the basic findings in [11] generalizes to more location and proximity based sensors. Specifically, that the gain in knowing the past is significant, which indicates interesting potential for context-aware mobile applications relying on forecasting for example GSM and WLAN associations.

Finally, we showed that the prediction of human mobility generalizes to shorter time scales than the one hour time scale previously studied in [11]. In particular, we showed that the collected GSM and WLAN have the same optimal time scales for prediction, specifically 3-4 minute range. Despite this encouraging result, the exact interpretation and relevance of the patterns at the suggested scale needs further investigation and analysis, for example using explicit modeling.

## References

1. MyLifeBits (accessed May 1st 2010). `http://research.microsoft.com/en-us/projects/mylifebits/default.aspx`

2. Bell, G., Gemmell, J.: A digital life. Scientific American 296(3), 5865 (March 2007)
3. Bialek, W., Nemenman, I., Tishby, N.: Predictability, complexity, and learning. Neural Comput. 13(11), 2409–2463 (2001)
4. Eagle, N., (Sandy) Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
5. Farrahi, K., Gatica-Perez, D.: Discovering human routines from cell phone data with topic models. 2008 12th IEEE International Symposium on Wearable Computers pp. 29–32 (2008)
6. Gao, Y., Kontoyiannis, I., Bienenstock, E.: Estimating the entropy of binary time series: Methodology, some theory and a simulation study. Entropy 10(2), 71–99 (2008)
7. Jensen, B.S., Larsen, J.E., Jensen, K., Larsen, J., Hansen, L.K.: Estimating human predictability from mobile sensor data. In: IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010) (2010)
8. Kontoyiannis, I., Algoet, P., Suhov, Y.M., Wyner, A.: Nonparametric entropy estimation for stationary process and random fields, with applications to english text. IEEE Transactions on Information Theory 44(3), 1319–1327 (1998)
9. Kwok, R.: Personal technology: Phoning in data. Nature 458(7241), 959–961 (2009)
10. Larsen, J.E., Jensen, K.: Mobile context toolbox - an extensible context framework for s60 mobile phones. In: Proceedings of the 4th European Conference on Smart Sensing and Context (EuroSSC) (2009)
11. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. Science 327(5968), 1018–1021 (2010)
12. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility - supplementary material (2010)

Appendix C

# Effecient Preference Learning with Pairwise Countinious Observation and Gaussian Processes

---

Bjørn Sand Jensen, J.B. Nielsen, and Jan Larsen. Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes. IEEE International Workshop on Machine Learning for Signal Processing, 2011.

# EFFICIENT PREFERENCE LEARNING WITH PAIRWISE CONTINUOUS OBSERVATIONS AND GAUSSIAN PROCESSES

*Bjørn Sand Jensen*[1], *Jens Brehm Nielsen*[1,2] *& Jan Larsen*[1]

[1]Technical University of Denmark,
Department of Informatics and
Mathematical Modeling,
Richard Petersens Plads B321
2800 Lyngby, Denmark

[2]Widex A/S,
Nymøllevej 6
3540 Lynge, Denmark

## ABSTRACT

Human preferences can effectively be elicited using pairwise comparisons and in this paper current state-of-the-art based on binary decisions is extended by a new paradigm which allows subjects to convey their degree of preference as a continuous but bounded response. For this purpose, a novel Beta-type likelihood is proposed and applied in a Bayesian regression framework using Gaussian Process priors. Posterior estimation and inference is performed using a Laplace approximation.

The potential of the paradigm is demonstrated and discussed in terms of learning rates and robustness by evaluating the predictive performance under various noise conditions on a synthetic dataset. It is demonstrated that the learning rate of the novel paradigm is not only faster under ideal conditions, where continuous responses are naturally more informative than binary decisions, but also under adverse conditions where it seemingly preserves the robustness of the binary paradigm, suggesting that the new paradigm is robust to human inconsistency.

***Index Terms*—** Pairwise Comparisons, Continuous Response, Gaussian Processes, Laplace Approximation

## 1. INTRODUCTION

Traditionally, various aspects of human perception and cognition are assumed to be related to absolute psychological magnitudes or intensities. This includes the classical findings by Weber, Fechner and Stevens who, for example, investigated the perception of light intensity. However, recently Lockhead [1] has argued that every aspect of perception is relative, even those apparently absolute aspects investigated by Weber, Fechner and Stevens. In accordance with the theory in [1], we investigate human perception from a relative viewpoint and examine one such highly relative aspect, namely preference.

Formal treatment of relative aspects goes back to the ideas of Thurnstone [2] and the principle of comparative judgments. In the present context it was revisited by Chu *et al.* [3] who formulated a Bayesian approach to preference learning using Gaussian Process (GP) priors. This formulation has initiated a number of related studies and applications, such as audiological preference [4], multi-subject food preference [5] and an extension for semi-supervised, active learning settings [6].

In this work we extend the likelihood model in [3] to support observations which in effect measure *the perceived degree to which one option is preferred over another*. This degree of preference can be obtained from a traditional paired comparison test, which implies that a subject is asked to give a subjective assessment of the degree to whether $A$ or $B$ is preferred over the other. Specifically, we model the observed degrees of preferences through a likelihood conditioned on a functional value difference and support inconsistent observations by applying a re-parameterized Beta distribution.

In a traditional setting, users would not be trusted to be able to quantify such an abstract and difficult aspect as degree of preference. Instead, we would rely on massive repetitions of a standard binary experiment to estimate the proportion of $A \succ B$ using this as an expression of the degree of any preferences. However, we want to exploit the extra information from continuous responses to get a faster method for preference elicitation without jeopardizing the robustness from standard binary responses. The hypothesis is that we are able to learn faster by (indirectly) observing the perceived probability of $A \succ B$ as opposed to a binary decision. Applying appropriate priors and noise modeling should ensure this to be true also under adverse conditions.

In order to examine this hypothesis, we apply the novel likelihood in a flexible Bayesian setup similar to [3] in which the prior on the underlying preference function is defined by a GP with a potentially complex covariance structure. The Laplace approximation is used for inference and model selection by *maximum-a-posteriori* (MAP) estimates. This provides a consistent probabilistic framework for making predictions and evaluating the predictive uncertainty. We use simulations with different synthetic noise scenarios in order

to compare a standard binary decision with the novel model. The performance of both methods is evaluated using the predictive performance.

## 2. MODELS FOR PAIRWISE OBSERVATIONS

In the previous section, we motivated pairwise comparisons from a cognitive perspective, yet pairwise comparisons can be considered more broadly. It is usually possible to describe any aspect of a pairwise comparison, such as preference, real difference, or perceived similarity in terms of a latent function [2].

In the following we will model the preference of two distinct inputs, $u \in \mathcal{X}$ and $v \in \mathcal{X}$, in terms of the difference between two functional values, $f(u)$ and $f(v)$. This implies a function, $f : \mathcal{X} \to \mathbb{R}$, which defines an internal, but latent absolute preference.

The general setup is as follows: We consider $n$ distinct inputs $x_i \in \mathcal{X}$ denoted $\mathcal{X} = \{x_i | i = 1, ..., n\}$, and a set of $m$ responses on pairwise comparisons between any two inputs in $\mathcal{X}$, denoted by

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

where $y_k \in \mathbb{Y}$. $u_k \in \mathcal{X}$ and $v_k \in \mathcal{X}$ are option one and two in the $k$'th pairwise comparison, respectively. The main topic of this paper is how the domain of the response variable influences the learning rate of the latent function $f$ in relation to the number of paired comparisons. As previously indicated, we will consider two cases:

- **binary** where $y_k = d_k, d_k \in \{-1, 1\}$
- **continuous and bounded** where $y_k = \pi_k, \pi_k \in ]0, 1[$.

In both cases we consider $y$ a stochastic variable, informally implying the definition of the conditional density given by $p(y_k | f_k(u_k), f(v_k))$, denoted by $p(y_k|\mathbf{f}_k)$ with $\mathbf{f}_k = [f(u_k), f(v_k)]^\top$.

### 2.1. Binary Response

When restricting the response variable to be a discrete, two-alternatives, forced choice, paired-comparison between the two presented options, we define the response variable as $d_k \in \{-1, 1\}$. A preference for either $u_k$ or $v_k$ is indicated by $-1$ or $+1$, respectively.

When considering noise on the forced decisions the resulting random variable can be modeled by a classic choice model such as the Logit or Probit [7, chapter 6]. In the current setting we restrict ourself to the Probit model mainly for analytical reasons.

Given a function, $f$, we can define the likelihood of observing a discrete choice $d_k$ directly as the conditional density.

$$p(d_k|\mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L}) = \Phi\left(d_k \frac{f(v_k) - f(u_k)}{\sqrt{2}\sigma}\right), \tag{1}$$

where $\Phi(x)$ is the cumulative Gaussian (with zero mean and unity variance) and $\boldsymbol{\theta}_\mathcal{L} = \{\sigma\}$. This classic Probit likelihood is by no means a new invention and can be dated back to Thurstone and his fundamental definition of *The Law of Comparative Judgment*[2]. However, it was first considered with GPs in [3] and later in e.g. [5] and [6].

### 2.2. Continuous Response

The primary contribution of this paper is a novel response model allowing for more subtle judgments, where the response variable describes the degree to which the prevailing option is preferred.

For this purpose we formally define a continuous but bounded response $\pi \in ]0; 1[$ observed when comparing $u$ and $v$. The first option, $u$, is preferred for $\pi < 0.5$. The second option, $v$, is preferred for $\pi > 0.5$ and none is preferred for $\pi = 0.5$. Hence, the response captures both the choice between $u$ and $v$, and the degree of the preference.

Instead of using the Probit function directly as the choice model, it is used as a link function mapping from functional differences to continues bounded responses. More precisely, the Probit is used as a mean function for a Beta type distribution with parameterized shape parameters $\alpha$ and $\beta$, thus

$$p(\pi_k|\mathbf{f}_k) = \text{Beta}(\pi_k | \alpha(\mathbf{f}_k), \beta(\mathbf{f}_k)).$$

To express the shape parameters of the Beta distribution as a function of the Probit mean function $\mu(\mathbf{f}_k)$, we apply a well-known re-parametrization of the Beta distribution [8].

$$\alpha(\mathbf{f}_k) = \nu\mu(\mathbf{f}_k), \qquad \beta(\mathbf{f}_k) = \nu(1 - \mu(\mathbf{f}_k)), \tag{2}$$

where $\nu$ relates to the precision of the Beta distribution and is not parameterized by $f$. Finally, our novel likelihood depicted in Fig. 1 is described by

$$p(\pi_k|\mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L}) = \text{Beta}(\pi_k | \nu\mu(\mathbf{f}_k, \sigma), \nu(1 - \mu(\mathbf{f}_k, \sigma))), \tag{3}$$

where $\boldsymbol{\theta}_\mathcal{L} = \{\sigma, \nu\}$ and $\mu(\mathbf{f}_k, \sigma)$ is given by

$$\mu(\mathbf{f}_k, \sigma) = \Phi\left(\frac{f(v_k) - f(u_k)}{\sqrt{2}\sigma}\right).$$

The precision term $\nu$ in Eq. (2) and Eq. (3) is inversely related to the observation noise on the continuous bounded responses. In general, $\nu$ can be viewed as a measure of how consistent the scale is used in a given comparison.

### 2.3. Gaussian Process Priors

At this point we have not specified any form, order or shape of $f$, but referred to $f$ as an abstract function. We maintain the abstraction by considering a non-parametric approach and use a Gaussian process (GP) to formulate our beliefs about $f$.

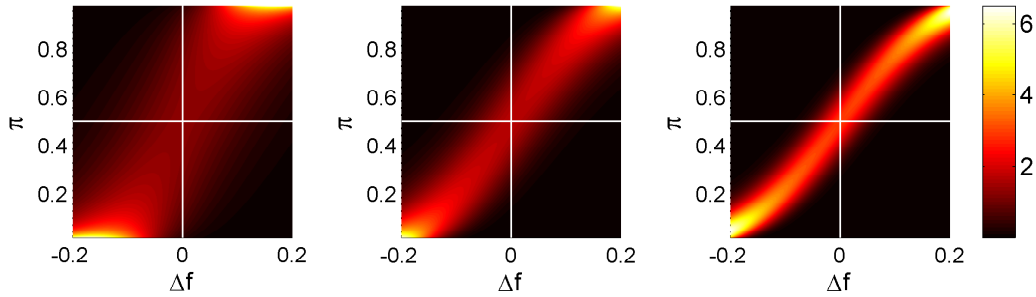



**Fig. 1**. Illustration of the proposed likelihood with $p(\pi_k|\mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L})$ shown as a color level. The likelihood parameters $\boldsymbol{\theta}_\mathcal{L}$ are $\sigma = 0.1$ and left: $\nu = 3$, middle: $\nu = 10$ and right: $\nu = 30$

A GP is typically defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [9]. Following [9] we denote a function drawn from a GP as $f(x) \sim \mathcal{GP}\left(\mathbf{0}, k(\cdot,\cdot)_{\boldsymbol{\theta}_c}\right)$ with a zero mean function, and $k(\cdot,\cdot)_{\boldsymbol{\theta}_c}$ referring to the covariance function with hyper-parameters $\boldsymbol{\theta}_c$, which defines the covariance between the random variables as a function of the inputs $\mathcal{X}$. The fundamental consequence of this formulation is that the GP can be considered a distribution over functions, i.e., $p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c)$, with hyper-parameters $\boldsymbol{\theta}_c$ and $\mathbf{f} = [f(x_1), f(x_2), ..., f(x_n)]^T$, i.e., dependent on $\mathcal{X}$.

In a Bayesian setting we can directly place the GP as a prior on the function defining the likelihood. This leads us directly to a formulation given Bayes relation with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\mathcal{L}, \boldsymbol{\theta}_c\}$

$$p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) = \frac{p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_\mathcal{L})\, p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c)}{p(\mathcal{Y}|\boldsymbol{\theta}, \mathcal{X})}. \tag{4}$$

The prior $p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c)$ is given by the GP and the likelihood $p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_\mathcal{L})$ is either of the two likelihoods defined previously, with the assumption that the likelihood factorizes as usual, i.e., $p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_\mathcal{L}) = \prod_{k=1:m} p(y_k|f(u_k), f(v_k), \boldsymbol{\theta}_\mathcal{L})$

The posterior of interest, $p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$, is directly defined when equipped with the likelihood and the prior, but it is unfortunately not of any known analytical form in either the binary nor the continuous case.

## 3. INFERENCE & PREDICTIONS

Since the likelihoods considered in this paper do not result in closed form solutions to the posterior in Eq. (4), we must resort to approximations, such as the Laplace approximation, Expectation Propagation or sampling. Since the main focus of this work is to examine the general properties of the likelihood proposed in Sec. 2.2, we use the well-know and relatively simple Laplace approximation. The required steps have previously been derived for the binary likelihood [3] (see [10] for a detailed derivation), and in the following it will be derived for the proposed likelihood from Sec. 2.2.

### 3.1. Laplace Approximation

The main idea is to approximate the posterior by a single Gaussian distribution, such that $p(\mathbf{f}|\mathcal{Y}) \approx \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \mathbf{A}^{-1})$. Where $\hat{\mathbf{f}}$ is the mode of the posterior and $\mathbf{A}$ is the Hessian of the negative log-likelihood at the mode. The mode is found as $\hat{\mathbf{f}} = \arg\max_\mathbf{f} p(\mathbf{f}|\mathcal{Y}) = \arg\max_\mathbf{f} p(\mathcal{Y}|\mathbf{f})\, p(\mathbf{f})$.

The general solution to the problem can be found by considering the unnormalized log-posterior and the resulting cost function which is to be maximized, is given by

$$\begin{aligned}\psi(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) = &\log p(\mathcal{Y}|\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_\mathcal{L}) - \frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} \\ &- \frac{1}{2}\log|\mathbf{K}| - \frac{N}{2}\log 2\pi.\end{aligned} \tag{5}$$

where $\mathbf{K}_{i,j} = k(x_i, x_j)_{\boldsymbol{\theta}_c}$. We use a damped Newton method with soft linesearch to maximize Eq. (5). In our case the basic damped Newton step (with adaptive damping factor $\lambda$) can be calculated without inversion of the Hessian (see [10])

$$\begin{aligned}\mathbf{f}^{new} = &\left(K^{-1} + \mathbf{W} - \lambda\mathbf{I}\right)^{-1} \\ &\cdot\left[(\mathbf{W} - \lambda\mathbf{I}) - \mathbf{f} + \nabla\log p(\mathcal{Y}|\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_\mathcal{L})\right],\end{aligned} \tag{6}$$

Using the notation $\nabla\nabla_{i,j} = \frac{\partial^2}{\partial f(x_i)\partial f(x_j)}$ we apply the definition $\mathbf{W}_{i,j} = -\sum_k \nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L})$. We note that the term $\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L})$ is only nonzero when both $x_i$ and $x_j$ occur as either $v_k$ or $u_k$ in $\mathbf{f}_k$. In contrast to standard binary GP classification the Hessian $\mathbf{W}$ is not diagonal, which makes the approximation slightly more involved.

When converged, the resulting approximation is

$$p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) \approx \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, \left(\mathbf{W} + \mathbf{K}^{-1}\right)^{-1}\right). \tag{7}$$

In the Beta case the required two first derivatives of the like-

lihood are given by:

$$\nabla_i \log p(\pi_k \mid \mathbf{f}_k, \boldsymbol{\theta}_{\mathcal{L}}) = \mathbb{I}(x_i) \cdot \nu \cdot \mathcal{N}(\mathbf{f}_k) \cdot [\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha) + \psi(\beta)] \text{ and} \quad (8)$$

$$\nabla\nabla_{i,j} \log p(\pi_k \mid \mathbf{f}_k, \boldsymbol{\theta}_{\mathcal{L}}) = -\mathbb{I}(x_i)\mathbb{I}(x_j) \cdot \nu^2 \cdot \mathcal{N}(\mathbf{f}_k), \cdot \left[\mathcal{N}(\mathbf{f}_k) \cdot \left(\psi^{(1)}(\alpha) + \psi^{(1)}(\beta)\right) + \frac{f(v_k) - f(u_k)}{2\nu\sigma^2} \cdot (\log(\pi_k) - \log(1-\pi_k) - \psi(\alpha) + \psi(\beta))\right], \quad (9)$$

where we for convenience write $\alpha$ and $\beta$ without the dependency on $\mathbf{f}_k$ Eq. (2). $\psi(z)$ and $\psi^{(1)}(z)$ are the digamma function of zero'th and first order, respectively, $\mathcal{N}(\mathbf{f}_k) = \mathcal{N}\left(\frac{f(v_k)-f(u_k)}{\sqrt{2}\sigma}\middle| 0, 1\right)$ and $\mathbb{I}(z)$ is an indicator function defined by

$$\mathbb{I}(z) = \begin{cases} 1 & \text{if } z = u_k \\ -1 & \text{if } z = v_k \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

We refer to [10] for a full derivation and for the required derivatives for the binary case as first described in [3].

### 3.2. Hyper-parameter Estimation

So far we have simply considered the hyper-parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathcal{L}}, \boldsymbol{\theta}_c\}$ variables on which we can condition the primary posterior, and not worried about their values or distributions. In the following, we consider the hyper-parameters random variables on which we place a prior and the full posterior would be $p(\mathbf{f}, \boldsymbol{\theta}|\mathcal{Y})$. However, since the focus in this work is $p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$ we only use the prior on $\boldsymbol{\theta}$ to make point estimates of the hyper-parameters in terms of *maximum-a-posteriori* (MAP) estimates.

We obtain the MAP estimates by iterating between the Laplace approximation with fixed hyper-parameters, i.e. finding $p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}^{\text{MAP}})$, followed by a maximization step in which $\boldsymbol{\theta}^{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X}, \mathbf{f})$.

We first consider the standard evidence approach which seeks to optimize the marginal likelihood given by

$$p(\mathcal{Y}|\boldsymbol{\theta}, \mathcal{X}) = \int p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c) d\mathbf{f} = p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X}) p(\mathcal{Y}|\mathcal{X}) / p(\boldsymbol{\theta}|\mathcal{X}). \quad (11)$$

Our interest is in the posterior term, $p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X})$, so considering Eq. (11) in terms of the log-posterior of $\boldsymbol{\theta}$ we obtain $\log p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X}) = \log p(\boldsymbol{\theta}|\mathcal{X}) + \log p(\mathcal{Y}|\boldsymbol{\theta}, \mathcal{X}) - \log p(\mathcal{Y}|\mathcal{X})$, where $p(\boldsymbol{\theta}|\mathcal{X})$ is the prior and typical considered independent of $\mathcal{X}$. The evidence term, $\log p(\mathcal{Y}|\boldsymbol{\theta}, \mathcal{X})$, is analytical intractable in both likelihood cases, but we can approximate it using the existing Laplace approximation to obtain [10] $\log p(\mathcal{Y}|\boldsymbol{\theta}) \approx \log p(\mathcal{Y}|\hat{\mathbf{f}}, \boldsymbol{\theta}_{\mathcal{L}}) - \frac{1}{2}\hat{\mathbf{f}}^T \mathbf{K}^{-1} \hat{\mathbf{f}} - \frac{1}{2} \log |I + \mathbf{K}W|$. Now $\boldsymbol{\theta}^{\text{MAP}}$ is found by maximizing $\log p(\boldsymbol{\theta}|\mathcal{Y}, \mathcal{X})$ with respect to $\boldsymbol{\theta}$ and noting that $p(\mathcal{Y}|\mathcal{X})$ is independent of $\boldsymbol{\theta}$. We perform the optimization using a BFGS gradient method. The required derivatives and details are provided in [10].

The choice of particular priors is left for the simulations in Sec. 4, however, if $p(\boldsymbol{\theta})$ is the Uniform distribution, we obtain the traditional evidence optimization [9] as expected. It is noted that the complexity of the posterior inference is of the same order as standard GP regression described in [9].

### 3.3. Prediction

The main task is to estimate the latent function, $f$, with the end goal to do predictions of the observable variable $y$ for a pair of test inputs $r \in \mathcal{X}_t$ and $s \in \mathcal{X}_t$. In this paper, we are especially interested in the discrete decision, i.e., whether $r \succ s$ or $s \succ r$. This can be obtained from both likelihood models, thus allowing for direct comparison of the two formulations in terms of predictive performance.

We first consider the predictive distribution of $f$ which is required in both cases, and for notational convenience we omit the conditioning on $\mathcal{X}$ and $\mathcal{X}_t$. Given the GP, we can write the joint prior distribution between $\mathbf{f} \sim p(\mathbf{f}|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}})$ and the test variables $\mathbf{f}_t = [f(r), f(s)]^T$ as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_t \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_t \\ \mathbf{k}_t^T & \mathbf{K}_t \end{bmatrix}\right), \quad (12)$$

where $\mathbf{k}_t$ is a matrix with elements $\mathbf{k}_{2,i} = k(s, x_i)_{\boldsymbol{\theta}_c^{\text{MAP}}}$ and $\mathbf{k}_{1,i} = k(r, x_i)_{\boldsymbol{\theta}_c^{\text{MAP}}}$ with $x_i$ being a training input. The conditional $p(\mathbf{f}_t|\mathbf{f})$ is obviously Gaussian as well and can be obtained directly from Eq. (12). The predictive distribution is given as $p(\mathbf{f}_t|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}}) = \int p(\mathbf{f}_t|\mathbf{f}) p(\mathbf{f}|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}}) d\mathbf{f}$. With the posterior approximated with the Gaussian from the Laplace approximation then $p(\mathbf{f}_t|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}})$ will be Gaussian too and is given as $\mathcal{N}(\mathbf{f}_t|\mu^*, \mathbf{K}^*)$ with $\mu^* = [\mu_r^*, \mu_s^*]^T = \mathbf{k}_t \mathbf{K}^{-1} \hat{\mathbf{f}}$ and

$$\mathbf{K}^* = \begin{bmatrix} \mathbf{K}_{rr}^* & \mathbf{K}_{rs}^* \\ \mathbf{K}_{sr}^* & \mathbf{K}_{ss}^* \end{bmatrix} = \mathbf{K}_t - \mathbf{k}_t^T (\mathbf{I} + \mathbf{W}\mathbf{K}) \mathbf{k}_t,$$

where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from Eq. (7). With the predictive distribution for $\mathbf{f}_t$, the final prediction of the observed variable is available from

$$p(y_t|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}}) = \int p(y_t|\mathbf{f}_t, \boldsymbol{\theta}_{\mathcal{L}}^{\text{MAP}}) p(\mathbf{f}_t|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}}) d\mathbf{f}_t \quad (13)$$

If the likelihood is an odd function, as in both our cases, the binary preference decision between $r$ and $s$ can be made directly from $p(\mathbf{f}_t|\mathcal{Y})$. In contrast, evaluation of the integral in Eq. (13) is required for, e.g., soft decisions, reject options and sequential designs.

#### *3.3.1. Binary Likelihood*

If $p(\mathbf{f}_t|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}})$ is Gaussian and we consider the Probit likelihood, the integral in Eq. (13) can be evaluated in closed form as a modified Probit function given by [3]

$$P(r \succ s|\mathcal{Y}) = \Phi\left((\mu_r^* - \mu_s^*)/\sigma^*\right) \quad (14)$$

with $(\sigma^*)^2 = 2\sigma^2 + \mathbf{K}_{rr}^* + \mathbf{K}_{ss}^* - \mathbf{K}_{rs}^* - \mathbf{K}_{sr}^*$

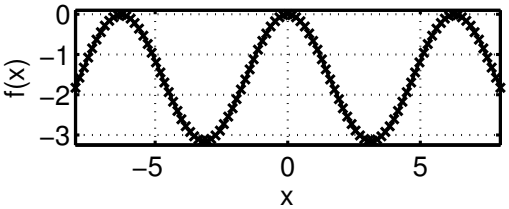



**Fig. 2**. The Griewangk function used to evaluate the predictive performance. Crosses indicate discrete samples. The center peak is slightly higher than the two others.

| Simulation | Data Noise | $\boldsymbol{\theta}_{\mathcal{L}}$ | | $\boldsymbol{\theta}_c$ | |
|---|---|---|---|---|---|
| | $\nu_D$ | $\sigma$ | $\nu$ | $\sigma_f$ | $l$ |
| BR NoiseFree | No Noise | $\delta_1$ | | $\delta_{\text{ideal}}$ | $\delta_{\text{ideal}}$ |
| BR | $\{3, 10, 30\}$ | $\delta_1$ | | $\mathcal{U}_1$ | $\mathcal{U}_1$ |
| CBR NoiseFree | No Noise | $\delta_1$ | $\delta_{\to\infty}$ | $\delta_{\text{ideal}}$ | $\delta_{\text{ideal}}$ |
| CBR Ideal | $\{3, 10, 30\}$ | $\delta_1$ | $\delta_{\{3,10,30\}}$ | $\delta_{\text{ideal}}$ | $\delta_{\text{ideal}}$ |
| CBR | $\{3, 10, 30\}$ | $\delta_1$ | $\mathcal{G}(1,\eta)_{\{3,10,30\}}$ | $\mathcal{U}_1$ | $\mathcal{U}_1$ |

**Table 1**. Simulation conditions. $\delta_x$ is a point-mass, thus the parameter is constantly equal to $x$. The $\delta_{\text{ideal}}$ value is learned as $m \to \infty$. $\mathcal{U}_x$ is an uniform prior over $]0; \infty[$ with the parameter initialized to $x$. $\mathcal{G}(1,\eta)_x$ is a Gamma prior with inverse scale parameter $\eta = 0.05$ and initialization $x$.

### *3.3.2. Continuous Likelihood*

In the continuous case the observed variable, $\pi$, does not directly define the discrete observation which is the main focus of this work. However, a binary preference can be derived from the continuous likelihood via the predictive distribution over $\pi$. With the suggested likelihood and mean function in Sec. 2.2 the probability of the binary choice is obtained as $P(r \succ s|\mathcal{Y}, \boldsymbol{\theta}_{\mathcal{L}}) = \int_{\pi=0}^{\pi=1/2} p(\pi_t|\mathcal{Y}, \boldsymbol{\theta}_{\mathcal{L}})\, d\pi_t$, thus

$$P(r \succ s|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}}) = \int p(\mathbf{f}_t|\mathcal{Y}, \boldsymbol{\theta}^{\text{MAP}}) \text{Betacdf}\left(\frac{1}{2}\middle|\alpha(\mathbf{f}_t), \beta(\mathbf{f}_t)\right) d\mathbf{f}_t \quad (15)$$

In the ideal case of a noise-free user, i.e., $\nu \to \infty$, the Beta distribution reduces to a point mass at the mean defined by the Probit function. Hence, in the limit of a completely consistent user, the predictions from Eq. (15) reduces to a classical choice model with predictions that follows Eq. (14).

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

To study the performance of the models in a controlled setting, we use a synthetic dataset generated from the deterministic *Griewangk function* depicted in Fig. 2. We use the predictive performance of the binary decision to compare the learning rates of the binary response (BR) model as the baseline and the continuous bounded response (CBR) model. In each comparison, the two inputs are drawn randomly among 101 input points sampled uniformly from $x = [-8; 8]$.

The training points $\pi_k$ are drawn from a Beta distribution with the parameterization from Sec. 2.2 with the Probit link function in Eq. (4), $\sigma = 1$, and the Griewangk function values as the two inputs. The noise level on the training data is defined by the parameter $\nu_D$ corresponding to $\nu$ in the CBR model. The binary decision $d_k$ is determined by whether $\pi_k$ is smaller or larger than 0.5. For evaluation, we generate an independent binary test set located equidistantly in between the training points. Initial experiments showed that in order to get a robust predictive model for all noise level, it is important to learn the $\nu$ parameter in the CBR model. The initial experiments also indicated that it is vital not to underestimate the noise, while an overestimation is not as crucial and provides overall good predictive performance. This suggests a prior with a monotonic increasing likelihood towards the highest noise level. A natural choice is a Gamma(1,$\eta$) prior with inverse scale parameter $\eta$.

The considered models, priors and parameters are listed in Table 1 where the covariance parameters, $\theta_c$, are applied in a GP prior with a covariance function defined by the *squared exponential* kernel $k_{SE}(x, x') = \sigma_f^2 \exp(-l^{-2}\|x - x'\|^2)$. When a specific prior is not a point-mass/constant indicated by $\delta_x$ in Table 1, the hyper-parameters are estimated (MAP) either for each training set size (realistic scenario) or for $m = 500$ (ideal scenario). The latter is indicated by $\delta_{\text{ideal}}$.

The learning curves from Fig. 3 show that under ideal conditions with nearly noise-free observations and a correct noise setting (Fig. 3, right plot) the CBR model outperforms the BR models as expected, since a continuous response will essentially provide more information from each experiment under ideal conditions than a binary response will. Also, in both high and moderate noise conditions (Fig. 3, left and middle plot) the CBR model with a correct noise setting (CBR Ideal) outperforms the corresponding BR model significantly in terms of learning rates and actually shows similar learning rates as the BR model under noise-free conditions. Finally and most importantly, the learning rates are only slightly lower when $\nu$ has been inferred from data via the MAP procedure (with different initializations) than when it is specified correctly, which suggests that the parameter inference framework with independent priors is robust in real-life-scenarios without ideal model and noise conditions.

We have focused on a controlled example to highlight properties of the model and inference, leaving a real-world validation for future work. Future work also includes the extension of the mean function, Eq. (4), using a mixture of Probit functions to account for different user behavior such as centering and contraction bias. For a real-world setting, a natural extension is a suitable active learning criteria, such as the *expected value of information* framework applied recently in e.g. [5] for the BR model.

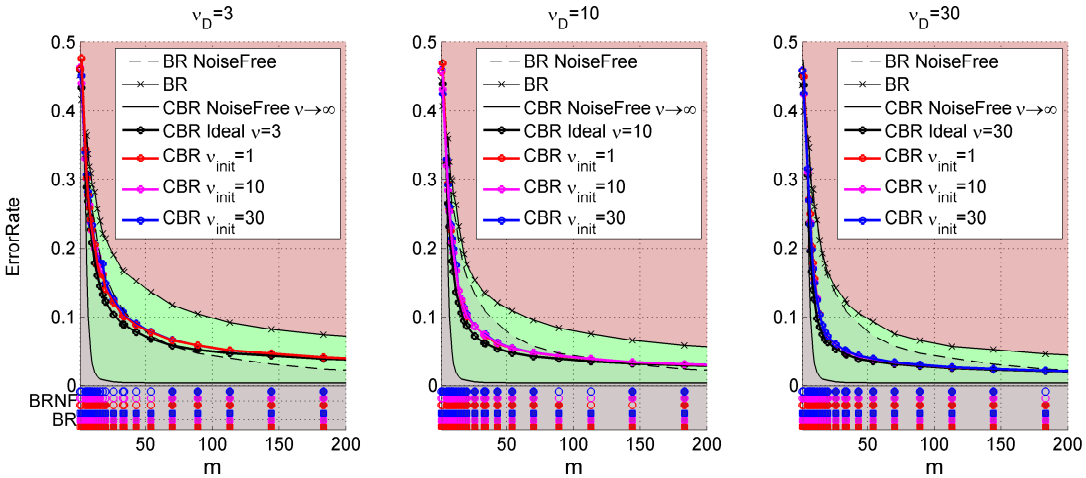



**Fig. 3**. Mean error test rates (MER) as a function of the number of experiments over 100 different realizations of the training set generated with different $\nu_D$. In the red and top green area MER are worse and better, respectively, than those obtained with the BR model on the noisy data. In the lower green area MER are also better than those obtained by the BR NoiseFree, and finally, the grey area corresponds to unrealistic MER better than those obtained with a CBR NoiseFree model with $\nu \to \infty$ evaluated with $\nu = 10^3$ on a noise-free data set. The six rows of markers indicate if the MER of the corresponding CBR model are significantly different from those resulting from the BR (squares) and from the BR NoiseFree (circles). If solid, the zero-hypothesis of the two means being equal is rejected at the 5% level using a paired t-test.

## 5. CONCLUSION AND PERSPECTIVES

We have proposed a new model for preference learning with Gaussian Process priors with the main purpose to increase the learning rate compared to the standard binary model applied in [3]. We have outlined a robust and flexible inference framework for the new model based on suitable priors and the Laplace approximation. Simulations were used to present properties and performance, which showed a significant information increase from each experiment under ideal conditions as expected but more importantly also under adverse conditions. The performance is especially increased in a certain window of opportunity.

Acknowledgement: This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

## 6. REFERENCES

[1] G. R. Lockhead, "Absolute Judgments Are Relative: A Reinterpretation of Some Psychophysical Ideas.," *Review of General Psychology*, vol. 8, no. 4, pp. 265–272, 2004.

[2] L. L. Thurstone, "A law of comparative judgement.," *Psychological Review*, vol. 34, 1927.

[3] W. Chu and Z. Ghahramani, "Preference learning with Gaussian Processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.

[4] P. Groot, T. Heskes, T. Dijkstra, and J. Kates, "Predicting preference judgments of individual normal and hearing-impaired listeners with Gaussian Processes," *IEEE Transactions on Audio, Sound, and Language Processing*, 2010.

[5] E. Bonilla, S. Guo, and S. Sanner, "Gaussian Process preference elicitation," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, Eds., pp. 262–270. 2010.

[6] W. Chu and Z. Ghahramani, "Extensions of Gaussian Processes for ranking: semi-supervised and active learning," in *Workshop Learning to Rank at Advances in Neural Information Processing Systems 18*, 2005.

[7] R. D. Bock and J. V. Jones, "The measurement and prediction of judgment and choice.," 1968.

[8] S. Ferrari and F. Cribari-Neto, "Beta Regression for Modelling Rates and Proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, Aug. 2004.

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

[10] Bjørn Sand Jensen and Jens Brehm Nielsen, *Pairwise Judgements and Absolute Ratings with Gaussian Process Priors*, Technical Report, DTU Informatics, September 2011.

Appendix D

# A Predictive Model of Music Preference using Pairwise Comparisons

---

Bjørn Sand Jensen, Javier Saez Gallego and Jan Larsen. A Predictive model of music preference using pairwise comparisons. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2012. Published.

# A PREDICTIVE MODEL OF MUSIC PREFERENCE USING PAIRWISE COMPARISONS

*Bjørn Sand Jensen, Javier Saez Gallego & Jan Larsen*[*]

Technical University of Denmark, Department of Informatics and
Mathematical Modeling, Richard Petersens Plads B321
2800 Lyngby, Denmark

## ABSTRACT

Music recommendation is an important aspect of many streaming services and multi-media systems, however, it is typically based on so-called collaborative filtering methods. In this paper we consider the recommendation task from a personal viewpoint and examine to which degree music preference can be elicited and predicted using simple and robust queries such as pairwise comparisons. We propose to model - and in turn predict - the pairwise music preference using a very flexible model based on Gaussian Process priors for which we describe the required inference. We further propose a specific covariance function and evaluate the predictive performance on a novel dataset. In a recommendation style setting we obtain a leave-one-out accuracy of 74% compared to 50% with random predictions, showing potential for further refinement and evaluation.

***Index Terms*—** Music Preference, Kernel Methods, Gaussian Process Priors, Recommendation

## 1. INTRODUCTION

Methods for music recommendation has received a great deal of attention the last decade with most approaches typically being classified as collaborative filtering (*top-down*) or content-based (*bottom-up)* methods, with hybrid methods (see e.g. [1]) comprising both. Such hybrid systems exploits both ratings and contents to make recommendations, but the focus is still on the recommendation itself and not the basic questions of preference. Although from a fundamental point of view it is also interesting how well human preference can be elicited and represented without relying on the help of others. This also includes the aim to answer basic questions such as which properties of music determines the persons music preference. Obviously the potential power of collaborative filtering should not be discarded, but exploited in a principled manner in order to answer basic questions and hopefully provide an even better predictive model of individual music preference.

Based on these observations we consider music preference in a personalized setting by applying a Gaussian Process regression model which takes into account both human ratings and audio features. In contrast to many audio rating systems it is not based on absolute ratings of a single track, but on a pairwise comparisons between tracks, which is typically considered robust and have a low cognitive load (see e.g. [2]).

We furthermore propose to use a covariance function motivated from a generative view of audio features with a potential multi-task part which lead to similar capabilities as standard collaborative filtering, but with the added information level provided by subject features. Posterior inference in the resulting non-parametric Bayesian regression model is performed using a Laplace approximation of the otherwise intractable distribution. Any hyperparameters in the model can be learned using an empirical Bayes approach.

We evaluate the resulting model by its predictive power on a small scale, public available dataset [3] where 10 subjects evaluate 30 tracks in 3 genres. We report and discuss a number of aspects of the performance such as the learning curves as a function of the number of pairwise comparisons and learning curves when leaving out a track as test set.

## 2. METHODS

In this work we focus on modeling preference elicited by pairwise queries, i.e., given two inputs tracks $u$ and $v$ we obtain a response, $y \in \{-1, 1\}$, where $y = -1$ corresponds to a preference for $u$, and $+1$ corresponds to a preference for $v$. We consider $n$ distinct input tracks $x_i \in \mathcal{X}$ denoted $\mathcal{X} = \{x_i | i = 1, ..., n\}$, and a set of $m$ responses on pairwise comparisons between any two inputs in $\mathcal{X}$, denoted by

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

where $y_k \in \{-1, 1\}$. $u_k \in \mathcal{X}$ and $v_k \in \mathcal{X}$ are option one and two in the $k$'th pairwise comparison.

We consider $y_k$ as a stochastic variable and we can then formulate the likelihood of observing a given response as cu-

[*]This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views. [**] 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.

mulative normal distribution.

$$p\left(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}}\right)=\Phi\left(y_k\frac{f\left(v_k\right)-f\left(u_k\right)}{\sqrt{2}\sigma}\right), \tag{1}$$

with $\mathbf{f}_k=[f\left(u_k\right),f\left(v_k\right)]$, $\Phi(x)$ defines a cumulative Gaussian (with zero mean and unity variance) and $\boldsymbol{\theta}_{\mathcal{L}}=\{\sigma\}$. This is in turn the well known Probit classification model, where the argument is the difference between two latent variables (functional values) and not just a single latent variable. This in effect implies that the $f(\cdot)$ encodes an internal, but latent preference function which can be elicited by pairwise comparisons via the likelihood model in Eq.( 1). This idea was already considered by [4], but recently suggested in a Gaussian Process context by [5].

### 2.1. Gaussian Process Prior

The real question remains, namely how $f$ is modelled. We will follow the principle suggested by [5] in which $f$ is considered an abstract function and we can in turn place a prior distribution over it. A natural prior is a Gaussian Process (GP) defined as "a collection of random variables, any finite number of which have a (consistent) joint Gaussian distribution" [6]. Following [6] we denote a function drawn from a GP as $f\left(x\right)\sim\mathcal{GP}\left(\mathbf{0},k(\cdot,\cdot)_{\boldsymbol{\theta}_c}\right)$ with a zero mean function, and $k(\cdot,\cdot)_{\boldsymbol{\theta}_c}$ referring to the covariance function with hyper-parameters $\boldsymbol{\theta}_c$, which defines the covariance between the random variables as a function of the inputs $\mathcal{X}$. The consequence of this formulation is that the GP can be considered a distribution over functions, i.e., $p\left(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c\right)$, with hyper-parameters $\boldsymbol{\theta}_c$ and $\mathbf{f}=[f(x_1),f(x_2),...,f(x_n)]^T$.

In a Bayesian setting we can directly place the GP as a prior on the function defining the likelihood. This leads us directly to a formulation given Bayes relation with $\boldsymbol{\theta}=\{\boldsymbol{\theta}_{\mathcal{L}},\boldsymbol{\theta}_c\}$

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)=\frac{p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)p(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c)}{p\left(\mathcal{Y}|\boldsymbol{\theta},\mathcal{X}\right)}. \tag{2}$$

The prior $p(\mathbf{f}|\mathcal{X},\boldsymbol{\theta}_c)$ is given by the GP and the likelihood $p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)$ is the two likelihood defined previously, with the usual assumption that the likelihood factorizes, i.e., $p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)=\prod_{k=1:m}p\left(y_k|f(u_k),f(v_k),\boldsymbol{\theta}_{\mathcal{L}}\right)$

The posterior of interest, $p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)$, is defined when equipped with the likelihood and the prior, but it is unfortunately not of any known analytical form, thus we rely on the Laplace approximation.

### 2.2. Inference & Hyperparameters

We apply the Laplace approximation and approximate the posterior by a multivariate Gaussian distribution, such that $p\left(\mathbf{f}|\mathcal{Y}\right)\approx\mathcal{N}(\mathbf{f}|\hat{\mathbf{f}},\mathbf{A}^{-1})$. Where $\hat{\mathbf{f}}$ is the mode of the posterior and $\mathbf{A}$ is the Hessian of the negative log-likelihood at the mode.

The mode is found as $\hat{\mathbf{f}}=\arg\max_{\mathbf{f}}\,p\left(\mathcal{Y}|\mathbf{f}\right)p\left(\mathbf{f}\right)$. We solve the problem by considering the unnormalized log-posterior and the resulting cost function which is to be maximized, is given by

$$\begin{aligned}\psi\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)=&\log p\left(\mathcal{Y}|\mathbf{f},\mathcal{X},\boldsymbol{\theta}_{\mathcal{L}}\right)-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\\&-\frac{1}{2}\log|\mathbf{K}|-\frac{N}{2}\log 2\pi.\end{aligned} \tag{3}$$

where $\mathbf{K}_{i,j}=k(x_i,x_j)_{\boldsymbol{\theta}_c}$. We use a damped Newton method with soft linesearch to maximize Eq. (3). In our case the basic damped Newton step (with adaptive damping factor $\lambda$) can be calculated without inversion of the Hessian (see [7])

$$\begin{aligned}\mathbf{f}^{new}=&\left(\mathbf{K}^{-1}+\mathbf{W}-\lambda\mathbf{I}\right)^{-1}\\&\cdot\left[\left(\mathbf{W}-\lambda\mathbf{I}\right)\mathbf{f}+\nabla\log p(\mathcal{Y}|\mathbf{f},\mathcal{X},\boldsymbol{\theta}_{\mathcal{L}})\right],\end{aligned} \tag{4}$$

Using the notation $\nabla\nabla_{i,j}=\frac{\partial^2}{\partial f(x_i)\partial f(x_j)}$ we apply the definition $\mathbf{W}_{i,j}=-\sum_k\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}})$. We note that the term $\nabla\nabla_{i,j}\log p(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}})$ is only nonzero when both $x_i$ and $x_j$ occur as either $v_k$ or $u_k$ in $\mathbf{f}_k$. In contrast to standard binary GP classification, the negative Hessian, $\mathbf{W}$ is not diagonal, which makes the approximation slightly more involved. When converged, the resulting approximation is

$$p\left(\mathbf{f}|\mathcal{Y},\mathcal{X},\boldsymbol{\theta}\right)\approx\mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}},\left(\mathbf{W}+\mathbf{K}^{-1}\right)^{-1}\right). \tag{5}$$

We refer to [7] for a full derivation and for the required derivatives as first outlined in [5]. Parameters in the likelihood and covariance function, collected in $\theta$, are found by evidence optimization using a standard BFGS method.

### 2.3. Predictions & Evaluations

Given the model, in essence defined by $f$, we wish to make predictions of the observed variable $y$ for a pair of test inputs $r\in\mathcal{X}_t$ and $s\in\mathcal{X}_t$. We are especially interested in the discrete decision, i.e., whether $r$ is preferred over $s$ denoted by $r\succ s$, or vice versa. Omitting the conditioning on $\mathcal{X}$ and $\mathcal{X}_t$, we can write the joint prior distribution between $\mathbf{f}\sim p\left(\mathbf{f}|\mathcal{Y},\boldsymbol{\theta}\right)$ and the test variables $\mathbf{f}_t=[f\left(r\right),f\left(s\right)]^T$ as

$$\begin{bmatrix}\mathbf{f}\\\mathbf{f}_t\end{bmatrix}=\mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\\mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{K}&\mathbf{k}_t\\\mathbf{k}_t^T&\mathbf{K}_t\end{bmatrix}\right), \tag{6}$$

where $\mathbf{k}_t$ is a matrix with elements $\mathbf{k}_{2,i}=k(s,x_i)_{\boldsymbol{\theta}_c}$ and $\mathbf{k}_{1,i}=k(r,x_i)_{\boldsymbol{\theta}_c}$ with $x_i$ being a training input. The conditional $p\left(\mathbf{f}_t|\mathbf{f}\right)$ is obviously Gaussian as well and can be obtained directly from Eq. (6). The predictive distribution is given as $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)=\int p\left(\mathbf{f}_t|\mathbf{f}\right)p\left(\mathbf{f}|\mathcal{Y},\boldsymbol{\theta}\right)d\mathbf{f}$. With the posterior approximated with the Gaussian from the Laplace approximation, then $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)$ will be Gaussian too and is given as $\mathcal{N}\left(\mathbf{f}_t|\mu^*,\mathbf{K}^*\right)$ with $\mu^*=[\mu_r^*,\mu_s^*]^T=\mathbf{k}_t\mathbf{K}^{-1}\hat{\mathbf{f}}$ and

$$\mathbf{K}^*=\begin{bmatrix}\mathbf{K}_{rr}^*&\mathbf{K}_{rs}^*\\\mathbf{K}_{sr}^*&\mathbf{K}_{ss}^*\end{bmatrix}=\mathbf{K}_t-\mathbf{k}_t^T\left(\mathbf{I}+\mathbf{W}\mathbf{K}\right)\mathbf{k}_t,$$

where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from Eq. (5). With the predictive distribution for $\mathbf{f}_t$, the final prediction of the observed variable is available from

$$p\left(y_t|\mathcal{Y},\boldsymbol{\theta}\right) = \int p\left(y_t|\mathbf{f}_t,\boldsymbol{\theta}_{\mathcal{L}}\right) p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right) d\mathbf{f}_t \tag{7}$$

If the likelihood is an odd function, as in our case, the binary preference decision between $r$ and $s$ can be made directly from $p\left(\mathbf{f}_t|\mathcal{Y}\right)$.

If $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)$ is Gaussian and we consider the Probit likelihood, the integral in Eq. (7) can be evaluated in closed form as a modified Probit function given by [5]

$$P\left(r \succ s|\mathcal{Y}\right) = \Phi\left(\left(\mu_r^* - \mu_s^*\right)/\sigma^*\right) \tag{8}$$

with $\left(\sigma^*\right)^2 = 2\sigma^2 + \mathbf{K}_{rr}^* + \mathbf{K}_{ss}^* - \mathbf{K}_{rs}^* - \mathbf{K}_{sr}^*$

### 2.4. Kernels for Audio Preference

We suggest a general purpose covariance function for audio modeling tasks with GPs. It can easily integrate different modalities and meta-data types, such as audio features, tags, lyrics and subject features. The general covariance function is defined as

$$k\left(x,x'\right) = \left(\sum\nolimits_{i=1}^{N_a} k_i\left(x_a,{x_a}'\right)\right) k_u\left(x_u,{x_u}'\right) \tag{9}$$

where the first factor is the sum of all the $N_a$ covariance functions defining the correlation structure of the audio part, $x_a$, of the complete instance, $x$. The second factor, or multi-task part, is the covariance function defining the covariance structure of the subject meta-data part, $x_u$. The practical evaluation is limited to the a individualized setting using only $x_a$, thus $k\left(x,x'\right) = k\left(x_a,{x_a}'\right)$, where we apply the probability product kernel formulation [8]. The probability product kernel is defined directly as an inner product, i.e., $k\left(x_a,{x_a}'\right) = \int \left[p\left(x_a\right) p\left({x_a}'\right)\right]^q dx$, where $p\left(x_a\right)$ is a density estimate of each audio track feature distribution. In this evaluation we fix $q = 1/2$, leading to the Hellinger divergence [8]. As custom in the audio community, see e.g. [9], we will resort to a (finite) Gaussian Mixture Model (GMM) in order to model the feature distribution. So $p(x)$ is in general given by $p\left(x\right) = \sum_{z=1}^{N_z} p\left(z\right) p\left(x|z\right)$, where $p\left(x|z\right) = \mathcal{N}\left(x|\mu_z,\sigma_z\right)$ is a standard Gaussian distribution. The kernel can be calculated in closed form [8] as.

$$k\left(p_a\left(x\right),p_a\left(x\right)\right) = \sum_z \sum_{z'} \left(p_a\left(z\right) p_{a'}\left(z'\right)\right)^q \tilde{k}\left(p\left(x|\theta_z\right),p\left(x|\theta_{z'}\right)\right) \tag{10}$$

where $\tilde{k}\left(p\left(x|\theta_z\right),p\left(x|\theta_{z'}\right)\right)$ is the probability product kernel between two single components, which is also available in closed form [8].

## 3. EXPERIMENT

In order to evaluate the model proposed in section in 2, we consider a small-scale dataset which is publicly available [3]. Specifically it consist of 10 test subjects, 30 audio tracks and 10 audio tracks per genre. The genres are Classical, Heavy Metal and Rock/Pop.

The experiment is based on a partial, full pairwise design, so that $155$ out of the $420$ combinations was evaluated by each of the 10 subjects. We extract standard audio features from the audio tracks, namely MFCCs (26 incl. delta coefficients). A GMM was fitted to each track distribution with a fixed model complexity of $N_z = 2$ and each components restricted to a diagonal covariance structure. Parameters where fitted using a standard maximum likelihood based EM algorithm using K-means initialization.

The experiment itself was conducted using a Matlab interface in a 2-Alternative-Forced-Choice setup inline with the model. The interface allowed subjects to listen to the two presented tracks as many times they wanted before making a choice between them. A questionnaire gathered subject meta-data such as, age, musical training, context and a priori genre preference. This data is, however, not used in this individualized evaluation, but can easily be applied in the multi-task kernel suggested in Sec. 2.4.

In the evaluation we are primarily interested in two aspects. The first, and main result, is an estimate of the generalization error on new unseen tracks, e.g., relevant for recommendation purposes. In order to evaluate this, we make an extensive cross-validation using a 30-fold cross-validation in which each track (incl. all connected comparisons) is left out once; the model with $\sigma = 1$ is then trained on 10 random subsets of tracks for each training set size, which results in an estimated of the average test error. The resulting learning curve is shown in Fig. 1 with the box plot illustrating the distribution of the average subject performance. When considering $N_{\text{tracks}} = 29$ we obtain an average prediction performance of $74.2\%$ , which is the main result in a typical (individual) recommendation scenario.

Secondly, we investigate how many pairwise comparisons the model requires in order to learn the individual preferences. This is evaluated using a 10-fold cross-validation over the comparisons which gives the learning curve in Fig. 2. We notice that on average we only require approximately $40\%$ or 56 comparisons in order to reach the $25\%$ level, corresponding to approximately two comparisons per track.

## 4. DISCUSSION & CONCLUSION

We have outlined a pairwise regression model based on Gaussian Process priors for modeling and predicting the pairwise preference of music. We proposed an appropriate covariance structure suitable for audio features (such as MFCCs) based on generative models of audio features. The general version

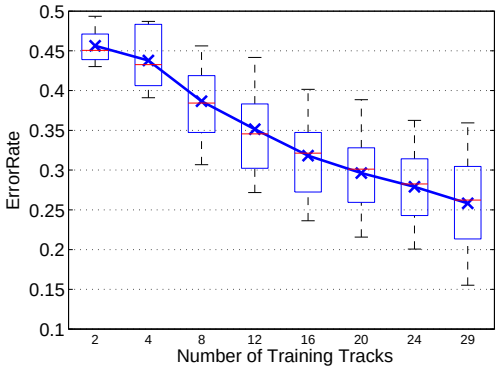



**Fig. 1**. Mean learning curve (blue line) and box plot over subjects. Leave-one-(track)-out test error as a function of the number of tracks in the training set. Thus, there can maximum be 29 tracks in the training set to predict the preference between the left out track and the rest. The baseline is 0.5 corresponding to random guessing.

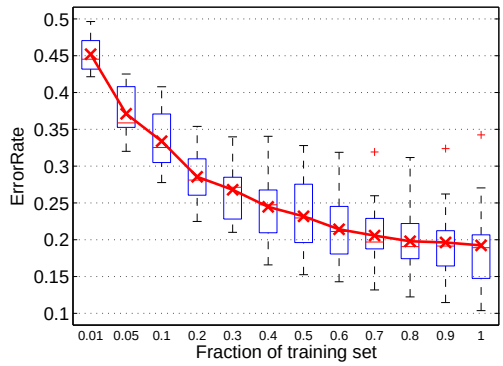



**Fig. 2**. Mean learning curve (red line) and box plot over the subjects mean performance. Test error rate as a function of the number of pairwise comparisons in the the training set. Notice that a fraction of one corresponds to $(155 \cdot 90\%)/420 \sim 33.2\%$ of all possible pairwise experiments.

of the covariance function allows for multi-task scenarios and feature integration. We evaluated the setup in a individual scenario in which we showed a 74% average accuracy. This indicates that there might very well be a promising upper bound on the number of required pairwise comparisons in this music setting, in effect implying that the specified correlation structure makes sense. This will ensure that the required number of pairwise comparisons does not scale quadratically when including more tracks.

We furthermore observe a difference among the different subjects indicating that some subjects may have a very consistent preference, possibly aligning well with the applied covariance function, while others seem very difficult to predict (observed as outliers in the box plot). We speculate that the pairwise approach to music preference is only possible for certain groups of subjects and/or in special contexts, which is to be investigated in future research.

The current model is intended for modeling personal preferences over a small/medium size dataset. For large datasets with millions of tracks, we see sparse techniques using pseudo-inputs and sequential selection as a powerful combination to scale the model and only use informative comparisons. Furthermore, a direct comparison between classic collaborative filtering with absolute ratings is obvious when a suitable dataset supporting is available.

In conclusion we have proposed a novel rating and modeling paradigm for eliciting music preference using pairwise comparisons. We conducted a preliminary evaluation of the performance on a small dataset and find the results promising for robust elicitation of music and audio preference in general.

## 5. REFERENCES


[1] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno, "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 435–447, 2008.

[2] R.D. Bock and J.V. Jones, "The measurement and prediction of judgment and choice," 1968.

[3] B. S. Jensen, J. S. Gallego, and J. Larsen, "A Predictive Model of Music Preference using Pairwise Comparisons - Supporting Material and Dataset," www.imm.dtu.dk/pubdb/p.php?6143.

[4] L L Thurstone, "A law of comparative judgement," *Psychological Review*, vol. 34, 1927.

[5] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.

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

[7] B. S. Jensen and J. B. Nielsen, "Pairwise Judgements and Absolute Ratings with Gaussian Process priors," Techinal Report, DTU Informatics, September 2011.

[8] T. Jebara and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[9] A. Meng and J. Shawe-Taylor, "An investigation of feature models for music genre classification using the support vector classifier," in *International Conference on Music Information Retrieval*, 2005, pp. 604–609.

Appendix E

# Towards Predicting Expressed Emotion in Music from Pairwise Comparisons

# 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 $\tau$ 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 $\tau$.

## 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 [12]. The other direction is to continuously measure expressed emotions in music directly in e.g the arousal and valence space (AV space) [11] 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 $\tau$ 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:

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

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

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

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 $\lambda$ 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 $\mathbf{K}^* = \mathbf{K}_t - \mathbf{k}_t^T (\mathbf{I} + \mathbf{W}\mathbf{K})\, \mathbf{k}_t$, 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 $\sigma_f$ is a variance term and $\sigma_l$ 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 $\theta$ 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.

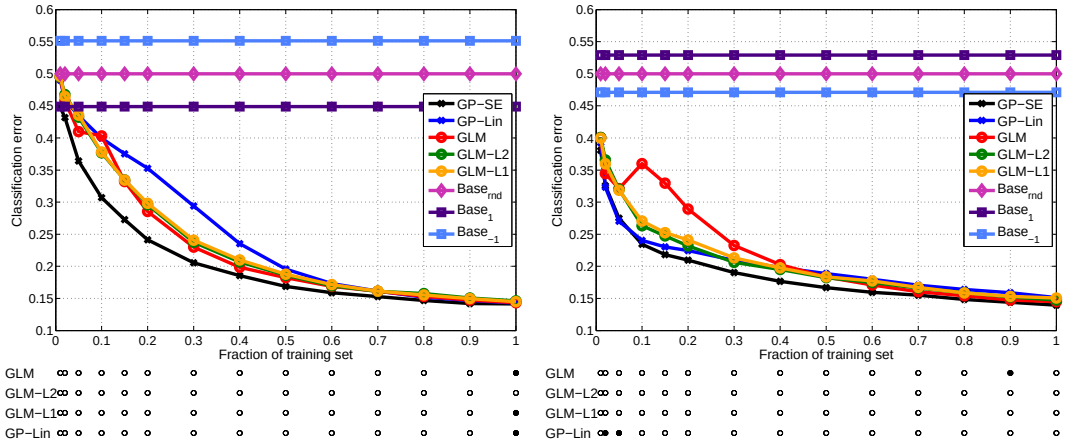
[2] More details can be found in e.g. [19].

[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

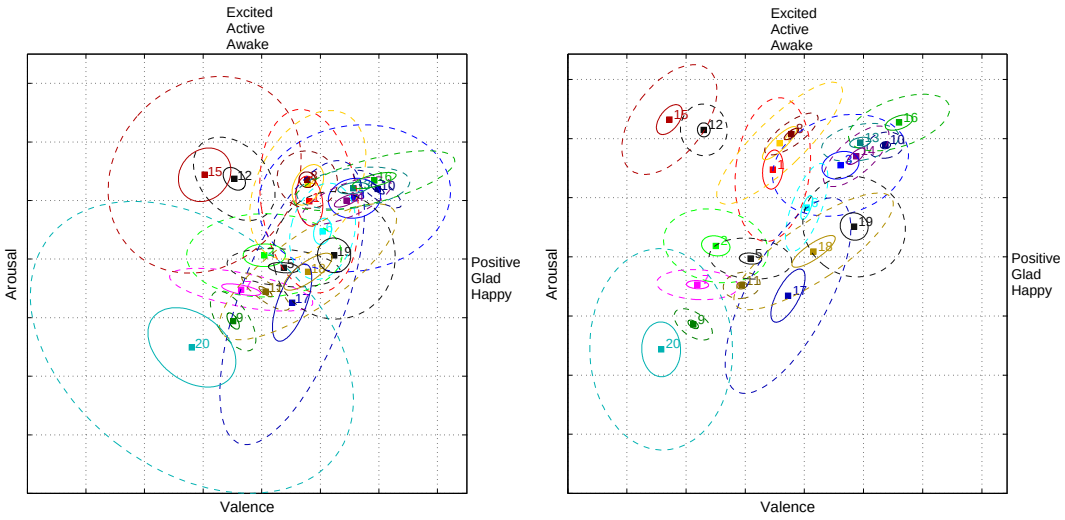
The learning curves show the performance of the models when predicting unseen comparisons. However, it may be difficult to interpret in terms of the typical AV space as one know from direct scaling experiments. To address this we show that both the GLM and the GP models can provide an internal, but unit 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 $\tau$ 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 $\tau$ 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 $\tau$ above $0.95$. Between the GLM and the regularized models the Kendall's $\tau$ 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 $\tau$ 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.

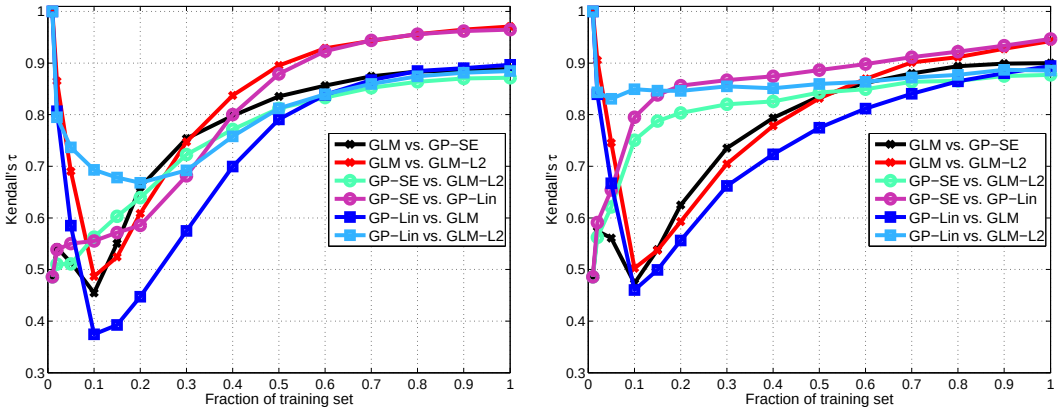
We have established that there is a difference in ranking of excerpts on the dimensions of valence and arousal given which models is chosen. As was shown in Figure 2 there is also a large difference in ranking across subjects, alternatively these individual differences can be quantified using the rank correlation. Using the GP-SE model trained on all the dataset, the Kendall's $\tau$ 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 $\tau$ 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 $\tau$ rank correlation coefficient. Curves are an average of the Kendall's $\tau$ 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 $\tau$ 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 $\tau$ 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 $\tau$ 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

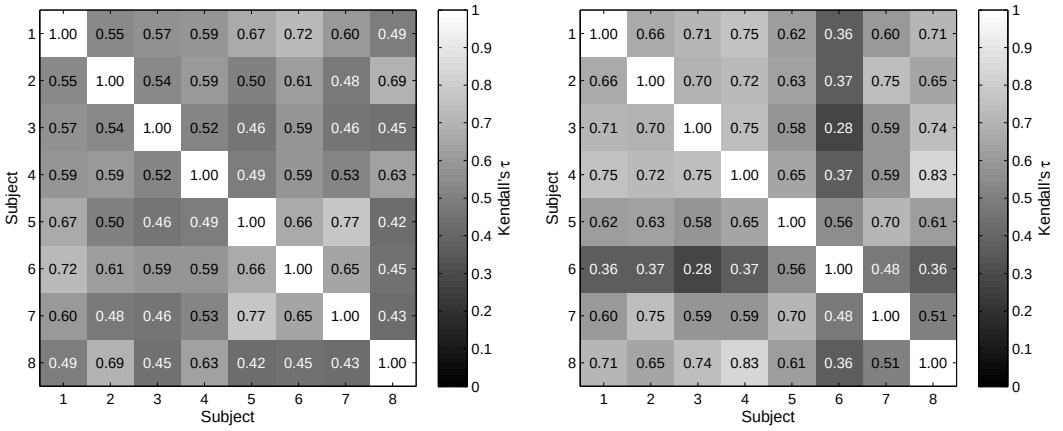

[1] K. Trochidis, C. Delbé, and E. Bigand, "Investigation of the relationships between audio features and induced emotions in Contemporary Western music," *8th Sound and Music Computing Conference*, 2011.

[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 $\tau$ on the latent regression function output of the GP-SE model trained on the entire dataset. The values are averages of the Kendall's $\tau$ 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. Schubert, "Measurement and time series analysis of emotion in music," Ph.D. dissertation, University of New South Wales, 1999.

[12] 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.

[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.

Appendix F

# Modeling Expressed Emotions in Music using Pairwise Comparisons

---

Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen, Modeling Expressed Emotions in Music using Pairwise Comparisons, 9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012), pages: 526-533, 2012.

# Modeling Expressed Emotions in Music using Pairwise Comparisons

Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen, and Jan Larsen *

Technical University of Denmark,
Department of Informatics and Mathematical Modeling,
Richard Petersens Plads B321, 2800 Lyngby, Denmark
{jenma; jenb; bjje; jl}@imm.dtu.dk

**Abstract.** We introduce a two-alternative forced-choice experimental paradigm to quantify expressed emotions in music using the two well-known arousal and valence (AV) dimensions. In order to produce AV scores from the pairwise comparisons and to visualize the locations of excerpts in the AV space, we introduce a flexible Gaussian process (GP) framework which learns from the pairwise comparisons directly. A novel dataset is used to evaluate the proposed framework and learning curves show that the proposed framework needs relative few comparisons in order to achieve satisfactory performance. This is further supported by visualizing the learned locations of excerpts in the AV space. Finally, by examining the predictive performance of the user-specific models we show the importance of modeling subjects individually due to significant subjective differences.

**Keywords:** expressed emotion, pairwise comparison, Gaussian process

## 1 Introduction

In recent years Music Emotion Recognition has gathered increasing attention within the Music Information Retrieval (MIR) community and is motivated by the possibility to recommend music that expresses a certain mood or emotion.

The design approach to automatically predict the expressed emotion in music has been to describe music by structural information such as audio features and/or lyrical features. Different models of emotion, e.g., categorical [1] or dimensional [2], have been chosen and depending on these, various approaches have been taken to gather emotional ground truth data [3]. When using dimensional models such as the well established *arousal* and *valence* (AV) model [2] the majority of approaches has been to use different variations of self-report direct scaling listening experiments [4].

* This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886, and 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.

Direct-scaling methods are fast ways of obtaining a large amount of data. However, the inherent subjective nature of both induced and expressed emotion, often makes anchors difficult to define and the use of them inappropriate due to risks of unexpected communication biases. These biases occur because users become uncertain about the meaning of scales, anchors or labels [5]. On the other hand, lack of anchors and reference points makes direct-scaling experiments susceptible to drift and inconsistent ratings. These effects are almost impossible to get rid of, but are rarely modeled directly. Instead, the issue is typically addressed through outlier removal or simply by averaging across users [6], thus neglecting individual user interpretation and user behavior in the assessment of expressed emotion in music.

Pairwise experiments eliminates the need for an absolute reference anchor, due to the embedded relative nature of pairwise comparisons which persists the relation to previous comparisons. However, pairwise experiments scale badly with the number of musical excerpts which they accommodate in [7] by a tournament based approach that limits the number of comparisons and transforms the pairwise judgments into possible rankings. Subsequently, they use the transformed rankings to model emotions.

In this paper, we present a novel dataset obtained by conducting a controlled pairwise experiment measuring expressed emotion in music on the dimensions of valence and arousal. In contrast to previous work, we learn from pairwise comparisons, directly, in a principled probabilistic manner using a flexible Gaussian process model which implies a latent but interpretable valence and arousal function. Using this latent function we visualize excerpts in a 2D valance and arousal space which is directly available from the principled modeling framework. Furthermore the framework accounts for inconsistent pairwise judgments by participants and their individual differences when quantifying the expressed emotion in music. We show that the framework needs relatively few comparisons in order to predict comparisons satisfactory, which is shown using computed learning curves. The learning curves show the misclassification error as a function of the number of (randomly chosen) pairwise comparisons.

## 2 Experiment

A listening experiment was conducted to obtain pairwise comparisons of expressed emotion in music using a two-alternative forced-choice paradigm. 20 different 15 second excerpts were chosen from the USPOP2002[1] dataset. The 20 excerpts were chosen such that a linear regression model developed in previous work [8] maps exactly 5 excerpts into each quadrant of the two dimensional AV space. A subjective evaluation was performed to verify that the emotional expression throughout each excerpt was considered constant.

A sound booth provided neutral surroundings for the experiment and the excerpts were played back using headphones to the 8 participants (2 female,

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

6 male). Written and verbal instructions were given prior to each session to ensure that subjects understood the purpose of the experiment and were familiar with the two emotional dimensions of 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 were evaluated individually in random order. The details of the experiment are available in [9].

## 3 Pairwise-Observation based Regression

We aim to construct a model for the dataset given the audio excerpts in the set $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ with $n = 20$ distinct excerpts, each described by an input vector $\mathbf{x}_i$ of audio features extracted from the excerpt. For each test subject the dataset comprises of all $m = 190$ 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 (for each subject) as $\mathcal{Y} = \{(d_k; u_k, v_k) | k = 1, ..., m\}$, where $d_k \in \{-1, 1\}$ indicates which of the two excerpts that had the highest valence or arousal. $d_k = -1$ means that the $u_k$'th excerpt is picked over the $v_k$'th and visa versa when $d_k = 1$.

We model the pairwise choice, $d_k$, between two distinct excerpts, $u$ and $v$, as a function of the difference between two functional values, $f(\mathbf{x}_u)$ and $f(\mathbf{x}_v)$. The function $f : \mathcal{X} \to \mathbb{R}$ thereby defines an internal, but latent absolute reference of either valence or arousal as a function of the excerpt represented by the audio features.

Given a function, $f(\cdot)$, we can define the likelihood of observing the choice $d_k$ directly as the conditional distribution.

$$p\left(d_k | \mathbf{f}_k\right) = \Phi\left(d_k \frac{f\left(\mathbf{x}_{v_k}\right) - f\left(\mathbf{x}_{u_k}\right)}{\sqrt{2}}\right), \tag{1}$$

where $\Phi(x)$ is the cumulative Gaussian (with zero mean and unity variance) and $\mathbf{f}_k = \left[f\left(\mathbf{x}_{u_k}\right), f\left(\mathbf{x}_{v_k}\right)\right]^\top$. This classical choice model can be dated back to Thurstone and his fundamental definition of *The Law of Comparative Judgment* [10].

We consider the likelihood in a Bayesian setting such that $p\left(\mathbf{f} | \mathcal{Y}, \mathcal{X}\right) = p\left(\mathcal{Y} | \mathbf{f}\right) p(\mathbf{f} | \mathcal{X}) / p\left(\mathcal{Y} | \mathcal{X}\right)$ where we assume that the likelihood factorizes, i.e., $p\left(\mathcal{Y} | \mathbf{f}\right) = \prod_{k=1}^{m} p\left(d_k | \mathbf{f}_k\right)$.

In this work we consider a specific prior, namely a Gaussian Process (GP), first considered with the pairwise likelihood in [11]. A GP is typically defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" [12]. By $f\left(\mathbf{x}\right) \sim \mathcal{GP}\left(\mathbf{0}, k(\mathbf{x}, \mathbf{x}')\right)$ we denote that the function $f(\mathbf{x})$ is modeled by a zero-mean GP with covariance function $k(\mathbf{x}, \mathbf{x}')$. The fundamental consequence of this formulation is that the GP can be considered a distribution over functions, defined as $p\left(\mathbf{f} | \mathcal{X}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$ for any finite set of of function values $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$, 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 analytical tractable. Instead, we use the *Laplace Approximation* to approximate the posterior with a multivariate Gaussian distribution[1].

To predict the pairwise choice $d_t$ 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 $f(\mathbf{x}_s)$. Given the GP, we can write the joint distribution between $\mathbf{f} \sim p(\mathbf{f}|\mathcal{Y}, \mathcal{X})$ and the test variables $\mathbf{f}_t = [f(\mathbf{x}_r), f(\mathbf{x}_s)]^T$ as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_t \end{bmatrix} = \mathcal{N}\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{k}_t \\ \mathbf{k}_t^T & \mathbf{K}_t \end{bmatrix} \right), \tag{2}$$

where $\mathbf{k}_t$ 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.

The conditional $p(\mathbf{f}_t|\mathbf{f})$ is directly available from Eq. (2) as a Gaussian too. The predictive distribution 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 $\mathcal{N}(\mathbf{f}_t|\boldsymbol{\mu}^*, \mathbf{K}^*)$ with $\boldsymbol{\mu}^* = \mathbf{k}_t^T \mathbf{K}^{-1} \hat{\mathbf{f}}$ and $\mathbf{K}^* = \mathbf{K}_t - \mathbf{k}_t^T (\mathbf{I} + \mathbf{W}\mathbf{K})\, \mathbf{k}_t$, where $\hat{\mathbf{f}}$ and $\mathbf{W}$ are obtained from the Laplace approximation (see [13]). In this paper we are only interested in the binary choice $d_t$, which is determined by which of $f(\mathbf{x}_r)$ or $f(\mathbf{x}_s)$ that dominates[2].

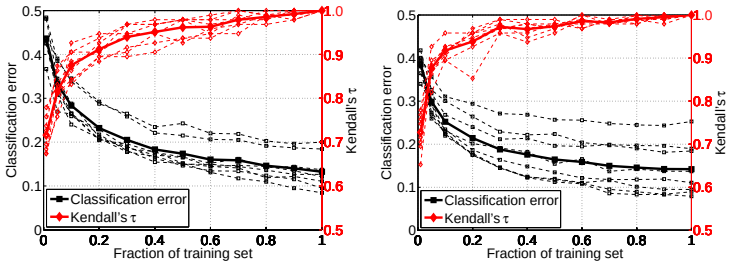
The zero-mean GP is fully defined by the covariance function, $k(\mathbf{x}, \mathbf{x}')$. In the emotion dataset each input instance is an excerpt described by the vector $\mathbf{x}$ containing the audio features for each time frame which is naturally modeled with a probability density, $p(\mathbf{x})$. We apply the probability product (PP) kernel [14] in order to support these types of distributional inputs. The PP kernel is defined directly as an inner product as $k(\mathbf{x}, \mathbf{x}') = \int [p(\mathbf{x})\, p(\mathbf{x}')]^q d\mathbf{x}$. We fix $q = 1/2$, leading to the Hellinger divergence [14]. In order to model the audio feature distribution for each excerpt, we resort to a (finite) Gaussian Mixture Model (GMM). Hence, $p(\mathbf{x})$ is given by $p(\mathbf{x}) = \sum_{z=1}^{N_z} p(z)\, p(\mathbf{x}|z)$, where $p(\mathbf{x}|z) = \mathcal{N}(\mathbf{x}|\mu_z, \sigma_z)$ is a standard Gaussian distribution. The kernel is expressed in closed form [14] as $k(p(\mathbf{x}), p(\mathbf{x}')) = \sum_z \sum_{z'} (p(z)\, p(z'))^q \tilde{k}(p(\mathbf{x}|\theta_z), p(\mathbf{x}'|\theta_{z'}))$ where $\tilde{k}(p(\mathbf{x}|\theta_z), p(\mathbf{x}'|\theta_{z'}))$ is the probability product kernel between two single components - also available in closed form [14].

## 4 Modeling Expressed Emotion

In this section we evaluate the ability of the proposed framework to capture the underlying structure of expressed emotions based on pairwise comparisons, directly. We apply the GP model using the probability product (PP) kernel described in Section 3 with the inputs based on a set of audio features extracted

[1] More details can be found in e.g. [13].

[2] With the pairwise GP model the predictive distribution of $d_t$ can also be computed analytically (see [13]) and used to express the uncertainty in the prediction relevant for e.g. sequential designs, reject regions etc.




(a) Valence (b) Arousal

**Fig. 1.** Classification error learning curves and Kendall's $\tau$ for 10-fold CV on comparisons. Bold lines are mean curves across subjects and dash lines are curves for individual subjects. Notice, that for the classification error learning curves, the baseline performance corresponds to an error of 0.5, obtained by simply randomly guessing the pairwise outcome.

from the 20 excerpts. By investigating various combinations of features we obtained the best performance using two sets of commonly used audio features. The first set is the Mel-frequency cepstral coefficients (MFCC), which describe the short-term power spectrum of the signal. Secondly, we included spectral contrast features and features describing the spectrum of the Hanning windowed audio. Based on an initial evaluation, we fix the number of components in the GMM used in the PP Kernel to $N_z = 3$ components and train the individual GMMs by a standard EM algorithm with K-means initialization. Alternatively, measures such as the Bayesian Information Criterion (BIC) could be used to objectively set the model complexity for each excerpt.

### 4.1 Results: Learning Curves

Learning curves for the individual subjects are computed using 10-fold cross validation (CV) in which a fraction (90%) of the total number of pairwise comparisons constitutes the complete training set. Each point on the learning curve is an average over 10 randomly chosen and equally-sized subsets from the complete training set. The Kendall's $\tau$ rank correlation coefficient is computed in order to relate our results to that of e.g. [7] and other typical ranking based applications. The Kendall's $\tau$ 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. The reported Kendall's $\tau$ is in all cases calculated with respect to the predicted ranks using all the excerpts.

Figure 1 displays the computed learning curves. With the entire training set included the mean classification errors across subjects for valence and arousal are 0.13 and 0.14, respectively. On average this corresponds to a misclassified comparison in every 7.5 and 7'th comparison for valence and arousal, respectively.

For valence, the mean classification error across users is below 0.2 with 40% of the training data included, whereas only 30% of the training data is needed to obtain similar performance for arousal. This indicates that the model for arousal can be learned slightly faster than valence. Using 30% of the training data the Kendall's $\tau$ is 0.94 and 0.97, respectively, indicating a good ranking performance using only a fraction of the training data.

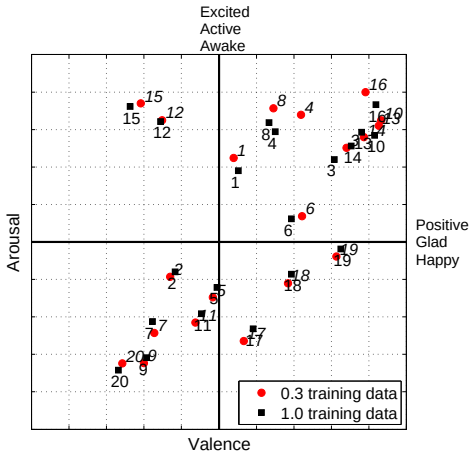
When considering the learning curves for individual users we notice significant individual differences between users—especially for arousal. Using the entire training set in the arousal experiment, the user for which the model performs best results in an error of 0.08 whereas the worst results in an error of 0.25. In the valence experiment the best and worst performances result in classification errors of 0.08 and 0.2, respectively.

### 4.2 Results: AV space

The learning curves show the pure predictive power of the model on unseen comparisons, but may be difficult to interpret in terms of the typical AV space. To address this we show that the latent regression function $f(\cdot)$ provides an internal but unit free representation of the AV scores. The only step required is a normalization which ensures that the latent values are comparable across folds and subjects. In Figure 2 the predicted AV scores are shown when the entire training set is included and when only 30% is included. The latter corresponds to 51 comparisons in total or an average of 2.5 comparisons per excerpt. The results are summarized by averaging across the predicted values for each user. 15 of the 20 excerpts are positioned in the typical high-valence high-arousal and low-valence low-arousal quadrants, 2 excerpts are clearly in the low-valance high-arousal quadrant and 3 excerpts are in the high-valance low-arousal quadrant of the AV space. The minor difference in predictive performance between 30% and the entire training dataset does not lead to any significant change in AV scores, which is in line with the reported Kendall's $\tau$ measure.

### 4.3 Discussion

The results clearly indicate that it is possible to model expressed emotions in music by directly modeling pairwise comparisons in the proposed Gaussian process framework using subject specific models. An interesting point is the large difference in predictive performance between subjects given the specific models. These differences can be attributed to the specific model choice (including kernel) or simply to subject inconsistency in the pairwise decisions. The less impressive predictive performance for certain subjects is presumably a combination of the two effects, although given the very flexible nature of the Gaussian process model, we mainly attribute the effect to subjects being inconsistent due to for example mental drift. Hence, individual user behavior, consistency and discriminative ability are important aspects of modeling expressed emotion in music and other cognitive experiments, and thus also a critical part when aggregating subjects in large datasets.




| 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 |

**Fig. 2.** AV values computed by averaging the latent function across folds and repetitions and normalizing for each individual model for each participant. Red circles: 30% of training set is used. Black squares: entire training set is used.

The flexibility and interpolation abilities of Gaussian Processes allow the number of comparisons to be significantly lower than the otherwise quadratic scaling of unique comparisons. This aspect and the overall performance should of course be examined further by considering a large scale dataset and the use of several model variations. In addition, the learning rates can be improved by combining the pairwise approach with active learning or sequential design methods, which in turn select only pairwise comparisons that maximize some information criterion.

We plan to investigate how to apply multi-task (MT) or transfer learning to the special case of pairwise comparisons, such that we learn one unifying model taking subjects differences into account instead of multiple independent subject-specific models. A very appealing method is to include MT learning in the kernel of the GP [15], but this might not be directly applicable in the pairwise case.

## 5 Conclusion

We introduced a two-alternative forced-choice experimental paradigm for quantifying expressed emotions in music in the typical arousal and valance (AV) dimensions. We proposed a flexible probabilistic Gaussian process framework to model the latent AV scales directly from the pairwise comparisons. The framework was evaluated on a novel dataset and resulted in promising error rates for both arousal and valence using as little as 30% of the training set corresponding to 2.5 comparisons per excerpt. We visualized AV scores in the well-known two dimensional AV space by exploiting the latent function in the Gaussian process

model, showing the application of the model in a standard scenario. Finally we especially draw attention to the importance of maintaining individual models for subjects due to the apparent inconsistency of certain subjects and general subject differences.

## References


1. K. Hevner, "Experimental studies of the elements of expression in music," *American journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.
2. J.A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.
3. Y.E. Kim, E.M. Schmidt, Raymond Migneco, B.G. Morton, Patrick Richardson, Jeffrey Scott, J.A. Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proc. of the 11th Intl. Society for Music Information Retrieval (ISMIR) Conf*, 2010, pp. 255–266.
4. E. Schubert, *Measurement and time series analysis of emotion in music*, Ph.D. thesis, University of New South Wales, 1999.
5. M. Zentner and T. Eerola, *Handbook of Music and Emotion - Theory, Research, Application*, chapter 8 - Self-report measures and models, Oxford University Press, 2010.
6. A. Huq, J. P. Bello, and R. Rowe, "Automated Music Emotion Recognition: A Systematic Evaluation," *Journal of New Music Research*, vol. 39, no. 3, pp. 227–244, Sept. 2010.
7. Y.-H. Yang and H.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.
8. J. Madsen, *Modeling of Emotions expressed in Music using Audio features*, DTU Informatics, Master Thesis, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6036, 2011.
9. J. Madsen, *Experimental Protocol for Modelling Expressed Emotion in Music*, DTU Informatics, http://www2.imm.dtu.dk/pubdb/p.php?6246, 2012.
10. L. L. Thurstone, "A law of comparative judgement.," *Psychological Review*, vol. 34, 1927.
11. W. Chu and Z. Ghahramani, "Preference learning with Gaussian Processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.
12. C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
13. B.S. Jensen and J.B. 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.
14. T. Jebara and A. Howard, "Probability Product Kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.
15. E.V. Bonilla, F.V. Agakov, and C.K.I. Williams, "Kernel multi-task learning using task-specific features," *Proceedings of the 11th AISTATS*, 2007.

Appendix G

# Pseudo Inputs For Pairwise Learning With Gaussian Processes

---

Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen, Pseudo Inputs For Pairwise Learning With Gaussian Processes, IEEE International Workshop on Machine Learning for Signal Processing 2012.

2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

# PSEUDO INPUTS FOR PAIRWISE LEARNING WITH GAUSSIAN PROCESSES

*Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen*

DTU Informatics
Technical University of Denmark
Asmussens Alle B305, 2800 Kgs. Lyngby, Denmark
{jenb,bjje,jl}@imm.dtu.dk

## ABSTRACT

We consider learning and prediction of pairwise comparisons between instances. The problem is motivated from a perceptual view point, where pairwise comparisons serve as an effective and extensively used paradigm. A state-of-the-art method for modeling pairwise data in high dimensional domains is based on a classical pairwise probit likelihood imposed with a Gaussian process prior. While extremely flexible, this non-parametric method struggles with an inconvenient $\mathcal{O}\left(n^3\right)$ scaling in terms of the $n$ input instances which limits the method only to smaller problems. To overcome this, we derive a specific sparse extension of the classical pairwise likelihood using the pseudo-input formulation. The behavior of the proposed extension is demonstrated on a toy example and on two real-world data sets which outlines the potential gain and pitfalls of the approach. Finally, we discuss the relation to other similar approximations that have been applied in standard Gaussian process regression and classification problems such as FI(T)C and PI(T)C.

## 1. INTRODUCTION

The pairwise learning setting has several application areas such as preference learning and ranking [1], metric learning [2] and general pairwise comparison paradigms. Pairwise comparisons are naturally motivated from a perceptual point of view, where human subjects make a sequence of pairwise (subjective) preference decisions in relation to sound quality, music taste, etc. The main advantage is that pairwise relations are relatively easy for subjects to convey consistently since subjects do not need an internal reference.

The theory underlying pairwise comparisons was first formulated in a principle manner in [3] stating *The Law of Comparative Judgments* building on cognitive and perceptual ideas. The basic idea is that a choice is determined by the difference in the response from a latent stochastic process. The resulting likelihood function in its simplest form—which is also by far the most common one—was first put into the flexible framework of Gaussian processes priors in [4].

Gaussian process based models are flexible and thus desirable for pairwise learning, but struggle with an inconvenient $\mathcal{O}\left(n^3\right)$ scaling in terms of the number of input instances $n$. This makes their use impractical for large-scale problems. Several suggestions have been proposed to remedy this issue for the standard Gaussian process regression case by using a smaller set of inputs that is either a subset of the original input set [5, 6] or a completely new set of *pseudo inputs* [7, 8, 9]. An unifying view of the latter family of models is given in [10] and extended in [11] leading to the well-known FI(T)C and PI(T)C approximations for standard regression and classification models.

In the standard case the explicit formulation of pseudo inputs can easily and without further considerations be turned into a conditional Gaussian process prior with an easy to invert covariance matrix. However, in the pairwise case the likelihood function depends on two variables. Therefore, we cannot immediately and without consideration use the standard approximations in the covariance as done in [12]. Instead, our quest to derive a sparse approximation for pairwise problems starts from the original pseudo-input formulation presented in [7]. Using this direct approach, our objective is to extend the pairwise likelihood model to allow for explicit sparsity in input space achieved by extending the model by a set of pseudo inputs—or inducing points—of size $l \ll n$. Essentially, the pseudo inputs are used to integrate out the two original variables of the classical pairwise likelihood function. In effect the Gaussian process prior is now placed over the function values of the pseudo inputs often resulting in a considerably lower computational load. Posterior inference relies on a Laplace approximation and the pseudo inputs can be found by evidence optimization for example initialized by k-means.

We give insight and intuition about the behavior and performance of the sparse model compared with the standard model by considering the *Boston housing* data set and a *wine-quality* data set. Examination of the out-of-sample error rates

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

978-1-4673-1026-0/12/$31.00 

is the basis for discussing the potential and limitations of the sparse model.

## 2. MODEL & EXTENSIONS

In this section we describe the general setup and frame the pairwise model in a Bayesian non-parametric setting. Each input instance $i$ is described by a feature vector $\mathbf{x} \in \mathbb{R}^d$ and $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$. Next, we consider a data set $\mathcal{Y} = \{y_k; u_k, v_k | k = 1, ..., m\}$ of pairwise relations $y \in \{-1, +1\}$ between the $u$'th and the $v$'th instance of $\mathcal{X}$, hence $\mathbf{x}_{u_k}, \mathbf{x}_{v_k} \in \mathcal{X}$ [1] . The two opposite choices picking either the $u$'th or the $v$'th instance are denoted by $y = -1$ and $y = +1$, respectively.

Given two latent function values $\mathbf{f}_k = [f(\mathbf{x}_{u_k}), f(\mathbf{x}_{v_k})]^\top$, the observations are modeled by a pairwise likelihood function $p\left(y_k | \mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L}\right)$ with parameter(s) $\boldsymbol{\theta}_\mathcal{L}$. The function $f$ is an latent function, which in a Thurstonian context [13], models the mean absolute response from the internal cognitive process when the subject is exposed to an input instance. The function parametrization admits that we directly place a zero-mean Gaussian process [14] prior on $f$ allowing for a flexible predictive model for the pairwise responses. Formally, we write $f(\mathbf{x}_i) \sim \mathcal{GP}\left(0, k_{\boldsymbol{\theta}_{\mathcal{GP}}}(\mathbf{x}_i, \cdot)\right)$, where $k(\cdot, \cdot)$ denotes a covariance function, or kernel, with parameter(s) $\boldsymbol{\theta}_{\mathcal{GP}}$, which generally speaking restricts the smoothness of the function. The fundamental consequence of a Gaussian process is that the joint distribution of a finite set of function values $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), f(\mathbf{x}_3), ..., f(\mathbf{x}_n)]^\top$ has a multivariate Gaussian distribution defined by $p\left(\mathbf{f} | \mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{K}_{\mathcal{X}\mathcal{X}}\right)$, where the elements of the covariance matrix are given as $[\mathbf{K}_{\mathcal{X}\mathcal{X}}]_{i,j} = k_{\boldsymbol{\theta}_{\mathcal{GP}}}(\mathbf{x}_i, \mathbf{x}_j)$. Given a standard Bayesian framework and assuming i.i.d. comparisons we now obtain the posterior over the function values

$$p\left(\mathbf{f} | \mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}\right) \propto p\left(\mathbf{f} | \mathcal{X}, \boldsymbol{\theta}_{\mathcal{GP}}\right) \prod_{k=1}^{m} p\left(y_k | \mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L}\right)$$

with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\mathcal{L}, \boldsymbol{\theta}_{\mathcal{GP}}\}$. The main computational issue in the Gaussian process framework is to calculate/approximate the posterior posing a $\mathcal{O}\left(n^3\right)$ scaling challenge due to the inversion of the kernel matrix.

### 2.1. Standard Pairwise Likelihood Function

The pairwise likelihood function described in a general pairwise context by [13] and used with Gaussian processes by e.g. [4] and [15] is given by

$$p\left(y_k | \mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L}\right) = \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma}\right), \tag{1}$$

where $\Phi(\cdot)$ defines a cumulative Gaussian (with zero mean and unity variance) and $\boldsymbol{\theta}_\mathcal{L} = \{\sigma\}$. The use of a Gaussian process prior in connection with this likelihood function was first proposed in [4].

[1] We will without loss of generality assume that the set $\mathcal{Y}$ involves all $n$ inputs instances in $\mathcal{X}$.

### 2.2. Sparse Pairwise Likelihood Function

To obtain sparsity in input space, we generally follow the ideas in [7]. Hence, given a set of pseudo inputs $\bar{\mathbf{X}}$, their functional values $\bar{\mathbf{f}}$ must originate from the same Gaussian process that was used for $\mathbf{f}$. Therefore, we can directly place a Gaussian process prior over $\bar{\mathbf{f}}$, i.e., $p\left(\bar{\mathbf{f}} | \bar{\mathbf{X}}\right) = \mathcal{N}\left(\bar{\mathbf{f}} | \mathbf{0}, \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}\right)$, where the matrix $\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}$ is the covariance matrix of the $l$ pseudo inputs collected in the matrix $\bar{\mathbf{X}} = [\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_l]$.

The overall idea of the pseudo-input formalism is now to refine the likelihood function from Eq. (1) such that the real $\mathbf{f}$ values that enter directly in the original, non-sparse likelihood function (through $\mathbf{f}_k$), exist only in the form of predictions from the pseudo inputs $\bar{\mathbf{f}}(\bar{\mathbf{X}})$. Given the listed assumptions, we formally have that $\mathbf{f}$ and $\bar{\mathbf{f}}$ are jointly Gaussian, hence

$$\begin{bmatrix} \mathbf{f}_k \\ \bar{\mathbf{f}} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\mathbf{x}_k\mathbf{x}_k} & \mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k}^\top \\ \mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k} & \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}} \end{bmatrix}\right), \tag{2}$$

where we define the following matrices and vectors

$$\mathbf{K}_{\mathbf{x}_k\mathbf{x}_k} = \begin{bmatrix} k(\mathbf{x}_{u_k}, \mathbf{x}_{u_k}) & k(\mathbf{x}_{u_k}, \mathbf{x}_{v_k}) \\ k(\mathbf{x}_{v_k}, \mathbf{x}_{u_k}) & k(\mathbf{x}_{v_k}, \mathbf{x}_{v_k}) \end{bmatrix} \tag{3}$$

$$\mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k} = [\mathbf{k}_{u_k}, \mathbf{k}_{v_k}] \tag{4}$$

with $[\mathbf{k}_{u_k}]_i = k(\bar{\mathbf{x}}_i, \mathbf{x}_{u_k})$ and $[\mathbf{k}_{v_k}]_i = k(\bar{\mathbf{x}}_i, \mathbf{x}_{v_k})$. From Eq. (2) it is trivial to find the conditional distribution of $\mathbf{f}_k$ given $\bar{\mathbf{f}}$, hence the sparse likelihood function can be derived in terms of $\bar{\mathbf{f}}$ by integrating over $\mathbf{f}_k$, thus

$$\begin{aligned} & p\left(y_k | \mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \bar{\mathbf{X}}, \bar{\mathbf{f}}, \boldsymbol{\theta}\right) \\ &= \int p\left(y_k | \mathbf{f}_k, \boldsymbol{\theta}_\mathcal{L}\right) p\left(\mathbf{f}_k | \bar{\mathbf{f}}, \bar{\mathbf{X}}\right) d\mathbf{f}_k \\ &= \int \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma}\right) \mathcal{N}\left(\mathbf{f}_k | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) d\mathbf{f}_k \\ &= \Phi\left(y_k \frac{\mu_{u_k} - \mu_{v_k}}{\sigma_k^*}\right) \end{aligned}$$

where $\boldsymbol{\mu}_k = [\mu_{u_k}, \mu_{v_k}]^\top$, $\mu_{u_k} = \mathbf{k}_{u_k}^\top \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}}$, $\mu_{v_k} = \mathbf{k}_{v_k}^\top \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}}$ and

$$\boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{u_k u_k} & \sigma_{u_k v_k} \\ \sigma_{v_k u_k} & \sigma_{v_k v_k} \end{bmatrix} = \mathbf{K}_{\mathbf{x}_k\mathbf{x}_k} - \mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k}^\top \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \mathbf{K}_{\bar{\mathbf{X}}\mathbf{x}_k}$$

Furthermore, $(\sigma_k^*)^2 = 2\sigma^2 + \sigma_{u_k u_k} + \sigma_{v_k v_k} - \sigma_{u_k v_k} - \sigma_{v_k u_k}$, which all together results in the pseudo-input likelihood

$$p\left(y_k | \mathbf{x}_{u_k}, \mathbf{x}_{v_k}, \bar{\mathbf{X}}, \bar{\mathbf{f}}, \boldsymbol{\theta}\right) = \Phi\left(z_k\right), \tag{5}$$

with $z_k = y_k \left(\mathbf{k}_{u_k}^T - \mathbf{k}_{v_k}^T\right) \mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} \bar{\mathbf{f}} / \sigma_k^*$.

### 2.3. Inference & Predictions

The likelihood functions described in Section 2.1 and 2.2 lead to intractable posteriors and call for approximation techniques or sampling methods. Our goal in this initial study is to examine the sparse model and its properties—not to provide the optimal approximation—hence, we only explore inference based on the Laplace approximation.

#### *2.3.1. Posterior Approximation*

Inference using the Laplace approximation has also been applied in [16] for the standard model. The general solution to the approximation problem can be found by maximizing the unnormalized log-posterior $\psi\left(\bar{\mathbf{f}}|\mathcal{Y},\mathcal{X},\bar{\mathbf{X}},\boldsymbol{\theta}\right) = \log p\left(\mathcal{Y}|\bar{\mathbf{f}},\mathcal{X},\bar{\mathbf{X}},\boldsymbol{\theta}\right) - \frac{1}{2}\bar{\mathbf{f}}^T\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1}\bar{\mathbf{f}} - \frac{1}{2}\log|\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}| - \frac{l}{2}\log 2\pi$ with regards to $\bar{\mathbf{f}}$. For the maximization we use a damped Newton method in which the damped step (with adaptive damping factor $\lambda$) can be calculated without inversion of the Hessian

$$\bar{\mathbf{f}}^{new} = \left(\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1} + \mathbf{W} - \lambda\mathbf{I}\right)^{-1} \left[\left(\mathbf{W} - \lambda\mathbf{I}\right)\bar{\mathbf{f}} + \nabla\log p(\mathcal{Y}|\bar{\mathbf{f}},\mathcal{X},\bar{\mathbf{X}},\boldsymbol{\theta})\right]. \quad (6)$$

Using the notation $\nabla\nabla_{i,j} = \frac{\partial^2}{\partial f(x_i)\partial f(x_j)}$ we apply the definition $\mathbf{W}_{i,j} = -\sum_k \nabla\nabla_{i,j}\log p(y_k|\mathbf{x}_{u_k},\mathbf{x}_{v_k},\bar{\mathbf{X}},\bar{\mathbf{f}},\boldsymbol{\theta})$. When converged, the resulting approximation can be shown to be $p\left(\bar{\mathbf{f}}|\mathcal{Y},\mathcal{X},\bar{\mathbf{X}},\boldsymbol{\theta}\right) \approx \mathcal{N}\left(\bar{\mathbf{f}}|\hat{\mathbf{f}},\left(\mathbf{W}+\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1}\right)^{-1}\right)$. The damped Newton step requires the Jacobian and Hessian of the new pseudo-input log-likelihood from Eq. (5), which require the following two derivatives

$$\frac{\partial}{\partial\bar{\mathbf{f}}}p\left(y_k|...\right) = y_k\frac{\mathcal{N}\left(z_k\right)}{\sigma_k^*\Phi\left(z_k\right)}\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1}\left(\mathbf{k}_{u_k}-\mathbf{k}_{v_k}\right) \quad (7)$$

$$\frac{\partial^2}{\partial\bar{\mathbf{f}}\bar{\mathbf{f}}^\top}p\left(y_k|...\right) = -y_k^2\frac{\mathcal{N}\left(z_k\right)}{(\sigma_k^*)^2\Phi\left(z_k\right)}\left[z_k+\frac{\mathcal{N}\left(z_k\right)}{\Phi\left(z_k\right)}\right] \cdot\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1}\left(\mathbf{k}_{u_k}-\mathbf{k}_{v_k}\right)\left(\mathbf{k}_{u_k}-\mathbf{k}_{v_k}\right)^\top\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1}. \quad (8)$$

#### *2.3.2. Evidence / Hyperparameter Optimization*

So far we have simply considered the hyperparameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathcal{L}},\boldsymbol{\theta}_{\mathcal{GP}}\}$ and pseudo inputs $\bar{\mathbf{X}}$ as fixed parameters, but their values have a crucial influence on the model performance. Here, we resort to point estimates and find (possible locally) optimal values by iterating between the Laplace approximation with fixed hyperparameters, i.e., finding $p\left(\bar{\mathbf{f}}|\mathcal{Y},\mathcal{X},\bar{\mathbf{X}},\boldsymbol{\theta}\right)$, followed by an evidence maximization step in which $(\boldsymbol{\theta},\bar{\mathbf{X}}) = \arg\max_{(\boldsymbol{\theta},\bar{\mathbf{X}})}p\left(\mathcal{Y}|\boldsymbol{\theta},\bar{\mathbf{X}}\right)$. The log-evidence $\log p(\mathcal{Y}|\boldsymbol{\theta},\bar{\mathbf{X}})$ has to be approximated in our case, which in terms of the existing Laplace approximation yields

$$\log p\left(\mathcal{Y}|\boldsymbol{\theta},\bar{\mathbf{X}}\right) \approx \log q\left(\mathcal{Y}|\bar{\mathbf{X}},\boldsymbol{\theta}\right) = \log p(\mathcal{Y}|\hat{\mathbf{f}},\bar{\mathbf{X}},\mathcal{X},\boldsymbol{\theta}) - \frac{1}{2}\hat{\mathbf{f}}^T\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\mathbf{I}+\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}\mathbf{W}|. \quad (9)$$

We further allow for fixed hyperpriors on the individual hyperparameters serving as regularization, which results in a procedure referenced to as MAP-II which provides more robust estimation. Consequently, the MAP-II is given by $\log q_{\text{MAP-II}}\left(\mathcal{Y}|\bar{\mathbf{X}},\boldsymbol{\theta}\right) = \log q\left(\mathcal{Y}|\bar{\mathbf{X}},\boldsymbol{\theta}\right) + \log p\left(\boldsymbol{\theta},\bar{\mathbf{X}}|\xi\right)$, where $\xi$ is a set of fixed parameters in the hyperprior.

The optimization requires the derivatives of the evidence approximation. These turn out to be rather tedious and involved, and we refer to the appendix for details. The pseudo-input model poses a number of difficulties since $\bar{\mathbf{X}}$ are also to be considered hyperparameters. Typically, this will—as noted by [7] and [17]—lead to a large number of local maxima providing potentially suboptimal solutions. It is not our aim to resolve nor document this issue, and we will take a pragmatic view and simply accept evidence optimization methods as is. Like [17] we recommend starting out with a fixed set of pseudo inputs initialized by a standard unsupervised clustering, such as k-means with restarts, followed by evidence optimization.

#### *2.3.3. Predictions*

The main task is to infer the latent function values $\bar{\mathbf{f}}$ with the end objective to make predictions of the observable variable $y$ for a pair of test inputs $\mathbf{x}_r \in \mathcal{X}_t$ and $\mathbf{x}_s \in \mathcal{X}_t$ denoted $\mathbf{x}_t = [\mathbf{x}_r,\mathbf{x}_s]^T$. We consider the joint distribution between $\bar{\mathbf{f}} \sim p\left(\bar{\mathbf{f}}|\mathcal{Y},\boldsymbol{\theta}\right)$ and the test variables $\mathbf{f}_t = [f\left(\mathbf{x}_r\right),f\left(\mathbf{x}_s\right)]^T$. With the posterior of $\bar{\mathbf{f}}$ approximated with the Gaussian from the Laplace approximation, the predictive distribution $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)$ will also be Gaussian given by $\mathcal{N}\left(\mathbf{f}_t|\mu^*,\mathbf{K}^*\right)$ with $\mu^* = [\mu_r^*,\mu_s^*]^T = \mathbf{k}_t\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}{}^{-1}\bar{\mathbf{f}}$ and

$$\mathbf{K}^* = \begin{bmatrix}\sigma_{rr}^* & \sigma_{rs}^*\\ \sigma_{sr}^* & \sigma_{ss}^*\end{bmatrix} = \mathbf{K}_t - \mathbf{k}_t^T\left(\mathbf{I}+\mathbf{W}\mathbf{K}_{\bar{\mathbf{X}}\bar{\mathbf{X}}}\right)\mathbf{k}_t,$$

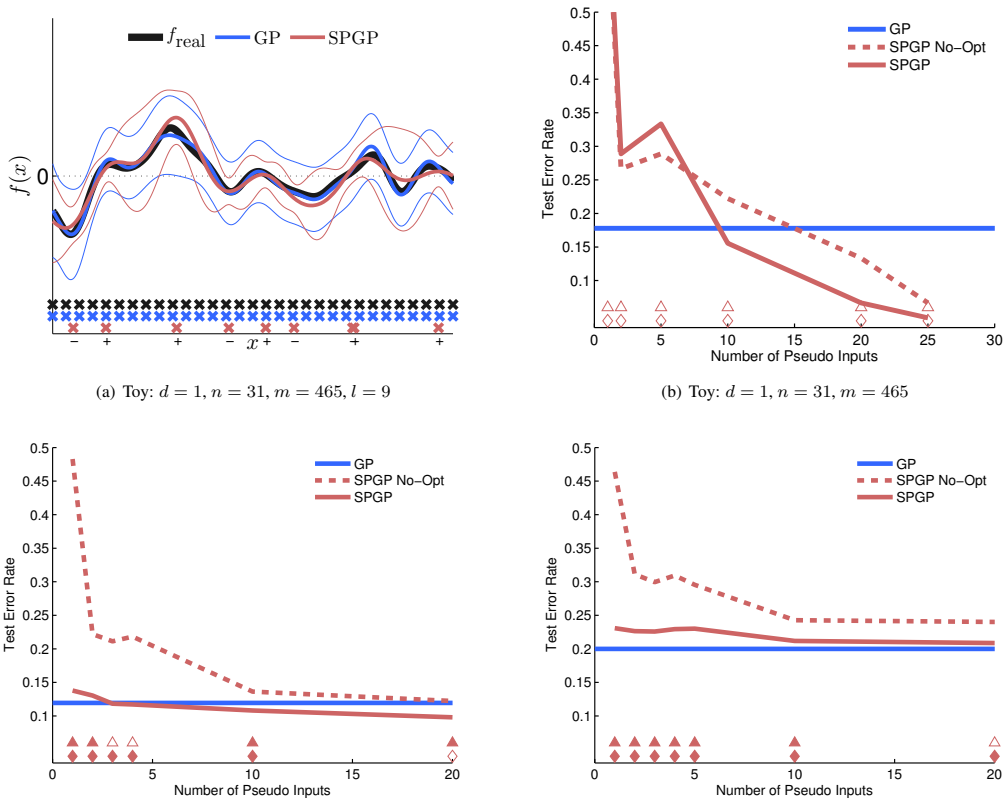
where $\mathbf{k}_t$ is the kernel between the test points and the pseudo inputs. With $p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)$, the prediction distribution of the observed variable is given as $p\left(y_t|\mathcal{Y},\boldsymbol{\theta}\right) = \int p\left(y_t|\mathbf{f}_t,\boldsymbol{\theta}_{\mathcal{L}}\right)p\left(\mathbf{f}_t|\mathcal{Y},\boldsymbol{\theta}\right)d\mathbf{f}_t$. The integral can be calculated in closed form as $P\left(\mathbf{x}_r \succ \mathbf{x}_s|\mathcal{Y},\boldsymbol{\theta}\right) = \Phi\left(\left(\mu_r^*-\mu_s^*\right)/\sigma^*\right)$ with $\left(\sigma^*\right)^2 = 2\sigma^2+\sigma_{rr}^*+\sigma_{ss}^*-\sigma_{rs}^*-\sigma_{sr}^*$.

## 3. SIMULATIONS & EXPERIMENTAL RESULTS

In this section we demonstrate the performance of the pseudo-input method on a toy example and provide predictive performance on two real-world data sets: *Boston housing* and *wine quality*. The main objective is not to achieve the overall best performance, but to compare the standard (GP) and the sparse (SPGP) formulations.

### 3.1. Toy Example

To illustrate the basics of the SPGP model, we draw a deterministic function $f_{\text{real}}$ (see Fig. 1(a)) from a zero-mean Gaus-




(a) Toy: $d = 1$, $n = 31$, $m = 465$, $l = 9$

(b) Toy: $d = 1$, $n = 31$, $m = 465$

(c) Boston Housing: $d = 10$, $n = 506$, $m = 127765$

(d) Wine Quality: $d = 11$, $n = 600$, $m = 179700$

**Fig. 1**. In general, blue graphs indicate the full model (GP) and red indicate the sparse model (SPGP). In **Fig. (a)** thick graphs indicate means and thin graphs indicate one standard deviation. The black graph indicates the real (deterministic) function used to generate the full pairwise data set between the instances marked with black crosses in the bottom. The two other colors sketch the predictive distribution of the GP and SPGP models using the (pseudo) inputs at the locations marked with the corresponding color in the bottom. **Fig. (b)-(d)** display the performance of the sparse model (SPGP) evaluated on the toy example and on the two real-world data sets as a function of the number of pseudo inputs for the sparse model (red). The performance of the standard model is included as a baseline. The **solid** and **dashed** red graphs show the average test error rate for the optimized and non-optimized SPGP model, respectively. The two rows of markers indicate whether the optimized (triangle) and non-optimized (diamond) SPGP models are significant different from the GP model using the McNemar test. The markers are solid if the null hypothesis that they are equal can be rejected at the 5% significance level.

sian process with a squared exponential covariance function. This function is then used to generate a pairwise data set consisting of all possible pairwise comparisons using the function values at equidistantly distributed locations marked with black crosses in Fig. 1(a). To model this data, we consider the two models: The GP model (Sec. 2.1) and the SPGP model with optimized pseudo inputs (Sec. 2.2). The $l = 9$ pseudo inputs are initialized equidistantly in the input interval, the length scale of the covariance function $\boldsymbol{\theta}_{\mathcal{GP}} = \{\sigma_\ell\}$ and the likelihood parameter $\boldsymbol{\theta}_{\mathcal{L}} = \{\sigma\}$ are learned by evidence optimization whereas $\sigma_f = 1$ of the covariance function is fixed. The results are presented in Fig. 1(a).

We notice that the SPGP model is capable of modeling the mean and thereby the actual pairwise relationships, whereas the predictive variance differs significantly from the GP variance. This is a characteristic and expected artifact also seen

in connection with the pseudo-input models for standard classification and regression.

### 3.2. Real World Examples

We compare the performance of the SPGP model to the GP model on two different real-world data sets.

The first data set is the well-known *Boston housing*[2] where we have constructed a full pairwise version by using all $m = 127765$ pairwise combinations of the $n = 506$ inputs base on the house price. For each input we use all available features except RAD, CHAS and NOX, thus $d = 10$.

The second data set is a subset of the *wine quality*[3] which is based on user ratings of wines. The subset is based on $n = 600$ instances of wines described by $d = 11$ features. We construct the set of unique pairwise comparisons from the ratings resulting in $m = 179700$ comparisons.

We use a squared exponential covariance function for both data sets which (based on initial experimentation) is initialized with $\sigma_f = 1$ and $\sigma_\ell = 1$. The covariance parameter $\sigma_f$ is fixed, whereas the likelihood parameter initialized as $\boldsymbol{\theta}_{\mathcal{L}} = \{\sigma = 1\}$ and $\boldsymbol{\theta}_{\mathcal{GP}} = \{\sigma_\ell\}$ are learned by MAP-II optimization using a uniform hyperprior and a half-student-t hyperprior with scale 6 and 4 degrees of freedom, respectively. Pseudo inputs are initialized with k-means (selecting the solution with minimum total squared distance out of five random initializations). We compare two SPGP models: one where the pseudo inputs are kept fixed following the k-means initialization (this model is identified with the No-Opt tag) and one where they are further fitted using MAP-II with a uniform hyperprior. With both data sets we use 20-fold cross validation on instances, such that a minimum of two instances are held out for testing and a randomly selected quarter of all remaining pairwise comparisons between training instances are used for training. Consequently, predictions are only performed on comparisons between instances that do not appear in the training data and the setting is thus a true predictive ranking scenario. In Fig. 1(c)-(d) we report the average error rate on the test set as a function of the number of pseudo inputs for the two SPGP models. The GP model is included as a baseline.

## 4. DISCUSSION

In the toy example (Fig. 1(a)) we see that the mean is well modeled by both the GP model and the SPGP model with $l = 9$ pseudo inputs, suggesting that the SPGP model performs nearly as good as the GP model. The main difference between the two models seems to be the predictive variance which differs significantly, yet this is an expected property of the sparse model. A way to improve the estimation of the predictive variance is by allowing the input instances and pseudo inputs to have different length scales [8][17].

Focusing on the predictive mean performance of the optimized SPGP model on the two real-world data sets (Fig. 1(c)-(d)), we see that a SPGP model with few pseudo inputs (as low as 1-5) performs only slightly worse than or equal to the GP model. This indicates that the two real-world problems do not constitute very complex pairwise problems. The performance is, however, highly dependent on the optimization of the locations of the pseudo inputs, seen since the non-optimized SPGP model requires more pseudo inputs due to the fixed locations. This illustrates the importance and power of the optimization.

By further adding pseudo inputs we can obtain better performance than the GP model. We believe that two effects come into play. The first effect is that the constraints induced in the SPGP model provide better regularization compared to the full Gaussian process prior meaning that it generalizes better. The second effect stems from the fact that the arbitrary placement of the pseudo inputs provides added flexibility, which effectively renders it more adequate for capturing the important regions of the underlying function when these locations are optimized appropriately. We speculate that the observed behavior is a combination of the two effects of course dependent on the application.

A further aspect to be investigated is the capability of the SPGP model to capture and approximate higher order moments of the predictive distribution. In line with previous work on the topic and with the variances observed in the toy example, we have observed fluctuating behavior of the predictive likelihoods as a function of $l$ for the SPGP models in the two real-world examples. Whether the behavior is due to the pairwise setting, specific application or a general property of the pseudo-input formulation is an open question.

In the current sparse formulation the original function values are dependent in pairs given the exact comparisons, whereas in FI(T)C all the original function values are independent given the pseudo inputs. We plan to investigate if this difference have any practical importance and to compare the current approximation to other traditional approaches—in particular the PI(T)C approximation.

## 5. CONCLUSION

In this paper we have derived a sparse version of the pairwise likelihood model using the pseudo-input formulation. We applied the Laplace approximation for both posterior and evidence approximation. We observe competitive predictive performance with the sparse model using only few pseudo inputs on a toy example and on two real-world data sets. A noticeable observation is the fact that we by adding more pseudo inputs are able to obtain better performance than the full GP model in the studied applications.

[2] `archive.ics.uci.edu/ml/datasets/Housing`
[3] `archive.ics.uci.edu/ml/datasets/Wine+Quality`

## 6. REFERENCES


[1] J. Fürnkranz and E. Hüllermeier, *Preference Learning*, Springer, 1st edition, 2011.

[2] B. McFee and G. Lanckriet, "Metric Learning to Rank," *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pp. 775–782, 2010.

[3] L. L. Thurstone, "A Law of Comparative Judgement," *Psychological Review*, vol. 34, 1927.

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

[5] N. Lawrence, M. Seeger, and R. Herbrich, "Fast Sparse Gaussian Process Methods: The Informative Vector Machine," in *Neural Information Processing Systems (NIPS)*, 2002, p. 8.

[6] L. Csato, *Gaussian Processes - Iterative Sparse Approximations*, Ph.D. thesis, Aston University, 2002.

[7] E. Snelson and Z. Ghahramani, "Sparse Gaussian Processes using Pseudo-Inputs," *Advances in neural information processing*, 2006.

[8] C. Walder, K. I. Kim, and B. Schölkopf, "Sparse Multiscale Gaussian Process Regression," in *Proceedings of the 25th international conference on Machine Learning*, 2008, pp. 1112–1119.

[9] M. Lazaro-Gredilla and A. Figueiras-Vidal, "Inter-Domain Gaussian Processes for Sparse Inference using Inducing Features," in *Advances in Neural Information Processing Systems 22*, pp. 1087–1095. 2009.

[10] J. Quiñonero-Candela and C.E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[11] E. Snelson and Z. Ghahramani, "Local and Global Sparse Gaussian Process Approximations," in *Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS*, 2007, pp. 524–531.

[12] J. Guiver and E. Snelson, "Learning to Rank with Softrank and Gaussian Processes," *Annual ACM Conference on Research and Development in Information Retrieval*, pp. 259–266, 2008.

[13] R. D. Bock and J. V. Jones, "The Measurement and Prediction of Judgment and Choice," 1968.

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

[15] E. Bonilla, S. Guo, and S. Sanner, "Gaussian Process Preference Elicitation," in *Advances in Neural Information Processing Systems 23*.

[16] W. Chu and Z. Ghahramani, "Extensions of Gaussian Processes for Ranking: Semi-Supervised and Active Learning," in *Workshop Learning to Rank at Advances in Neural Information Processing Systems 18*, 2005.

[17] Y. Qi, A. Abdel-Gawad, and T. Minka, "Sparse-Posterior Gaussian Processes for General Likelihoods," in *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, 2010.


## 7. APPENDIX - EVIDENCE DERIVATIVES

The derivatives of Eq. (9) are slightly different compared to the standard classification case [14, Sec 5.5.1] due to the pseudo-input model because the covariance parameters enter into the likelihood, and the fact that the covariance function also depends on $\bar{\mathbf{X}}$. We outline the derivations by noting that the Eq. (9) depends both explicitly and implicitly (due to the solution of $\hat{\mathbf{f}}$) on the parameters $\boldsymbol{\theta}$. We do not differentiate between likelihood and covariance parameters and $\bar{\mathbf{X}}$. Here, we simply denote a parameter by $\theta_j$. We can split the derivatives into an explicit and implicit part

$$\frac{\partial \log q\left(\mathcal{Y}|...\right)}{\partial \theta_i} = \left.\frac{\partial \log q\left(\mathcal{Y}|...\right)}{\partial \theta}\right|_{\text{explicit}} + \sum_j \frac{\partial \log q\left(\mathcal{Y}|...\right)}{\partial f_j}\frac{\partial f_j}{\partial \theta_i}.$$

Referring to the **explicit** term we obtain the following terms

$$\frac{\partial}{\partial \boldsymbol{\theta}_i}\log p\left(\mathcal{Y}|\hat{\mathbf{f}},\boldsymbol{\theta}\right)$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i}\hat{\mathbf{f}}^\top \mathbf{K}_{\boldsymbol{\theta}}{}^{-1}\hat{\mathbf{f}} = -\hat{\mathbf{f}}^\top\left(\mathbf{K}_{\boldsymbol{\theta}}{}^{-1}\frac{\partial \mathbf{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\mathbf{K}_{\boldsymbol{\theta}}{}^{-1}\right)\hat{\mathbf{f}}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i}\log|\mathbf{I}+\mathbf{W}_{\boldsymbol{\theta}}\mathbf{K}_{\boldsymbol{\theta}}| = Tr\left[(\mathbf{I}+\mathbf{K}_{\boldsymbol{\theta}}\mathbf{W}_{\boldsymbol{\theta}})^{-1}\cdot \left(\frac{\partial \mathbf{W}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\mathbf{K}_{\boldsymbol{\theta}} + \mathbf{W}_{\boldsymbol{\theta}}\frac{\partial \mathbf{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\right)\right]$$

Referring to the **implicit** term we have (without any assumptions regarding the type of parameter)

$$\frac{\partial \log q\left(\mathcal{Y}|\bar{\mathbf{X}},\mathcal{X},\boldsymbol{\theta}\right)}{\partial f_j} = -\frac{1}{2}Tr\left[(\mathbf{I}+\mathbf{K}_{\boldsymbol{\theta}}\mathbf{W}_{\boldsymbol{\theta}})^{-1}\left(\mathbf{K}_{\boldsymbol{\theta}}\frac{\partial \mathbf{W}_{\boldsymbol{\theta}}}{\partial f_j}\right)\right]$$

$\frac{\partial f_j}{\partial \boldsymbol{\theta}_i}$ is found by exploiting that $\hat{\mathbf{f}} = \mathbf{K}_{\boldsymbol{\theta}}\nabla \log p\left(\mathcal{Y}|\hat{\mathbf{f}},\boldsymbol{\theta}\right)$ at the current solution leading to the following result

$$\frac{\partial f_j}{\partial \boldsymbol{\theta}_i} = (\mathbf{I}+\mathbf{K}_{\boldsymbol{\theta}}\mathbf{W}_{\boldsymbol{\theta}})^{-1}\left(\frac{\partial \mathbf{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\right)\frac{\partial \log p\left(y|\hat{\mathbf{f}},\boldsymbol{\theta}\right)}{\partial \mathbf{f}} + (\mathbf{I}+\mathbf{K}_{\boldsymbol{\theta}}\mathbf{W}_{\boldsymbol{\theta}})^{-1}\mathbf{K}_{\boldsymbol{\theta}}\frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\frac{\partial \log p\left(y|\hat{\mathbf{f}},\boldsymbol{\theta}\right)}{\partial \mathbf{f}}\right)$$

We may exploit that the inverse of the common factor $(\mathbf{I}+\mathbf{K}_{\boldsymbol{\theta}}\mathbf{W}_{\boldsymbol{\theta}})$ can be computed using the Cholesky decomposition which enters robustly into the individual expressions for added numerical stability. The expression above is a general result and valid for both likelihood parameters, covariance parameters and pseudo inputs. In addition, the derivatives of the likelihood, Jacobian, Hessian and covariance function are required. One should be aware that some of the derivatives are zero depending on the actual parameter type (e.g. $\partial \mathbf{K}_{\boldsymbol{\theta}}/\partial \boldsymbol{\theta}_{\mathcal{L}}$). The gradients are based on the current Laplace approximation. Even though we take into account implicit dependencies, there is in general no guarantee for strictly monotonic behavior, thus a robust optimization method is required. In practice we have found the BFGS implementation in the `immoptibox`[4] robust.

[4] `www2.imm.dtu.dk/%7Ehbn/immoptibox/`

Appendix H

# Towards a Universal Representation for Audio Information Retrieval and Analysis

---

Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen and Lars Kai Hansen. Towards a Universal Representation for Audio Information Retrieval and Analysis. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2013. Published.

# TOWARDS A UNIVERSAL REPRESENTATION FOR AUDIO INFORMATION RETRIEVAL AND ANALYSIS

*Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen, and Lars Kai Hansen*

DTU Compute
Technical University of Denmark
Asmussens Allé B305, 2800 Kgs. Lyngby, Denmark
{bjje,rast,janla,lkai}@dtu.dk

## ABSTRACT

A fundamental and general representation of audio and music which integrates multi-modal data sources is important for both application and basic research purposes. In this paper we address this challenge by proposing a multi-modal version of the Latent Dirichlet Allocation model which provides a joint latent representation. We evaluate this representation on the Million Song Dataset by integrating three fundamentally different modalities, namely tags, lyrics, and audio features. We show how the resulting representation is aligned with common 'cognitive' variables such as tags, and provide some evidence for the common assumption that genres form an acceptable categorization when evaluating latent representations of music. We furthermore quantify the model by its predictive performance in terms of genre and style, providing benchmark results for the Million Song Dataset.

***Index Terms***— Audio representation, multi-modal LDA, Million Song Dataset, genre classification.

## 1. INTRODUCTION

Music representation and information retrieval are issues of great theoretical and practical importance. The theoretical interest relates in part to the close interplay between audio, human cognition and sociality, leading to heterogenous and highly multi-modal representations in music. The practical importance, on the other hand, is evident as current music business models suffer from the lack of efficient and user friendly navigation tools. We are interested in representations that directly support interactivity, thus representations based on latent variables that are well-aligned with cognitively (semantic) relevant variables [1]. User generated tags can be seen as such 'cognitive variables' since they represent decisions that express reflections on music content and context. Clearly, such tags are often extremely heterogenous, high-dimensional, and idiosyncratic as they may relate to any aspect of music use and understanding.

Moving towards broadly applicable and cognitively relevant representations of music data is clearly contingent on the ability to handle multi-modality. This is reflected in current music information research that use a large variety of representations and models, ranging from support vector machine (SVM) genre classifiers [2]; custom latent variable models models for tagging [3]; similarity based methods for recommendation based on Gaussian Mixture models [4]; and latent variable models for hybrid recommendation [5]. A significant step in the direction of flexible multi-modal representations was taken in the work of Law *et al.* [6] based on the probabilistic framework of Latent Dirichlet Allocation (LDA) topic modeling. Their topic model representation of tags allows capturing rich cognitive semantics as users are able to tag freely without being constrained by a fixed vocabulary. However, with a strong focus on automatic tagging Law *et al.* refrained from developing a universal representation - symmetric with respect to all modalities. A more symmetric representation is pursued in recent work by Weston *et al.* [7]; however, without a formal statistical framework it offers less flexibility, e.g., in relation to handling missing features or modalities. This is often a challenge encountered in real world music applications.

In this work we pursue a multi-modal view towards a unifying representation, focusing on latent representations informed symmetrically by all modalities based on a multi-modal version of the Latent Dirichlet Allocation model. In order to quantify the approach, we evaluate the model and representation in a large-scale setting using the million song dataset (MSD) [8], and consider a number of models trained on combinations of the three basic modalities: user tags (top-down view), lyrics (meta-data view) and content based audio features (bottom-up view). First, we show that the latent representation obtained by considering the audio and lyrics modalities is well aligned—in an unsupervised manner - with 'cognitive' variables by analyzing the mutual information

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. Bob L. Sturm, Aalborg University Copenhagen is acknowledged for suggestion of relevant references in music interpretation.

between the user generated tags and the representation itself. Secondly, with knowledge obtained in the first step, we evaluate auxiliary predictive tasks to demonstrate the predictive alignment of the latent representation with well-known human categories and metadata information. In particular we consider genre and styles provided by [9], none of which is used to learn the latent semantics themselves. This leads to benchmark results on the MSD and provides insight into the nature of generative genre and style classifiers.

Our work is related to a rich body of studies in music modeling, and multi-modal integration. In terms of non-probabilistic approaches this includes the already mentioned work of Weston *et al.* [7]. McFee *et al.* [10] showed how hypergraphs (see also [11]) can be used to combine multiple modalities with the possibilities to learn the importance of each modality for a particular task. Recently McVicar *et al.* [12] applied multi-way CCA to analyze emotional aspects of music based on the MSD.

In the topic modelling domain, Arenas-García *et al.* [13] proposed multi-modal PLSA as a way to integrate multiple descriptors of similarity such as genre and low-level audio features. Yoshii *et al.* [5, 14] suggested a similar approach for hybrid music recommendation integrating subject taste and timbre features. In [15], standard LDA was applied with audio words for the task of obtaining low-dimensional features (topic distributions) applied in a discriminative SVM classifier. For the particular task of genre classification *et al.* [16] applied the pLSA model as a generative genre classifier. Our work is a generalization and extension of these previous ideas and contributions based on the multi-modal LDA, multiple audio features, audio words and a generative classification view.

## 2. DATA & REPRESENTATION

The recently published million song dataset (MSD) [8] has highlighted some of the challenges in modern music information retrieval; and made it possible to evaluate top-down and bottom-up integration of data sources on a large scale. Hence, we naturally use the MSD and associated data sets to evaluate the merits of our approach. In defining the latent semantic representation, we integrate the following modalities/data sources.

The tags, or top-down features, are human annotations from `last.fm` often conveying information about genre and year of release. Since users have consciously annotated the music in an open vocabulary, such tags are considered an expressed view of the users cognitive representation. The metadata level, i.e., the lyrics, is of course nonexistent for for majority of certain genres, and in other cases simply missing for individual songs which is not a problem for the proposed model. The lyrics are represented in a *bag-of-words* style, i.e., no information about the order in which the terms occurs is included. The content based or bottom up features are derived from the audio itself. We rely on the Echonest feature extraction[1] already available in for the MSD, namely timbre, chroma, loudness, and tempo. These are orginally derived in event related segments, but we follow previous work [17] by beat aligning all features obtaining an meaningful alignment with music related aspects.

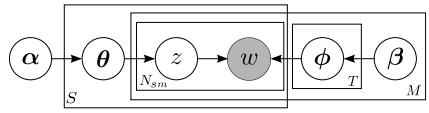



**Fig. 1**: Graphical model of the multi-modal LDA model

In order to allow for practical and efficient indexing and representation, we abandon the classic representation of using for example a Gaussian mixture model for representing each song in its respective feature space. Instead we turn to the so-called *audio word* approach (see e.g. [18, 19, 3, 17]) where each song is represented by a vector of counts of (finite) number of audio words (feature vector). We obtain these *audio words* by running a randomly initiated K-means algorithm on a 5% random subset of the MSD for timbre, chroma, loudness and tempo with 1024, 1024, 32, and 32 clusters, respectively. All beat segments in a all songs are then quantized into these audio words and the resulting counts, representing the four different audio features, are concatenated to yield the audio modality.

## 3. MULTI-MODAL MODEL

In order to model the heterogeneous modalities outline above, we turn to the framework of topic modeling. We propose to use a multi-modal modification of the standard LDA to represent the latent representation in a symmetric way relevant to many music applications. The multi-modal LDA, mmLDA, [20] is a straight forward extension of standard LDA topic model [21], as shown in Fig. 1. The model and notation is easily understood by the way it generates a new song by the different modalities, thus the following generative process defines the model:

- For each topic $z \in [1;T]$ in each modality $m \in [1;M]$
  Draw $\boldsymbol{\phi}_z^{(m)} \sim \textit{Dirichlet}(\boldsymbol{\beta}^{(m)})$.
  This is the parameters of the $z^{th}$ topic's distribution over vocabulary $[1;V^{(m)}]$ of modality $m$.
- For each song $s \in [1;S]$
  - Draw $\boldsymbol{\theta}_s \sim \textit{Dirichlet}(\boldsymbol{\alpha})$.
    This is the parameters of the $s^{th}$ song's distribution over topics $[1;T]$.
  - For each modality $m \in [1;M]$
    - For each word $w \in [1;N_{sm}]$
      - Draw a specific topic $z^{(m)} \sim \textit{Categorical}(\boldsymbol{\theta}_s)$

[1] `http://the.echonest.com`

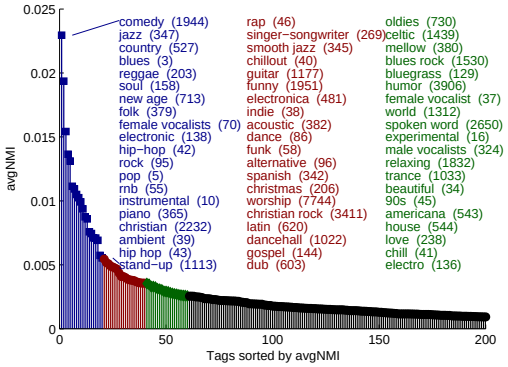
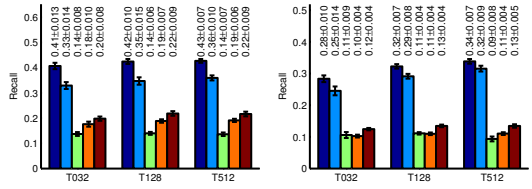



**Fig. 2**: Normalized average mutual information (avgNMI) between the latent representation defined by audio and lyrics for $T = 128$ topics and the 200 top-ranked tags. avgNMI is computed on the test set in each fold. The popularity of each tag is indicated in parenthesis.




(a) Genre (b) Style

**Fig. 3**: Classification accuracy for $T \in \{32, 128, 512\}$. Dark blue: Combined model; Light Blue: Tags; Green: Lyrics; Orange: Audio; Red: Audio+Lyrics.

· Draw a word $w^{(m)} \sim \textit{Categorical}(\boldsymbol{\phi}^{(m)}_{z^{(m)}})$

A main characteristic of mmLDA is the common topic proportions for all $M$ modalities in each song, $s$, and separate word-topic distributions $p(w^{(m)}|z)$ for each modality, where $z$ denotes a particular topic. Thus, each modality has its own definition of what a topic is in terms of its own vocabulary.

Model inference is performed using a collapsed Gibbs sampler [22] similar to the standard LDA. The Gibbs sampler is run for a limited number of complete sweeps through the training songs, and the model state with the highest model evidence within the last 50 iterations is regarded as the MAP estimate. From this MAP sample, point estimates of the topic-song distribution, $\hat{p}(z|s)$, and the modality, $m$, specific word-topic distribution, $\hat{p}(w^{(m)}|z)$, can be computed based on the expectations of the corresponding Dirichlet distributions.

Evaluation of model performance on a unknown test song, $s^*$, is performed using the procedure of fold-in [23, 24] by computing the point estimate of the topic distribution, $\hat{p}(z|s^*)$ for the new song, by keeping the all the word-topic counts fixed during a number of new Gibbs sweeps. Testing on a modality, not included in the training phase, requires a point estimate of the word-topic distribution, $p(w^{(m^*)}|z)$, of the held out modality, $m^*$, of the training data. This is obtained by fixing the song-topic counts while updating the word-topic counts for that specific modality. This is similar to the fold-in procedure used for test songs.

## 4. EXPERIMENTAL RESULTS & DISCUSSION

### 4.1. Alignment

The first aim is to evaluate the latent representation's alignment with a human 'cognitive' variable, which we previously argued could be the open vocabulary tags. We do this by including only the lower level modalities of audio and lyrics when estimating the model. Then the normalized mutual information between a single tag and the latent representations, i.e., the topics, is calculated for all the tags.

Thus for a single tag, ${w_i}^{(tag)}$ we can compute the mutual information between the tag and the topic distribution for a specific song, $s$ as:

$$\mathrm{MI}\left({w_i}^{(tag)}, z|s\right) = \mathrm{KL}\left(\hat{p}\left({w_i}^{(tag)}, z|s\right) || \hat{p}\left({w_i}^{(tag)}|s\right)\hat{p}\left(z|s\right)\right), \quad (1)$$

where $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler divergence. We normalize the MI to be in $[0; 1]$, i.e,

$$\mathrm{NMI}\left({w_i}^{(tag)}, z|s\right) = 2\frac{\mathrm{MI}\left({w_i}^{(tag)}, z|s\right)}{H\left({w_i}^{(tag)}|s\right) + H\left(z|s\right)},$$

where $H(\cdot)$ denotes the entropy. Finally, we compute the average over all songs to arrive at the final measure of alignment for a specific tag, given by $\mathrm{avgNMI}({w_i}^{(tag)}) = \frac{1}{S}\sum_s \mathrm{NMI}\left({w_i}^{(tag)}, z|s\right)$.

Fig. 2 shows a sorted list of tags, where tags with high alignment with the latent representation have higher average NMI (avgNMI). It is notable that the combination of the audio and lyrics modality, in defining the latent representation, seems to align well with genre-like and style-like tags. On the contrary, emotional and period tags are relatively less aligned with the representation. Also note that the alignment is not simply a matter of the tag being the most popular as can be seen from Fig. 2. Less popular tags are ranked higher by avgNMI than very popular tags, suggesting that some are more specialized in terms of the latent representation than others.

The result gives merit to the idea of using genre and styles as proxy for evaluating latent representation in comparison with other open vocabulary tags, since we - from lower level features, such as audio features and lyrics - can find latent representations which align well with high-level, 'cognitive' aspects in an unsupervised way. This is in line with many

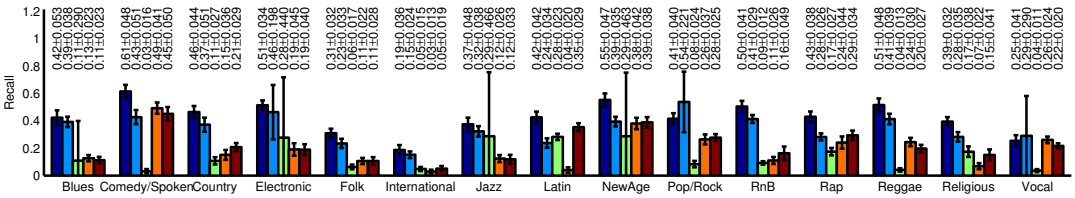
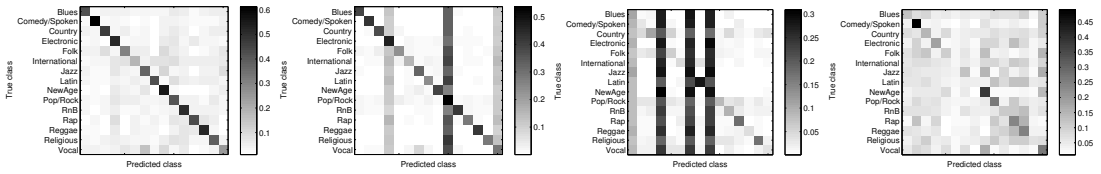



**Fig. 4**: Dark blue: Combined model, Light Blue: Tags, Green: Lyrics, Orange: Audio, Red: Audio+Lyrics, genre, $T = 128$.




(a) Combined Model (b) Tag Model (c) Lyrics Model (d) Audio Model

**Fig. 5**: Confusion matrices for genre and 128 topics. The color level indicates the classification accuracy.

studies in music informatics on western music (see e.g. [25, 26, 27]) which indicate coherence between genre and tag categories and cognitive understanding of music structure. In summary, the ranking of tag alignment using our modeling approach on the MSD provides some evidence in favor of such coherence.

### 4.2. Prediction

Given the evidence presented for genre and style being the relatively most appropriate human categories, our second aim is to evaluate the predictive performance of the multi-modal model for genre and style, and we turn to the recently published extension of the MSD [9] for reference test/train splits and genre and style labels. In particular, we use the balanced splits defined in [9].

For the genre case, this results in 2000 labeled examples per genre and 15 genres, thus resulting in $30,000$ songs. We estimate the predictive genre performance by 10-fold cross-validation. Fig. 4 shows the per-label classification accuracy (perfect classification equals 1). The total genre classification performance is illustrated in Fig. 3a. The corresponding result for style classification, based on a total of $50,000$ labeled examples, is shown in Fig. 3b. Both results are generated using $T = 128$ topics, 2000 Gibbs sweeps and predicting using the MAP estimate from the Gibbs sampler.

We first note that the combination of all modalities performs the best and significantly better than random as seen from Fig. 3, which is encouraging, and support the multi-modal approach. It is furthermore noted that the tag modality is able to perform very well. This indicates that despite the possibly noisy user expressed view, the model is able to find structure in line with the taxonomy defined in the reference labels of [9]. More interesting is perhaps the audio and lyric modalities and the combination of the two. This shows that lyrics performs the worst for genre, possibly due to the missing data in some tracks, while the combination is significantly better. For style there is no significant difference between audio and lyrics.

Looking at the genre specific performance in Fig. 4 we find a significant difference between the modalities. It appears that the importance of the modalities is partly in line with the fundamentally different characteristics of each specific genre. For example 'latin' is driven by very characteristic lyrics. Further insight can be obtained by considering the confusion matrices which show some systematic pattern of error in the individual modalities, whereas the combined model shows a distinct diagonal structure, highlighting the benefits of multi-modal integration.

## 5. CONCLUSION

In this paper, we proposed the multi-way LDA as a flexible model for analyzing and modeling multi-modal and heterogeneous music data in a large scale setting. Based on the analysis of tags and latent representation, we provided evidence for the common assumption that genre may be an acceptable proxy for cognitive categorization of (western) music. Finally, we demonstrated and analyzed the predictive performance of the generative model providing benchmark result for the Million Song Dataset, and a genre dependent performance was observed. In our current research, we are looking at purely supervised topic models trained for, e.g. genre prediction. In order to address truly multi-modal and multi-task scenarios such as [7], we are currently pursuing an extended probabilistic framework that include correlated topic models [28], multi-task models [29], and non-parametric priors [30].

## 6. REFERENCES


[1] L.K. Hansen, P. Ahrendt, and J. Larsen, "Towards cognitive component analysis," in *AKRR05-International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

[2] C. Xu, N.C. Maddage, and X. Shao, "Musical genre classification using support vector machines," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 429–432, 2003.

[3] M. Hoffman, D. Blei, and P. Cook, "Easy as CBA: A simple probabilistic model for tagging music," *Proc. of ISMIR*, pp. 369–374, 2009.

[4] F. Pachet and J.J. Aucouturier, "Improving timbre similarity: How high is the sky?," *Journal of negative results in speech and audio*, pp. 1–13, 2004.

[5] Y. Kazuyoshi, M. Goto, K. Komatani, R. Ogata, and H.G. Okuno, "Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR*, 2006, pp. 296–301.

[6] E. Law, B. Settles, and T. Mitchell, "Learning to tag from open vocabulary labels," *Machine Learning and Knowledge Discovery in Databases*, pp. 211–226, 2010.

[7] J. Weston, S. Bengio, and P. Hamel, "Multi-Tasking with Joint Semantic Spaces for Large- Scale Music Annotation and Retrieval Multi-Tasking with Joint Semantic Spaces for Large-Scale Music Annotation and Retrieval," *Journal of New Music Research*, , no. November 2012, pp. 37–41, 2011.

[8] T. Bertin-Mahieux, D.P.W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[9] A. Schindler, R. Mayer, and A. Rauber, "Facilitating comprehensive benchmarking experiments on the million song dataset," in *13th International Conference on Music Information Retrieval (ISMIR 2012)*. 2012.

[10] B. McFee and G. R. G. Lanckriet, "Hypergraph models of playlist dialects," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Fabien Gouyon, Perfecto Herrera, Luis Gustavo Martins, and Meinard Müller, Eds. 2012, pp. 343–348, FEUP Edições.

[11] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music Recommendation by Unified Hypergraph: Combining Social Media Information and Music Content," pp. 391–400, 2010.

[12] M. Mcvicar and T. de Bie, "CCA and a Multi-way Extension for Investigating Common Components between Audio , Lyrics and Tags .," in *CMMR*, 2012, number June, pp. 19–22.

[13] J. Arenas-García, A. Meng, K.B. Petersen, T. Lehn-Schiøler, L.K. Hansen, and J. Larsen, *Unveiling Music Structure Via PLSA Similarity Fusion*, pp. 419–424, IEEE, 2007.

[14] K. Yoshii and M. Goto, "Continuous pLSI and smoothing techniques for hybrid music recommendation," *International Society for Music Information Retrieval Conference*, pp. 339–344, 2009.

[15] S. K., S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," pp. 2–5, 2009.

[16] Zhi Zeng, Shuwu Zhang, Heping Li, W. Liang, and Haibo Zheng, "A novel approach to musical genre classification using probabilistic latent semantic analysis model," in *IEEE International Conference on Multimedia and Expo (ICME), 2009*, 2009, pp. 486–489.

[17] T. Bertin-Mahieux, "Clustering beat-chroma patterns in a large music database," in *International Society for Music Information Retrieval Conference*, 2010.

[18] Y. Cho and L.K. Saul, "Learning dictionaries of stable autoregressive models for audio scene analysis," *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pp. 1–8, 2009.

[19] K. Seyerlehner, G. Widmer, and P. Knees, "Frame level audio similarity-a codebook approach," *Conference on Digital Audio Effects*, pp. 1–8, 2008.

[20] D.M. Blei and M.I. Jordan, "Modeling annotated data," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, 2003.

[21] D. M. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[22] T.L. Griffiths and M. Steyvers, "Finding scientific topics.," *Proceedings of the National Academy of Sciences of the United States of America*, pp. 5228–35, Apr. 2004.

[23] H.M. Wallach, I. Murray, Ruslan Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, , no. d, pp. 1–8, 2009.

[24] T. Hofmann, "Probabilistic latent semantic analysis," *Proc. of Uncertainty in Artificial Intelligence, UAI*, p. 21, 1999.

[25] J.H. Lee and J..S Downie, "Survey of music information needs, uses, and seeking behaviours: Preliminary findings," in *Proc. of ISMIR*, 2004, pp. 441–446.

[26] J. Frow, *Genre*, Routledge, New York, NY, USA, 2005.

[27] E. Law, "Human computation for music classification," in *Music Data Mining*, T. Li, M. Ogihara, and G. Tzanetakis, Eds., pp. 281–301. CRC Press, 2011.

[28] S. Virtanen, Y. Jia, A. Klami, and T. Darrell, "Factorized Multi-Modal Topic Model," *auai.org*, 2010.

[29] A. Faisal, J. Gillberg, J. Peltonen, G. Leen, and S. Kaski, "Sparse Nonparametric Topic Model for Transfer Learning," *dice.ucl.ac.be*.

[30] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, 2004.

Appendix I

# Personalized Audio System - a Bayesian Approach

---

Jens Brehm Nielsen, Bjørn Sand Jensen, Toke Jansen Hansen and Jan Larsen, Personalized Audio System - a Bayesian Approach, 135th AES Convention, 2013.

Note: Only a pre-print of the published paper is included, the published version can be found via the publisher, http://www.aes.org/e-lib/browse.cfm?elib=17048.

# 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 users listening to the system. To obtain the best possible setting, the user is forced into multi-parameter optimization with respect to the users's own objective and preference. To address this, the present paper presents a general inter-active framework for personalization of such audio systems. The framework builds on Bayesian Gaussian process regression in which a model of the users's *objective function* is updated sequentially. The parameter setting to be evaluated in a given trial is selected by model-based sequential experimental design. A Gaussian process model is proposed which incorporates correlation among particular parameters providing better modeling capabilities compared to a standard model. A five-band equalizer is considered for demonstration purposes, in which the parameters are optimized using the proposed framework. Twelve test subjects obtain a personalized setting with the framework, and these settings are significantly preferred to those 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 systems 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 systems optimally. This is because users have to navigate in a high-dimensional parameter space, which makes it extremely difficult for users to find even a local optimum. 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

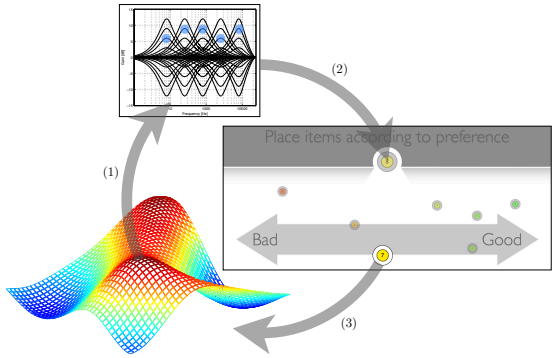



Figure 1: A conceptual overview of the interactive system. At step (1) we draw a new EQ from the current estimate of the user's objective function. Next, at step (2) this particular EQ is associated with a *ball*, in this case number *eight*, in the visualized user interface. Finally, after the user 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.

users's opinions concerning the adjustment of a particular system. It is therefore of great interest to find and evaluate fast and flexible tools for robustly optimizing user-adjustable parameters, with the aim to rapidly obtain a truly personalized audio system setting.

Prime examples of complex audio systems are hearing aids, where hundreds of parameters make up a unique and personal experience. It is therefore natural that this field has considered ways to learn an optimal setting based on preference (Kuk *et. al.* [8] and Baskent *et. al.* [1]), although these are currently based on non-probabilistic methods. Recently—and the closest related to our approach—Birlutiu *et. al.* [4, 3] have proposed two probabilistic approaches driven mainly by a *multi-task* formulation utilizing the information transfer among users, to learn a *complete* preference model accounting for *all* preference relations. For the purpose of optimizing parameters, it is not efficient to learn a complete model over a high-dimensional parameter space, because the model is only required to be accurate around possible optimal parameters.

In audio reproduction systems—like home entertainment and professional mixing equipment—preference learning approaches are relatively unknown, despite the clear evidence that personalization may be beneficial in for example equalization (Paterson [11] and Zhang *et. al.* [18]). Existing approaches such as Reed [13], Pardo *et. al.* [10], and Sabin *et. al.* [14], are based on non-probabilistic approaches, thus neglecting the highly stochastic nature of perceptual responses.

In this work we focus on audio reproduction systems and the outlined task of optimizing multiple parameters in such systems for an individual user listening to the output of the system. For this purpose, we propose and consider a combination of robust Bayesian modeling, an engaging user interface for user feedback and global optimization techniques (active learning) 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 [15]. The framework uses an intuitive and simple graphical user interface for obtaining user ratings, which allows the user to listen to previously rated settings thus serving as anchors/references for future ratings. In contrast to standard practice, we do however not only allow the user to listen to previous settings, but we also allow the user to change all the ratings of 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.

We furthermore consider the fact that certain parameters may be correlated with respect to the user's objective. An example could be the compression ratio, the attack time and the release time in a compressor. To exploit such correlation and obtain better modeling capabilities, we suggest a specific model which assumes correlation between specific input parameters.

To demonstrate the potential of the framework for personal audio system optimization, we use a five-band constant-Q equalizer (EQ) as the running example, because the parameters (gains) in an EQ is something that we (as professionals) more or less all can relate to. We are aware that any audio-engineering professional will probably be able to quickly tune the five parameters of the EQ to his own objective. However, this is actually not a very typical scenario. Typically, users of home entertainment systems are untrained, and thus, have very little intuition about the parameters that they have the opportunity to tune and close to no intuition at all about the interplay *between* parameters. Hence with for instance five parameters controlling a virtual surround sound system with virtual base enhancement, most users would seek an optimal setting using trial and error (random experimentation). This is the premises in which the EQ example should be considered and the EQ is just convenient for demonstration purposes.

Through model comparison, we first show that the model with assumed correlation between input parameters improves the modeling capabilities compared to a traditional GP model without assumed correlation. The analysis is performed on real-world data, where 21 test 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 (for the typically untrained user) without an efficient sequential design approach and correspondingly good modeling capabilities. 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 approach (Sec. 2.2). Secondly in Sec. 3.2, we show that the GP model extension provides better modeling capabilities 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 framework uses an interactive loop to discover the user's preferred setting of a particular audio system, 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 presents how a user's objective function over EQ settings is learned 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 [15]. 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 [12]. In effect, the GP is placed as a prior over any finite set of functional values $\mathbf{f} = [f_1, f_2, ..., f_n]^\top$, where $f_i = f(\mathbf{x}_i)$, resulting in a finite multivariate Gaussian distribution over the set as $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_{XX}})$, where each element of the covariance matrix $\mathbf{K_{XX}}$ is given by a covariance function $k(\cdot, \cdot)$ such that $[\mathbf{K_{XX}}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. 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 [12, Chapter 2]. 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,

$$
\begin{aligned}
\sigma_s|\theta_s &\sim \mathcal{U}(0, \infty) \\
\sigma_\ell|\theta_\ell &\sim \mathcal{U}(0, \infty) \\
\sigma|\theta_\ell &\sim \mathcal{U}(0, \infty) \\
f_i|\sigma_s, \sigma_\ell &\sim \mathcal{GP}\left(m\left(\mathbf{x}_i\right), \mathrm{k}\left(\mathbf{x}_i, \cdot\right)_{\sigma_s, \sigma_\ell}\right) \\
z_i|f_i &\sim \mathcal{N}\left(f_i, \sigma\right) \qquad (1) \\
z_i &= \Phi^{-1}\left(y_i\right), \qquad (2)
\end{aligned}
$$

where $\sigma_\ell$ is the length scale of the covariance function, $\sigma_s$ is the standard deviation of the latent function, and $\sigma$ is the noise standard deviation (in latent space). $\mathcal{U}(a, b)$ denotes a uniform hyper prior on the open interval from $a$ to $b$, i.e. an improper and *non-informative* prior. Alternatively, so-called *weakly-informative* hyper priors would apply—especially over the length scale $\sigma_\ell$—such as the half-student-t hyper prior [5, 16], which could be applied to provide a more robust inference and prediction scheme avoiding the GP model to fit hyperplanes with only few observations. We note that the observation noise, $\sigma$, 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 [12]

$$
\begin{aligned}
&k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\
&\quad = \sigma_s \exp\left(-\frac{1}{\sigma_\ell}(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{\Lambda}^{-1}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right). \qquad (3)
\end{aligned}
$$

In the first case, $\mathbf{\Lambda}$ is the identity matrix leading to the well-known (isotropic) squared exponential covariance function $k_{\text{iso}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_s \exp\left(-\frac{1}{\sigma_\ell}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$. In the second case, $\mathbf{\Lambda}$ is a general positive semi-definite matrix defining a correlation between parameters (input space) as explicit prior information. Here will denote this variant as the Mahalanobis covariance function[1], $k_{\text{mah}}(\mathbf{x}_i, \mathbf{x}_j)$ and set

$$\mathbf{\Lambda}_{\text{mah}} = \begin{bmatrix} 1 & 0.5 & 0.2 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0.2 & 0 \\ 0.2 & 0.5 & 1 & 0.5 & 0.2 \\ 0 & 0.2 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0.2 & 0.5 & 1 \end{bmatrix}. \tag{4}$$

The effect of the two options on the EQ example will be evaluated with reference to the standard case as `iso` and the Mahalanobis case as `mah`.

We turn to a standard GP inference scheme [12] in which the covariance and likelihood parameters, $\sigma_s, \sigma_\ell, \sigma$, 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 [15]. For the BFGS methods, the parameters are always initialized as $\sigma_s = 1, \sigma_\ell = 1, \sigma = 1$. The predictive mean and (co)variance of the latent function, $\mathbb{E}(\mathbf{f}^*)$ and $\mathbb{V}(\mathbf{f}^*)$, are given in standard form [12] as

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

$$\mathbb{V}\{\mathbf{f}^*\} = \mathbf{K}_{\mathbf{X}^*\mathbf{X}^*} - \mathbf{K}_{\mathbf{X}\mathbf{X}^*}^\top \left[\mathbf{K}_{\mathbf{X}\mathbf{X}} + \sigma_i^2 \mathbf{I}\right]^{-1} \mathbf{K}_{\mathbf{X}\mathbf{X}^*} \tag{6}$$

where $\mathbf{K}_{\mathbf{AB}}$ 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 Squares or random experimentation [9] become increasingly infeasible in high dimensions. As an alternative, we propose to use sequential design approaches which, by greedy selection of the most informative next sample, potentially achieve much faster convergence than fixed designs [7].

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

[1] Sometimes also referred to as a anisotropic (squared exponential) covariance functions [12].

version of the so-called *Expected Improvement* (EI) [7], a known criterion in the design of computer experiment (DACE) community. The expected improvement is for each candidate point, $\mathbf{x}_j$, defined as,

$$\mathrm{EI}(\mathbf{x}_j) = \sigma_{EI} \cdot \mathcal{N}\left(\frac{\mu_{EI}}{\sigma_{EI}}\right) + \mu_{EI} \cdot \Phi\left(\frac{\mu_{EI}}{\sigma_{EI}}\right), \tag{7}$$

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 $\mu_j$ and $\sigma_j$ is the predictive mean and variance of the test point and $\mu_{\max}$ and $\sigma_{max}$ 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. 5-6. The covariance 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 users 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 [10] is designed. Users can listen to previous settings (references) and are allowed to change previous ratings based on the new one. Obviously, this means that ratings are neither directly comparable across users nor between iterations. However, it is not of particular interest to use ratings across users to formulate one single optimal setting, but instead we are interested in personalized settings—one for each user.

# 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 four 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. 2. During part (1), the subjects rate ten randomly chosen balls to learn how to use 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 test 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 test subject. 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. 2). Each tournament (as defined in Fig. 2) was repeated twice resulting in ten tournaments for which the sequence was randomized.

The sound was played back to the test subjects 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 test subject primarily focus on the tonal qualities—not the loudness. An interval of 31.9 seconds in the beginning of the track "Sleeping with the Light on" by Teitur was used as the music piece.

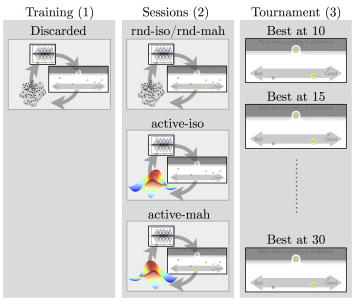



Figure 2: Visualization of the experiment with its 3 sessions: (1) Training, (2) Sessions and (3) Tournament.

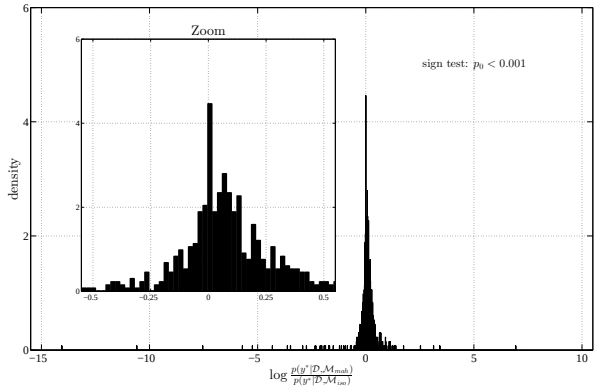



Figure 3: Predictive log-likelihood ratio (Bayes factor) over all leave-one-out cross-validation splits for all twenty-one test subjects. The $p_0$-value gives the probability of the null-hypothesis that the median is equal to zero (the to models are equally well) with the alternative hypothesis that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

## 3.2 Model Analysis

The interactive loop outlined in Sec. 2 has two critical blocks which influence the convergence of the optimization procedure: 1) the GP-model predictions of the subject's objective function at all inputs given only the rated inputs, and 2) 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 test 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 test subject contain only 30 ratings, we use leave-one-out cross validation (LOO-CV) [12] 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.

A proper Bayesian and statistical way of comparing two models [17, 6] 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* [6]. 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 $21 \times 30$ Bayes factor estimates shown in a histogram in Fig. 3. We see that on average, the `mah` model performs the best probabilistic predictions of test subjects's individual ratings and thus appears to be the most suitable model due to 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 null-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.

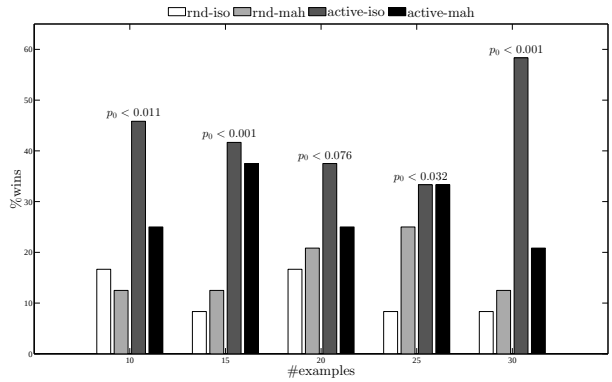
Averaged across test subjects and repetitions, the sequential design is significantly better than random design after any given number of examples, as shown by the $p_0$–values. This is without distinguishing between the two applied covariance functions. It demonstrates the potential of the Bayesian 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 only a few ratings available. This is generally not the case, although the specific Mahalanobis model possesses better generalization capabilities compared to the isotropic model as shown in Sec. 3.2.

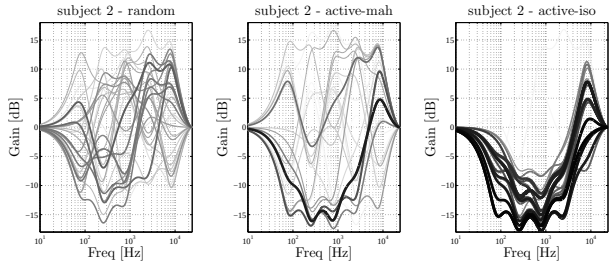
## 4 Discussion and Future Work

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 approach 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




(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 test subjects 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 test 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 ($y \to 0$).

speculate that at least two additions would improve the performance of the Mahalanobis model in the suggested framework. Firstly, the modeling capabilities could be improved by a parametrization of the correlation structure in the Mahalanobis kernel by a small set of parameters, which could then be in-

ferred 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 Gaussian process modeling approach will in general benefit from the addition of weakly informative hyper-priors over especially the length scale parameter, $\sigma_\ell$. This will result in a more robust inference scheme in which unrealistic hyper-plane predictions of the user's objective function would be avoided, thus aiding the sequential design. This is currently being introduced into the modeling framework.

The current evaluation is based on an absolute paradigm with adjustable anchors in terms of previous ratings. For the user, 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 on a larger scale.

# 5 Conclusion

We have proposed a framework for obtaining personalized systems—in particular audio systems—which utilizes a Bayesian probabilistic modeling approach in combination with sequential experimental design. This improves the high-dimensional preference optimization procedure in comparison to random (equivalent 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 advantage of 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 capabilities of the Mahalanobis kernel, it is nevertheless believed that future additions to the framework would be able to exploit these possibilities and hence 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. Multi-task preference learning with an application to hearing aid personalization. *Neurocomputing*, 73(9-9):1177, 2010.

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

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

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

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

[8] 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.

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

[10] 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.

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

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

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

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

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

[16] 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.

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

[18] 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.

Appendix J

# Bounded Gaussian Process Regression

---

Bjørn Sand Jensen, Jens Brehm Nielsen and Jan Larsen. Bounded Gaussian Process Regression. IEEE International Workshop on Machine Learning for Signal Processing, 2013.

2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAMPTON, UK

# BOUNDED GAUSSIAN PROCESS REGRESSION

*Bjørn Sand Jensen, Jens Brehm Nielsen and Jan Larsen*

Department of Applied Mathematics and Computer Science,
Technical University of Denmark,
Matematiktorvet Building 303B, 2800 Kongens Lyngby, Denmark
{bjje,jenb,janla}@dtu.dk

## ABSTRACT

We extend the Gaussian process (GP) framework for *bounded* regression by introducing two bounded likelihood functions that model the noise on the dependent variable explicitly. This is fundamentally different from the implicit noise assumption in the previously suggested warped GP framework. We approximate the intractable posterior distributions by the Laplace approximation and expectation propagation and show the properties of the models on an artificial example. We finally consider two real-world data sets originating from perceptual rating experiments which indicate a significant gain obtained with the proposed explicit noise-model extension.

## 1. INTRODUCTION

Regression is typically defined as learning a mapping from a possible multi-dimensional input to an effectively unbounded one-dimensional observational space, i.e., the space of the dependent variable. However, in many regression problems the observational space is clearly bounded. Examples of such problems include prediction of betting odds, data compression ratios and ratings from perceptual experiments. When the observational space is bounded, modeling the observations with a distribution having infinite support such as the Gaussian distribution, is clearly incorrect from a probabilistic point of view. In this work we will extend the GP framework to allow for principle modeling of such observations.

Gaussian processes (GPs) are currently considered a state-of-the-art Bayesian regression method due to its flexible and non-parametric nature. However, *bounded* regression with GPs has only indirectly been addressed by mapping or *warping* the bounded observations onto a latent unbounded space in which the observational noise can be assumed to be Gaussian [1]. Hereby, the observational model is only modeled implicitly through the warping function. In contrast, we consider observational models or likelihood functions that make assumptions about the noise directly in the observational space, and thus, model the observational noise explicitly.

Possibly, the simplest way to derive a bounded likelihood function is to use a truncated distribution. A natural choice is to use the truncated version of the Gaussian distribution considered in this work. Alternatively, a bounded likelihood function could be derived from a distribution that only has finite support. Of this type, we will consider the beta distribution and derive a bounded likelihood function based on a re-parameterization. For both models we perform inference and predictions based on the Laplace approximation and expectation propagation (EP).

Employing a toy example, we compare the predictive distributions of warped GPs with regression based on the bounded likelihood functions mentioned above. We show that, as expected, the model with the correct noise assumption provides the best expected predictive negative log likelihood (or, alternatively, generalization error). Two examples are used to justify the models in real-world regression scenarios and they show that the two likelihood models provide better model fits compared to the warped GP.

## 2. GAUSSIAN PROCESS REGRESSION

A 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. In effect, we may place the GP as a prior over any finite set of functional values $\mathbf{f} = [f_1, f_2, ..., f_n]^\top$, where $f_i = f(\mathbf{x}_i)$, resulting in a finite multivariate (zero-mean) Gaussian distribution over the set as $p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$, where each element of the covariance matrix $[\mathbf{K}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)_{\boldsymbol{\theta}_c}$ is given by a covariance function $k(\cdot,\cdot)_{\boldsymbol{\theta}_c}$ with parameters $\boldsymbol{\theta}_c$, and where $\mathcal{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ denotes the set of inputs. The GP is effectively used as a prior over functions in non-parametric Bayesian regression frameworks where either the outputs or a likelihood can be parameterized by a smooth and continuous function $f(\cdot)$. In the simplest case the set of observations, $\mathcal{Y} = \{y_i | i = 1, ..., n\}$, consists of the functional val-

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.

978-1-4799-1180-6/13/$31.00 

ues themselves with added i.i.d Gaussian noise with variance $\sigma_n^2$. Hereby, the likelihood function is a standard Gaussian likelihood function parameterized by $f(\cdot)$ defining the mean. Hence, $p(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}}) = \mathcal{N}(y_i|f_i, \sigma^2)$.

Bayes formula gives us—regardless of the likelihood function—the posterior distribution,

$$p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) = \frac{p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}})p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c)}{p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})},$$

where it is typically assumed that the likelihood factorizes over instances such that $p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) = \prod_{i=1}^{n} p(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}})$. The denominator, $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$, is called the *marginal likelihood* or *evidence* given as $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = \int p(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}) p(\mathbf{f}|\mathcal{X}, \boldsymbol{\theta}_c) d\mathbf{f}$. In empirical Bayesian methods the evidence is used to learn point estimates of both likelihood function and prior parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c, \boldsymbol{\theta}_{\mathcal{L}}\}$.

Provided that the likelihood is Gaussian, both the posterior and predictive distribution will be Gaussian (processes) available in closed form [2, Chapter 2]. However, not all real-world problems actually justify the observations to be Gaussian distributed. As mentioned, we consider *bounded* observations, meaning that they in contrast to Gaussian distributed observations do not have infinite support.

## 3. BOUNDED LIKELIHOOD FUNCTIONS

We consider a set $\mathcal{Y} = \{y_i|i = 1, ..., n\}$ of bounded responses $y_i \in ]a, b[$ to an input $\mathbf{x}_i$. In the following we will present three different observational models for this type of response. The first is the warped GP [1], where the likelihood describes warped observations rather than the bounded responses directly. Following this, we propose two different likelihood functions that directly model the bounded responses in a principle probabilistic fashion by assuming particular distributions of the observations defining the noise in the original bounded domain.

### 3.1. Warping

Snelson *et. al* [1] learn a warping, that transforms the original data $\mathcal{Y}$ into a form where the data is modeled by a traditional GP with a Gaussian noise model. Here, we will not consider how to learn the correct warping, but instead use a fixed warping that transforms the bounded responses $y_i$ into unbounded versions $z_i$. Several warping functions would apply, but to allow for direct comparison of all the models we use the inverse cumulative Gaussian (probit) $\Phi^{-1}(\cdot)$—with zero mean and unity variance—such that $z_i = \Phi^{-1}(y_i)$. The resulting model will be referred to as GP-WA.

### 3.2. Truncated Distributions

The simplest route to a bounded likelihood function is to use distributions with infinite support and truncate them to the bounded domain. There are a number of relevant distributions including the truncated student-t and of course the truncated Gaussian (TG) distribution, see e.g. [3]. As a representative for this type of bounding approach, we consider the TG and define the corresponding likelihood function as

$$\begin{aligned}\mathcal{L}_{TG} &\equiv p\left(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}}\right) \\ &= \frac{\nu\mathcal{N}\left(\nu\left(y_i - \mathrm{M}\left(f_i\right)\right)\right)}{\Phi\left(\nu\left(b - \mathrm{M}\left(f_i\right)\right)\right) - \Phi\left(\nu\left(a - \mathrm{M}\left(f_i\right)\right)\right)},\end{aligned} \tag{1}$$

where the distribution is parameterized by the mode $\mathrm{M}(f_i)$ and the domain limits $a$ and $b$ which we assume to be 0 and 1, respectably[1]. The mean of the TG distribution is given by

$$\begin{aligned}\mu\left(f_i\right) &= \mathrm{M}\left(f_i\right) \\ &+ \frac{1}{\nu}\frac{\mathcal{N}\left(\nu\left(a - \mathrm{M}\left(f_i\right)\right)\right) - \mathcal{N}\left(\nu\left(b - \mathrm{M}\left(f_i\right)\right)\right)}{\Phi\left(\nu\left(b - \mathrm{M}\left(f_i\right)\right)\right) - \Phi\left(\nu\left(a - \mathrm{M}\left(f_i\right)\right)\right)}.\end{aligned} \tag{2}$$

Eq. 2 in effect leaves two parametrization options in the sense that we may select the non-parametric function, $f(\cdot)$, to parameterize either the mode or the mean function. Both options are valid from a modeling perspective, but the easiest parametrization is by far the mode, $\mathrm{M}(f_i)$. For prediction speed it may be beneficial to indirectly parameterize the mean, but then the (unique) solution to the mode given the mean must be found numerically or approximately. The numerical approach will severely limit the effectiveness of the posterior approximation and in this work we will therefore focus on the mode parametrization for the TG. Thus, the likelihood function in Eq. 1 is parameterized by the mode as follows $\mathrm{M}(f_i) = \Phi(f_i)$ and the resulting model depicted in Fig. 1 will be referred to as GP-TG

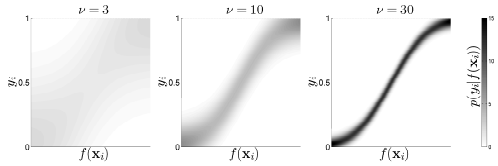



**Fig. 1**. Illustration of the proposed TG likelihood function with $p(y_i|f_i)$ shown as a gray-scale level. Left: $\nu = 3$, Middle: $\nu = 10$ and Right: $\nu = 30$.

### 3.3. Beta

A distribution that imposes bounded responses in a completely natural manner is the beta distribution which has also been applied in standard parametric settings [4, 5]. The beta distribution is therefore an obvious distribution for the bounded observations and we select a parametrization which

[1] We note that the truncated student-t has the same form as the TG and can easily be realized using the methods and implementations presented in this work.

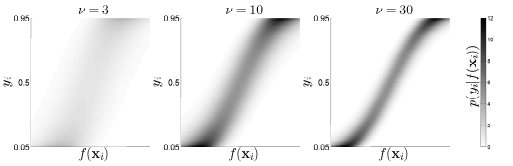



**Fig. 2**. Illustration of the proposed beta likelihood function with $p(y_i|f_i)$ shown as a gray-scale level. Left: $\nu = 3$, Middle: $\nu = 10$ and Right: $\nu = 30$.

expresses the shape parameters, $\alpha, \beta$, of the beta distribution, Beta$(\alpha, \beta)$, in terms of the mean $\mu$ such that

$$\alpha = \nu\mu, \qquad \beta = \nu\left(1-\mu\right).$$

We then parameterize the mean $\mu$ of the beta distribution by the cumulative Gaussian, such that $\mu(f_i) = \Phi(f_i)$. The re-parameterized beta likelihood depicted in Fig. 2 is thereby given by

$$\mathcal{L}_{\text{BE}} \equiv p(y_i|f_i, \boldsymbol{\theta}_{\mathcal{L}}) = \text{Beta}(y_i|\nu\Phi(f_i), \nu\left(1-\Phi(f_i)\right)),$$

and will be referred to as the GP-BE model. Note, that the $\nu$ parameter is an (inverse) dispersion parameter.

## 4. APPROXIMATE INFERENCE AND PREDICTION

For the GP-WA model the likelihood is effectively Gaussian, hence, inference is analytical tractable [1]. However, neither the GP-TG model nor the GP-BE model have analytical tractable posterior distributions. Instead, we must resort to approximations. We consider two different approximate inference schemes—the Laplace approximation and expectation propagation (EP). Both methods approximate the posterior distribution $p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$ with a single Gaussian $q(\mathbf{f})$. In the following we briefly give an overview of the two approximate inference schemes in relations to the bounded likelihood functions. For more details on the approximation schemes see for instance [2].

### 4.1. Laplace Approximation

Possibly, the simplest inference method is the Laplace approximation in which a multivariate Gaussian distribution is used to approximate the posterior, such that $p(\mathbf{f}|\mathcal{X}, \mathcal{Y}, \theta) \approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \mathbf{A}^{-1})$, where $\hat{\mathbf{f}}$ is the mode of the posterior and $\mathbf{A}$ is the Hessian of the negative log posterior at the mode. The mode is found as $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}\right) = \arg\max_{\mathbf{f}} p\left(\mathcal{Y}|\mathbf{f}, \boldsymbol{\theta}_{\mathcal{L}}\right) p\left(\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_c\right)$. The general solution to the problem can be found by considering the un-normalized log posterior and the resulting cost function which is to be maximized, is given by

$$\psi\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}\right) = \log p\left(\mathcal{Y}|\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_{\mathcal{L}}\right) - \frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \frac{N}{2}\log 2\pi,$$

where $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)_{\boldsymbol{\theta}_c}$. The maximization can be solved with a standard Newton-step algorithm given by

$$\hat{\mathbf{f}}^{new} = \left(\mathbf{K}^{-1} + \mathbf{W}\right)^{-1} \cdot \left[\mathbf{W}\hat{\mathbf{f}} + \nabla \log p(\mathcal{Y}|\mathbf{f}, \mathcal{X}, \boldsymbol{\theta}_{\mathcal{L}})\right],$$

where the Hessian $\mathbf{W} = -\nabla\nabla_{\mathbf{f}} \log p(\mathcal{Y}|\mathbf{f})$ is diagonal with elements defined by the second derivative of the log-likelihood function $[\mathbf{W}]_{i,i} = -\frac{\partial^2 \log p(y_i|f_i)}{\partial f_i^2}$. When converged, the resulting approximation is

$$p\left(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}\right) \approx \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, \boldsymbol{\Sigma}\right), \quad \text{where } \boldsymbol{\Sigma} = \left(\mathbf{W} + \mathbf{K}^{-1}\right)^{-1}.$$

Approximating the posterior of $\mathbf{f}$ by the Laplace approximation requires the first two derivatives of the log likelihood. For the TG we will report the general derivatives applicable for any truncated likelihood function based on symmetric densities for which the truncated density can be written as the TG, i.e. in the form

$$p\left(y_i|f_i\right) = \frac{r\left(g\left(y_i|f_i\right)\right)}{s\left(g\left(b|f_i\right)\right) - s\left(g\left(a|f_i\right)\right)}, \tag{3}$$

where we for the TG model defines $g\left(c|f_i\right) = \nu\left(c - \text{M}\left(f_i\right)\right)$. The resulting derivatives for the TG likelihood requires the following partial derivatives

$$\begin{aligned}
\frac{\partial r\left(\cdot\right)}{\partial f_i} &= \nu^2 g\left(y_i\right)\mathcal{N}\left(g\left(y_i\right)\right)\mathcal{N}\left(f_i\right), \\
\frac{\partial^2 r(\cdot)}{\partial^2 f_i} &= \nu^2\mathcal{N}\left(g\left(y_i\right)\right)\mathcal{N}\left(f_i\right) \\
&\quad \left[-\nu\mathcal{N}\left(f_i\right) + g\left(y_i\right)\left(\nu g\left(y_i\right)\mathcal{N}\left(f_i\right) - f_i\right)\right], \\
\frac{\partial s\left(\cdot\right)}{\partial f_i} &= -\nu\mathcal{N}\left(g\left(b\right)\right)\mathcal{N}\left(f_i\right) \quad \text{and} \\
\frac{\partial^2 s(\cdot)}{\partial^2 f_i} &= -\nu\mathcal{N}\left(g\left(b\right)\right)\mathcal{N}\left(f_i\right)\left[\nu g\left(b\right)\mathcal{N}\left(f_i\right) - f_i\right],
\end{aligned}$$

which enter into the derivatives of Eq. 3. The two required partial derivatives for the beta distribution are given by

$$\begin{aligned}
&\frac{\partial \log \text{Beta}(y_i|\cdot)}{\partial f_i} = \nu \cdot \mathcal{N}\left(f_i\right) \\
&\quad \cdot \left[\log(y_i) - \log(1-y_i) - \psi(\alpha) + \psi(\beta)\right] \quad \text{and} \\
&\frac{\partial^2 \log \text{Beta}(y_i|\cdot)}{\partial f_i^2} = \\
&\quad - \nu^2 \cdot \mathcal{N}\left(f_i\right) \cdot \left[\mathcal{N}\left(f_i\right) \cdot \left(\psi^{(1)}(\alpha) + \psi^{(1)}(\beta)\right)\right. \\
&\quad \left. + \frac{f_i}{\nu} \cdot \left(\log(y_i) - \log(1-y_i) - \psi(\alpha) + \psi(\beta)\right)\right],
\end{aligned}$$

where $\psi(\cdot)$ and $\psi^{(1)}(\cdot)$ are the digamma function of zero'th and first order, respectively.

### 4.2. Expectation Propagation

EP also approximates the posterior distribution with a single multivariate Gaussian distribution $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by factorizing the likelihood by $n$ Gaussian factors $t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\Sigma}_i) = \tilde{Z}_i\mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\Sigma}_i)$, where $i = 1, ..., n$. The EP approximation to the full posterior is thus given by

$$\begin{aligned} p(\mathbf{f}|\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) &\approx q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= p(\mathbf{f}, \mathcal{X})\mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{n} \tilde{Z}_i, \end{aligned}$$

where the means $\tilde{\mu}_i$ and variances $\tilde{\Sigma}_i$ have been collected into the vector $\tilde{\boldsymbol{\mu}}$ and diagonal matrix $\tilde{\boldsymbol{\Sigma}}$, respectively. The mean and covariance of the approximation are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}, \qquad \boldsymbol{\Sigma} = \left(\mathbf{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}.$$

EP updates each factor $t_i$ in turn by first removing the factor to yield what is called the *cavity distribution* $q_{-i}(f_i) = \mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})$, where $\mu_{-i} = \Sigma_{-i}([\boldsymbol{\Sigma}]_{i,i}^{-1}\mu_i - \tilde{\Sigma}_i^{-1}\tilde{\mu}_i)$ and $\Sigma_{-i} = ([\boldsymbol{\Sigma}]_{i,i}^{-1} - \tilde{\Sigma}_i^{-1})^{-1}$. Secondly, the factor $t_i$ is updated by projecting the cavity distribution multiplied with the true likelihood term onto a univariate Gaussian. The projection is effectively done by solving the following three integrals

$$Z_i = \int p(y_i|f_i)\mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})df_i, \tag{4}$$

$$\begin{aligned} \frac{dZ_i}{d\mu_{-i}} &= \frac{d}{d\mu_{-i}} \int p(y_i|f_i)\mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})df_i \\ &= \int p(y_i|f_i) \frac{d}{d\mu_{-i}} \left\{\mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})\right\} df_i, \end{aligned} \tag{5}$$

$$\begin{aligned} \frac{d^2Z_i}{d\mu_{-i}^2} &= \frac{d^2}{d\mu_{-i}^2} \int p(y_i|f_i)\mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})df_i \\ &= \int p(y_i|f_i) \frac{d^2}{d\mu_{-i}^2} \left\{\mathcal{N}(f_i|\mu_{-i}, \Sigma_{-i})\right\} df_i. \end{aligned} \tag{6}$$

Neither the beta likelihood nor the TG likelihood yield analytical tractable solutions for these three integrals, but the one-dimensional integrals can be solved numerically for the EP inference.

### 4.3. Predictive Distributions

Naturally, we want to predict future values of both the latent functional value $f^*$ and data label $y^*$. For all models the posterior distribution over $\mathbf{f}$ is effectively Gaussian[2]. Hence, the predictive distribution $p(f^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) = \mathcal{N}(f^*|\mu^*, \sigma_*^2)$ of latent functional values is Gaussian and is derived just as in the standard cases in a straight forward manner (see e.g. [2, Chapter 2-3]).

The predictive distribution of future targets $p(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*)$ involves computing the integral

$$p(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) = \int p(y^*|f^*)\mathcal{N}(f^*|\mu^*, \sigma_*^2)df^*.$$

For the GP-WA, the predictive distribution has a closed-form solution [1]

$$p_{\text{GP-WA}}(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*) = \frac{\mathcal{N}(\Phi^{-1}(y^*)|\mu^*, \sigma_*^2)}{\Phi(\Phi^{-1}(y^*))}.$$

In case of the GP-BE and GP-TG the predictive distribution is not given in closed form. Instead, the integral must be computed using numerical methods. Predictions of the mean, $\mathbb{E}(y) \in \,]\,0; 1\,[$, are in the bounded case given by

$$\begin{aligned} \mathbb{E}_{p(y^*|\cdot)}\{y^*\} &= \int_0^1 y^* p(y^*|\mathcal{Y}, \mathcal{X}, \mathbf{x}^*)dy^* \qquad (7) \\ &= \int \mathcal{N}(f^*|\mu^*, \sigma_*^2) \int_0^1 y^* p(y^*|f^*)dy^*df^* \\ &= \int \mathcal{N}(f^*|\mu^*, \sigma_*^2)\mathbb{E}_{p(y^*|f^*)}\{y^*\}df^*. \qquad (8) \end{aligned}$$

Given the cumulative Gaussian warping, Eq. 7 can be solved analytically for the GP-WA model. In Eq. 8 the mean of the likelihood occurs, which in the beta case is parameterized by a cumulative Gaussian and given the specific choice of warping this results in a closed form solution expressed by[3]

$$\mathbb{E}_{\text{GP-WA}}\{y^*\} = \mathbb{E}_{\text{GP-BE}}\{y^*\} = \Phi\left(\frac{\mu^*}{\sqrt{1 + (\sigma^*)^2}}\right).$$

In case of the GP-TG model, Eq. 7 has no analytical form and must be solved by one-dimensional numerical approximation.

## 5. SIMULATION EXAMPLE

In order to illustrate the difference between the warped and bounded likelihood approaches we consider an artificial example with added noise. It is generated by drawing a one-dimensional function from a zero-mean Gaussian process with a squared exponential (SE) kernel with length scale, $\sigma_l = 1$, and noise variance $\sigma_f = \exp(1)$. Three different types of noise are then added: The first type (WA) is i.i.d Gaussian noise added directly on $f$ and transformed through $\Phi(\cdot)$ which corresponds to the noise assumption in the warped

[2]For the warped GP the posterior is exactly Gaussian, whereas we for the two other models have approximated—either by Laplace or EP—the posterior with a Gaussian.

[3]Keep in mind that although there is an equal sign between the predictive mean of the cumulative-warped and the beta model, the means will in general be different due to difference in the *latent* predictive distributions of the GP.

| | Squared Exponential ($\sigma_f^2 = 2, \ell = 1$) | | |
|---|---|---|---|
| | WA | TG | BE |
| GP-WA | **-129.8** (6.4) | -82.0 (7.3) | -165.3 (31.1) |
| GP-TG | -91.0 (19.6) | **-96.8 (4.5)** | -81.8 (14.5) |
| GP-BE | -119.8 (7.7) | -91.2 (6.3) | **-195.2 (24.6)** |
| | Periodic ($\sigma_f^2 = 3, \ell = 0.8, \lambda = 5$) | | |
| | WA | TG | BE |
| GP-WA | **-93.2 (6.9)** | -80.6 (11.0) | -70.8 (10.4) |
| GP-TG | -76.2 (10.3) | **-91.6 (9.4)** | -66.0 (12.9) |
| GP-BE | -88.5 (3.8) | -84.5 (7.8) | **-99.8 (15.5)** |

**Table 1**. Expected predictive negative log likelihood (and standard deviation) for each of the three models (GP-WP, GP-TG, GP-BE)) evaluated on a specific function with additive noise from ten random realizations of the noise for each corresponding noise types: WA, BE and TG. The noise free function is drawn from a GP prior with the indicated covariance functions and parameter values (defined in [6])

GP. In the second case (TG), $f$ is transformed through $\Phi(\cdot)$ before adding noise based on the mode-parameterized TG distribution, thus corresponding to the noise assumption of the TG likelihood. In the third case (BE), we add noise based on the mean-parameterized beta distribution.

In order to visualize the special nature of bounded responses and the difference between the models, we have illustrated the WA noise case in Fig. 3, where all three bounded models are evaluated. Both the Laplace approximation and EP have been used for inference for the beta and TG model. The hyper-parameters are in all cases optimized using evidence maximization. The main difference of the three models occurs at the domain boundaries, where the GP-WA model concentrates the entire mass almost at the boundary. The predictive distribution of the GP-TG model generally has a similar shape over the entire domain with its mean always spaced significantly far from the boundary, whereas the GP-BE can also have its mean very close to the boundary as for the GP-WA model, but still retain mass away from the boundary. No significant differences between the two inference schemes are evident. Since the EP scheme requires numerical solutions to the integrals in Eq. 4-6, the Laplace approximation will be used in the reminder of this article.

We evaluate the ability of the models to model different noise distributions by comparing the predictive log likelihood for the previously mentioned dataset based on the Laplace approximation. A second example is added in which the function is drawn from a GP with a periodic covariance function. The predictive log likelihood for both examples is reported in Tab. 5 and is the average over ten realizations of the noise. As expected, we see that the model corresponding to the added noise type always results in the lowest negative likelihood, indicating a better model fit.

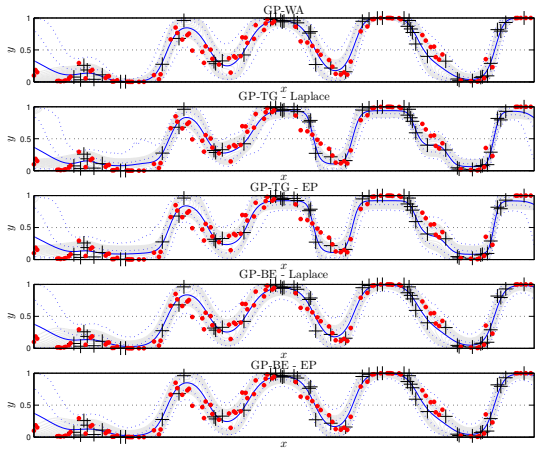



**Fig. 3**. Predictive distributions for the three models: GP-WA, GP-TG and GP-BE. For GP-TG and GP-BE both Laplace and EP inference are shown, where training data: $+$, test examples: $\cdot$, predictive mean: $-$ and 68% and 95% percentiles: $\cdots$. Also, contours of the predictive distribution are shown in gray, where the intensity reflects probability mass concentration.

## 6. PERCEPTUAL AUDIO EVALUATIONS

In order to demonstrate the difference between the three considered models in a real-world scenario, we have tested the three models on two datasets consisting of subjective ratings performed while listening to audio through a hearing aid (HA) compressor with different settings.

The first dataset [7], HA-I, contains six compression ratio settings (including one without compression) and three release-time settings. This results in sixteen non-trivial combinations of the settings with $\mathbf{x}^s \in \mathbb{R}^2$, that are rated three times by each of the seven test subjects, $u$, while listening to a speech signal. The dataset also contains an complete six point audiogram on both left and right ear, $\mathbf{x}^u \in \mathbb{R}^{2\times 6}$, of the hearing impaired test subjects. The audio signal resulting from each compressor setting is represented by standard audio features, namely thirty Mel frequency cepstral coefficients, $\mathbf{x}^a \in \mathbb{R}^{30}$. Thus, for one setting, $s$, each test subject, $u$, rated the audio signal, $a$. This results in a collection of inputs for this specific rating which we collect in $\mathbf{x} = \{\mathbf{x}^u, \mathbf{x}^a, \mathbf{x}^s\}$. We use the multi-task kernel formulation [8] and define the covariance function as $k(\mathbf{x}_i, \mathbf{x}_j) = k^u_{\text{SE-ARD}}(\mathbf{x}^u_i, \mathbf{x}^u_j)\left(k^a_{SE}(\mathbf{x}^a_i, \mathbf{x}^a_j) + k^s_{SE}(\mathbf{x}^s_i, \mathbf{x}^s_j)\right)$ where all covariance functions are squared exponential (SE), the first one with automatic relevance determination (SE-ARD).

The second dataset [9], HA-II, contains three input settings related to the compression ratio, attack time and release time of a HA dynamic range compressor, thus $\mathbf{x}^s \in \mathbb{R}^3$. Four

| | | GP-WA | GP-TG | GP-BE |
|---|---|---|---|---|
| HA-I | -log $p(y^*)$ | -66.1 | -96.1 | **-101.2** |
| | MSE | 0.013 | 0.001 | 0.010 |
| HA-II | -log $p(y^*)$ | -7.7 | -9.3 | **-14.1** |
| | MSE | 0.031 | 0.030 | 0.035 |

**Table 2**. **HA-I** Mean square error (MSE) and expected predictive negative log likelihood over 10 random sets. We find a significant difference in log likelihood at the 5% level between GP-TG and the two other models but not between GP-TG and GP-BE. For MSE the only significant difference is between GP-TG and GP-BE. **HA-II** Mean square error (MSE) and negative log likelihood over 10 folds. Considering the negative log likelihood only the GP-BE is significantly better than the GP-WA in a paired t-test. There is no significant difference between GP-TG and the other models. The GP-BE is significantly different in terms of MSE than the two others.

subjects have rated 50 combinations of inputs in relation to general preference while listening to a speech-in-background-noise signal. The dataset does not contain any data describing the subjects, hence we use only a single squared exponential covariance function.

We initialize the hyper-parameters in the (common) covariance function to the same value for all models, but initialize the likelihood noise parameter with multiple values in a grid pattern after which all the hyper-parameters are optimized using evidence maximization. We then report the performance of the model which yields the largest evidence after maximization. For the purpose of comparing the three models, we will simply consider the Laplace approximation and a retest scenario in which we train on a random repetition and test on another repetition for each setting. We repeat this three times and evaluate the resulting predictive likelihood and mean square error (MSE). The results are listed in Tab. 6. We note from the negative predictive log likelihood that the beta distribution provides a better fit to the noise compared to the other two models given the two real-world datasets presented here.

## 7. DISCUSSION AND CONCLUSION

In the present work, we outlined two bounded likelihood functions for bounded Gaussian process regression which in contrast to previous work make explicit assumptions about the noise in the bounded observation space. In the two considered examples we found the beta model to be better than the two other models in terms of the predictive log likelihood. These results together with the artificial examples support the application of all three models in the non-parametric Gaussian process framework. However, the optimal model obviously depends on the actual noise distribution in a given application. We therefore foresee addition and inclusion of other noise models based on other distribution with finite support.

Implementations of the various likelihoods are available [10] for use in the `gpml` toolbox [6] and can easily be extended to support more advanced link functions [11], which will make the models (both the bounded and the warped) even more flexible. In particular, we suggest to use a mixture of cumulative Gaussian link functions which do not complicate predictions significantly. Furthermore, we suggest to evaluate the performance of the deterministic approximations by the use of MCMC-sampling methods.

In conclusion, we have extended the Gaussian process framework to include bounded likelihood functions allowing for explicit specification of the likelihood model in applications where bounded observations are present and support an explicit noise model.

## 8. REFERENCES


[1] E. Snelson, C. E. Rasmussen, and Z. Ghahramani, "Warped Gaussian Processes," in *Advances in Neural Information Processing Systems*, vol. 16. MIT Press, 2004.

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

[3] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1 & 2, Wiley, 2nd edition, 1994-1995.

[4] S. Ferrari and F. Cribari-Neto, "Beta Regression for Modelling Rates and Proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, Aug. 2004.

[5] M. Smithsen and J. Verkuilen, "A Better Lemon-squeezer? Maximum Likelihood Regression with Beta-distributed Dependent Variables.," *Australian Journal Psychology*, vol. 57, pp. 98–98, 2005.

[6] C. E. Rasmussen and H. Nickisch, "Matlab GPML Toolbox," 2010.

[7] E. Schmidt, *Hearing Aid Processing of Loud Speech and Noise Signals: Consequences for Loudness Perception and Listening Comfort.*, Ph.D. thesis, Technical University of Denmark, 2006.

[8] E. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Advances in Neural Information Processing Systems*, vol. 20, pp. 153–160. MIT Press, 2008.

[9] J. B. Nielsen, "Preference based personalization of hearing aids," M.Sc. Thesis, 2010.

[10] J. B. Nielsen and B. S. Jensen, "Bounded Gaussian Process Regression - Supplementary Material," `www.imm.dtu.dk/pubdb/p.php?6683`, 2013.

[11] T. C. Martins Dias and C. A. R. Diniz, "The use of Several Link Functions on a Beta Regression Model: a Bayesian Approach.," *AIP Conference Proceedings*, vol. 1073, no. 1, pp. 144, 2008.

# Bibliography


[1] Last.fm dataset, the official song tags and song similarity collection for the Million Song Dataset. http://labrosa.ee.columbia.edu/millionsong/lastfm.

[2] The Echo Nest Taste profile subset, the official user data collection for the Million Song Dataset. http://labrosa.ee.columbia.edu/millionsong/tasteprofile.

[3] Samer A. Abdallah, Henrik Ekeus, Peter Foster, Andrew Robertson, and Mark D. Plumbley. Cognitive music modelling: An information dynamics approach. *2012 3rd International Workshop on Cognitive Information Processing, CIP 2012*, page 6232940, 2012.

[4] T. S. Alstrom, Bjørn S. Jensen, Mikkel N. Schmidt, Natalie V. Kostesha, and Jan Larsen. Haussdorff and hellinger for colorimetric sensor array classification. *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

[5] M. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for Vector-Valued Functions: A Review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

[6] Amazon. `http://amazon.com/`, 2012. [Online; accessed 17-12-2012].

[7] J. Arenas-García, A. Meng, K.B. Petersen, T. Lehn-Schiøler, L.K. Hansen, and J. Larsen. *Unveiling Music Structure Via pLSA Similarity Fusion*, pages 419–424. IEEE, 2007.

[8] A. Asuncion, M. Welling, P. Smyth, and Y.W. Teh. On smoothing and inference for topic models. *25th Conference on Uncertainty in Artificial Intelligence*, pages 27–34, 2009.

[9] J. J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of 3rd International Conference on Music Information Retrieval*, pages 157–163, Paris, France, 2002.

[10] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.

[11] L. Barrington, A. B. Chan, and G. Lanckriet. Dynamic texture models of music. *International Conference on Acoustics, Speech, and Signal Processing*, pages 1589–1592, 2009.

[12] L. Barrington, A.B. Chan, and G. Lanckriet. Modeling Music as a Dynamic Texture. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):602–612, March 2010.

[13] L. Barrington, D. Turnbull, and G. Lanckriet. Game-powered machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 109(17):6411–6, April 2012.

[14] L. Barrington and M. Yazdani. Combining feature kernels for semantic music retrieval. *International Conference on Music Information Retrieval*, 2008.

[15] M. A. Bartsch and G. H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, October 2001.

[16] S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application.* Cambridge University Press, 2006.

[17] S. Beggs, S. Cardell, and J. Hausman. Assessing the potential demand for electric cars. *Journal of Econometrics*, 17(1):1, 1981.

[18] T Bertin-Mahieux. Clustering beat-chroma patterns in a large music database. In *International Society for Music Information Retrieval Conference*, 2010.

[19] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A Model for Predicting Social Tags from Acoustic Features on Large Music Databases. *Journal of New Music Research*, 37(2):115–135, June 2008.

[20] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[21] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[22] A. Birlutiu, P. Groot, and T. Heskes. Multi-task preference learning with an application to hearing aid personalization. *Neurocomputing*, 73(7-9):1177–1185, March 2010.

[23] A. Birlutiu, P Groot, and T. Heskes. Multi-task preference learning with an application to hearing aid personalization. *Neurocomputing*, 73(7-9):1177–1185, March 2010.

[24] A. Birlutiu, P. Groot, and T. Heskes. Efficiently learning the preferences of people. *Machine Learning*, pages 1–28, 2012.

[25] C.M. Bishop. *Neural networks for pattern recognition*. Clarendon Press, 1995.

[26] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[27] D. Blei and J. Lafferty. Correlated Topic Models. In Y Weiss, B Schölkopf, and J Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 147–154. MIT Press, Cambridge, MA, 2006.

[28] D. Blei and J. McAuliffe. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, Cambridge, MA, 2008.

[29] D. M Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[30] R.D. Bock and J.V. Jones. The measurement and prediction of judgment and choice. 1968.

[31] E. Bonilla, S. Guo, and S. Sanner. Gaussian Process Preference Elicitation. In J Lafferty, C K I Williams, J Shawe-Taylor, R S Zemel, and A Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 262–270. 2010.

[32] E.. V. Bonilla, F. V. Agakov, and C. K.I. Williams. Kernel multi-task learning using task-specific features. *Journal of Machine Learning Research*, 2:43–50, 2007.

[33] Voice Box. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 2012. [Online; accessed 17-12-2012].

[34] R.A. Bradley. Rank analysis of incomplete block designs .2. additional tables for the method of paired comparisons. *Biometrika*, 41(3-4):502–537, 1954.

[35] R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs .1. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[36] P.B. Brockhoff and R.H.B. Christensen. Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference*, 21(3):330–338, 2010.

[37] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. *ACM International Conference Proceeding Series*, 119:89–96, 2005.

[38] Auditory Toolbox by M. Slaney. `https://engineering.purdue.edu/~malcolm/interval/1998-010/`, 2012. [Online; accessed 17-12-2012].

[39] CAL-500. `http://cosmal.ucsd.edu/cal/`, 2012. [Online; accessed 17-12-2012].

[40] Zhe Cao, Tao Qin, Tie-Yan Liu, Hang Li, and Ming-Feng Tsai. Learning to rank: From pairwise approach to listwise approach. *ACM International Conference Proceeding Series*, 227:129–136, 2007.

[41] G. Casella. An Introduction to Empirical Bayes Data Analysis. *The American Statistician*, 39(2):83–87, 1985.

[42] Ò. Celma. Foafing the music bridging the semantic gap in music recommendation. In *5th International Semantic Web Conference (ISWC)*, Athens, GA, USA, 2006.

[43] O. Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, New Work, 2010.

[44] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10:273–304, 1995.

[45] Gang Chen, Tian-Jiang Wang, and Perfecto Herrera. A novel music retrieval system with relevance feedback. *3rd International Conference on Innovative Computing Information and Control, ICICIC'08*, page 4603347, 2008.

[46] W. Chu and Z Ghahramani. Extensions of Gaussian processes for ranking: semi-supervised and active learning. In *Workshop Learning to Rank at Advances in Neural Information Processing Systems 18*, 2005.

[47] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6, 2005.

[48] W. Chu and Z. Ghahramani. Preference learning with Gaussian processes. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pages 137–144, 2005.

[49] D. A. Cohn, Z. Ghahramani, and M.I.. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[50] T.M. Cover and J.A. T. *Elements of information theory*. Wiley, 1991.

[51] A. M Croom. Music, neuroscience, and the psychology of well-being: a precis. *Frontiers in psychology.*, 2:393, 2012.

[52] I. Cross. Cognitive science and the cultural nature of music. *Topics in Cognitive Science*, 4(4):668, 2012.

[53] L. Csato. *Gaussian Processes - Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.

[54] CSIRAC. `http://news.bbc.co.uk/2/hi/technology/7458479.stm`, 2008. [Online; accessed 17-12-2012].

[55] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[56] J. R. Deller and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice Hall, New Jersey, 1987.

[57] D. Deutsch. *The psychology of music*. Academic Press, 1999. Academic Press Series in Cognition and Perception.

[58] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, 52(8):3913–3927, 2008.

[59] D. K. Duvenaud, H. Nickisch, and C.E. Rasmussen. Additive gaussian processes. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 226–234. 2011.

[60] The Echonest. `http://the.echonest.com/`, 2012. [Online; accessed 17-12-2012].

[61] D. Ellis. In *International Conference on Music Information Retrieval (ISMIR)*.

[62] D. P. W. Ellis and B. Whitman. The quest for ground truth in musical artist similarity. *SIGCHI International Symposium on Music Information Retrieval*, 2002.

[63] A Faisal, J Gillberg, J Peltonen, G Leen, and S Kaski. Sparse Nonparametric Topic Model for Transfer Learning. *dice.ucl.ac.be*.

[64] R.M. Fano. *Transmission of Information*. M.I.T. Press, first edition, 1961.

[65] V. V. Fedorov. *Theory of optimal experiments*. Academic Press, 1972.

[66] S. Ferrari and F. Cribari-Neto. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815, August 2004.

[67] Gaussian Process for Machine Learning (GPML toolbox). `http://www.gaussianprocess.org/gpml/code/matlab/doc/`, 2012. [Online; accessed 17-12-2012].

[68] J. Frow. *Genre*. Routledge, New York, NY, USA, 2005.

[69] T. Fujishima. Realtime chord recognition of musical sound: A system using common lisp music. *International Computer Music Conference (ICMC)*, pages 465–467, 1999.

[70] Y. Gao, I. Kontoyiannis, and E. Bienenstock. Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99, 2008.

[71] R. Garnett, M. A. Osborne, and S. J. Roberts. Bayesian optimization for sensor set selection. *Information Processing In Sensor Networks*, pages 209–219, 2010.

[72] G Geleijnse, M. Schedl, and P Knees. The quest for ground truth in musical artist tagging in the social web era. *International Conference on Music Information Retrieval*, 2007.

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

[74] M. Girolami and S. Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

[75] M. E. Glickman and Shane T. Jensen. Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, 127(1-2):279–293, 2005.

[76] B. Gold and N. Morgan. *Speech and Audio Signal Processing / Processing and Perception of Speech and Music*. John Wiley and sons Inc., 2000.

[77] F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005.

[78] T. L Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl:5228–35, April 2004.

[79] P Groot, A. Birlutiu, and T. Heskes. Bayesian Monte Carlo for the global optimization of expensive functions. *Proc. of ECAI*, 1(1.5):2, 2010.

[80] J. Guiver and E. Snelson. Bayesian inference for plackett-luce ranking models. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pages 377–384, 2009.

[81] B. Haasdonk and C. Bahlmann. Learning with distance substitution kernels. In *in Pattern Recognition, Proceedings of the 26th DAGM Symposium*, 2004.

[82] P. Hamel and D. Eck. Learning features from music audio with deep belief networks. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 339–344, August 2010.

[83] L.K. Hansen, P. Ahrendt, and J. Larsen. Towards cognitive component analysis. In *AKRR 2005-International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

[84] V.L. Hansen. *Functional analysis - entering Hilbert space*. World Scientific, 2006.

[85] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. Enser, and C. J. Sandom. Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. In *Mastering the Gap: From Information Extraction to Semantic Representation / 3rd European Semantic Web Conference*, 2006.

[86] R. Henao and O. Winther. Pass-gp: Predictive active set selection for gaussian processes. *IEEE International Workshop on Machine Learning for Signal Processing 2010 (MLSP 2010)*, pages 148–153, 2010.

[87] H. Hermansky. Perceptual linear predictive plp analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[88] K Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48:246–268, 1936.

[89] K. M. Higgins. Biology and culture in musical emotions. *Emotion Review (SAGE Journals)*, 4(3):273–282, 2012.

[90] M. Hoffman, D. Blei, and P Cook. Content-based musical similarity computation using the hierarchical dirichlet process. In *9th International Conference on Music Information Retrieval*, pages 349–354, 2008.

[91] M Hoffman, D Blei, and P Cook. Easy as CBA: A simple probabilistic model for tagging music. *International Society for Music Information Retrieval Conference*, pages 369–374, 2009.

[92] T. Hofmann. Probabilistic latent semantic analysis. *Proceedings of Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[93] N. Houlsby, Jose Miguel Hernandez-Lobato, Ferenc Huszar, and Z. Ghahramani. Collaborative gaussian processes for preference learning. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2105–2113. 2012.

[94] X. Hu and S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *11th International Society for Music Information Retrieval Conference*, pages 619–624, 2010.

[95] Digital Music Report 2012 IFPI. `http://http://www.ifpi.org/content/library/DMR2012.pdf/`, 2012. [Online; accessed 23-12-2012].

[96] Santner T. J., Brian J Williams, and William Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics, 2003.

[97] T. Jebara and A. Howard. Probability Product Kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[98] J. H. Jensen, D. P. W. Ellis, M G. Christensen, and S Holdt Jensen. Evaluation of distance measures between gaussian mixture models of mfccs. *8th International Conference on Music Information Retrieval*, 2007.

[99] N. Jiang, P. Grosche, V. Konz, and M. Müller. Analyzing chroma feature types for automated chord recognition. In *42nd AES Conference on Semantic Audio*, Ilmenau, Germany, 2011.

[100] T. Joachims. Optimizing search engines using clickthrough data. *International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

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

[102] D. R Jones, M. Schonlau, and William J Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[103] P. N. Juslin. Are musical emotions invariant across cultures? *EMOTION REVIEW*, 4(3):283–284, 2012.

[104] P. N. Juslin and P. Laukka. Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening. *Journal of New Music Research*, 33(3):217–238, September 2004.

[105] S K., S Narayanan, and S Sundaram. Acoustic topic model for audio information retrieval. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 37–40, 2009.

[106] Sofia K-Means. `http://code.google.com/p/sofia-ml/wiki/SofiaKMeans`, 2012. [Online; accessed 21-12-2012].

[107] Y.E. Kim, E. Schmidt, and L. Emelle. Moodswings: A collaborative game for music mood label collection. *International Conference on Music Information Retrieval*, pages 231–236, 2008.

[108] Y.E. Kim, E.M. Schmidt, Raymond Migneco, B.G. Morton, Patrick Richardson, Jeffrey Scott, J.A. Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *11th International Music Information Retrieval Conference*, pages 255–266, 2010.

[109] A. Klapuri and M. Davy. *Signal processing methods for music transcription*. Springer, 2006. Includes bibliographical references (p. [391]-428) and index.

[110] P. Knees and G. Widmer. Searching for music using natural language queries and relevance feedback. *Advances in Neural Information Processing Systems*, 4918:109–121, 2008.

[111] I. Kontoyiannis, P.H. Algoet, Yu M. Suhov, and A.J. Wyner. Nonparametric entropy estimation for stationary process and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.

[112] M. W. Krasilovsky, Sidney Shemel, John M. Gross, and Jonathan Feinstein. *This Business of Music, 10th Edition*. Billboard Books, 2007.

[113] A. Krause and C. Guestrin. Nonmyopic active learning of Gaussian processes: An exploration- exploitation approach. *24th International Conference on Machine Learning*, pages 449–456, 2007.

[114] A. Krithara, M. Amini, C. Goutte, and J. Renders. An extension of the aspect plsa model to active and semi-supervised learning for text classification. *ARTIFICIAL INTELLIGENCE: THEORIES, MODELS AND APPLICATIONS, PROCEEDINGS*, 6040:183–192, 2010.

[115] A. Krithara, C. Goutte, M.-R. Amini, and J.-M. Renders. Reducing the annotation burden in text classification. *International Conference on Multidisciplinary Information Sciences and Technologies*, 2(1):279–283, 2006.

[116] H.J. Kushner. New method of locating maximum point of arbitrary multipeak curve in presence of noise. *Joint Automatic Control Conference*, pages 69–79, 1963.

[117] S. Lacoste-Julien, Fei Sha, and M. Jordan. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 897–904. 2009.

[118] P. Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[119] last.fm. `http://last.fm/`, 2012. [Online; accessed 17-12-2012].

[120] E. Law and L. Von Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 2009.

[121] N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. *Advances in Neural Information Processing Systems*, (15):625–632, 2003.

[122] M. Lazaro-Gredilla. Bayesian warped gaussian processes. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1628–1636. 2012.

[123] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[124] H. Lee, L. Yan, P. Pham, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, pages 1096–1104, 2009.

[125] M. Levy and M. Sandler. Learning Latent Semantic Models for Music from Social Tags. *Journal of New Music Research*, 37(2):137–150, June 2008.

[126] M. Levy and M. Sandler. Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia*, 11(3):383–395, 2009.

[127] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19, February 2006.

[128] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

[129] D V Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, (4):986–1005, 1956.

[130] D. V. Lindley. Bayesian statistics. a review. *Regional Conference Series in Applied Mathematics*, (2):1–83, 1971.

[131] D.V. LINDLEY. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[132] Daniel Lizotte, Tao Wang, Michael Bowling, Dale Schuurmansdepartment, and Computing Science. Gaussian process regression for optimization. 2008.

[133] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *In International Symposium on Music Information Retrieval*, 2000.

[134] B. Logan and A. Salomon. A music similarity function based on signal analysis. *Multimedia and Expo, 2001, IEEE International Conference on*, 2001.

[135] R.D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley (New York), 1959.

[136] D. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

[137] D. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

[138] D.J.C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[139] D.J.C. MacKay. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.

[140] J. Madsen. *Modeling of Emotions expressed in Music using Audio features*. DTU Informatics, Master Thesis, http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6036, 2011.

[141] C. L. Mallows. Non-null ranking models i. *Biometrika*, 44(1/2):114–130, 1957.

[142] M. Mandel and D. Ellis. A Web-Based Game for Collecting Music Metadata. *Journal of New Music Research*, 37(2):151–165, June 2008.

[143] M. I. Mandel and Daniel P. W. Ellis. Song-Level Features and Support Vector Machines for Music Classification. In Joshua D. Reiss and Geraint A. Wiggins, editors, *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 594–599, September 2005.

[144] C. D. Manning, P Raghavan, and H. Schutze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[145] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman and Hall, 1989. Monographs on statistics and applied probability : 37.

[146] B. McFee, L. Barrington, and G. Lanckriet. Learning content similarity for music recommendation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2207–2218, 2012.

[147] B McFee and G. Lanckriet. Metric learning to rank. *International Conference on Machine Learning*, 2010.

[148] B. McFee and G. R. G. Lanckriet. Hypergraph models of playlist dialects. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, pages 343–348, 2012.

[149] Matt Mcvicar and Tijl De Bie. CCA and a multi-way extension for investigating common components between audio , lyrics and tags. In *CMMR*, number June, pages 19–22, 2012.

[150] A. Meng. *Temporal Feature Integration for Music Organisation*. PhD thesis, Technical University of Denmark, DTU, 2006.

[151] A. Meng, P. Ahrendt, and J. Larsen. Improving music genre classification by short-time feature integration. *IEEE International Confonference on Acoustics Speech Signal Processing*, 5:497–500, 2004.

[152] A. Meng and J. Shawe-Taylor. An investigation of feature models for music genre classification using the support vector classifier. *International Conference on Music Information Retrieval*, pages 604–609, 2005.

[153] P. Mermelstein. Distance Measures for Speech Recognition–Psychological and Instrumental. In *Joint Workshop on Pattern Recognition and Artificial Intelligence*, 1976.

[154] T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

[155] R. Miotto, L. Barrington, and G. Lanckriet. Improving Auto-tagging by Modeling Semantic Co-occurrences. In *International Society of Music Information Retrieval Conference*, pages 297–302, 2010.

[156] J. Mockus, V. Tiesis, and Zilinskas A. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117–129, 1978.

[157] L.L. Mølgaard. *Context based multimedia information retrieval*. PhD thesis, Technical Univeristy of Denmark, 2008.

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

[159] B. C. J. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition*. Academic Press, April 2003.

[160] M.. Muller. *Information retrieval for music and motion*. Springer, 2009.

[161] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

[162] I. Murray and R.P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. *arXiv preprint arXiv:1006.0868*, 2(1):1–9, 2010.

[163] T. Næ s, P.B. Brockhoff, and O. Tomic. *Statistics for sensory and consumer science*. Wiley, 2010.

[164] A Naish-Guzman. The generalized FITC approximation. *Advances in Neural Information Processing Systems*, pages 1–8, 2008.

[165] R.M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical report, 1997.

[166] Netflix. `http://netflix.com/`, 2012. [Online; accessed 17-12-2012].

[167] H. Nickisch and C.E. Rasmussen. Approximations for binary gaussian process classification. *JOURNAL OF MACHINE LEARNING RESEARCH*, 9(1):2035–2078, 2009.

[168] H. Nickisch and M. Seeger. Multiple Kernel Learning: A Unifying Probabilistic Viewpoint. *arXiv preprint arXiv:1103.0897*, 2011.

[169] A. B. Nielsen, L. K. Hansen, and U. Kjems. Pitch based sound classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 3:III788–III791, 2006.

[170] Jens Brehm Nielsen and J. Nielsen. Efficient individualization of hearing aid processes sound. *Submittet to the IEEE International Conference on Speech Audio and Signal Processing (ICASSP 2013)*, 2013.

[171] Nielsen.com. `http://www.nielsen.com/us/en/insights/press-room/2012/music-discovery-still-dominated-by-radio--says-nielsen-music-360.html`, 2012. [Online; accessed 23-12-2012].

[172] M. Osborne. *Bayesian Gaussian Processes for Sequential Prediction , Optimisation and Quadrature*. PhD thesis, University of Oxford, 2010.

[173] R. P. Paiva. *Melody detection in polyphonic audio*. PhD thesis, 2006.

[174] E. Pampalk, T. Pohle, and G. Widmer. Dynamic playlist generation based on skipping behavior. In *Proceedings of 6th International Conference on Music Information Retrieval*, London, UK, 2005.

[175] Y. Panagakis, C. Kotropoulos, and G.R. Arce. Music genre classification using locality preserving nonnegative tensor factorization and sparse representations. 2011.

[176] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Techical report, IRCAM, 2004.

[177] M. K. Petersen, M. Mørup, and L. K. Hansen. Sparse but emotional decomposition of lyrics. In *3rd International Workshop on Learning Semantics of Audio Signals (LSAS)*, pages 31–43, 2009.

[178] R.L. Plackett. The analysis of permutations. *Applied Statistics*, 24(2):193–202, 1975.

[179] J. Quiñonero Candela and C.E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[180] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. 2009.

[181] F. Rahman and J. Siddiqi. Semantic annotation of digital music. *J. Comput. Syst. Sci.*, 78(4):1219–1231, July 2012.

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

[183] J. Robinson. The expression and arousal of emotion in music. *Journal of Aesthetics and Art Criticism*, 1994.

[184] J.A. Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[185] O. Sacks. *Musicophilia: Tales of Music and the Brain*. Knopf Doubleday Publishing Group, 2007.

[186] J. Salamon and E. Gomez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[187] P. Sandhaus and S. Boll. Semantic analysis and retrieval in personal and social photo collections. *Multimedia Tools and Applications*, 51(1):5–33, December 2010.

[188] M. Schedl and Peter Knees. Personalization in Multimodal Music Retrieval. *9th International Workshop on Adaptive Multimedia Retrieval*, 2011.

[189] A Schindler, R Mayer, and A Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *13th International Conference on Music Information Retrieval (ISMIR 2012)*. 2012.

[190] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[191] M. Seeger, C. K. I. Williams, N. D. Lawrence, and Sheeld S Dp. Fast forward selection to speed up sparse gaussian process regression. In *in Workshop on AI and Statistics 9*, 2003.

[192] M. W. Seeger, N. D. Lawrence, and R. Herbrich. Efficient nonparametric bayesian modelling with sparse gaussian process approximations. Technical report, Max Planck Institute for Biological Cybernetics, Tubingen, Germany, 2006.

[193] W. A. Sethares. *Tuning, timbre, spectrum, scale*. Springer, 2nd ed. edition, 2005. Includes bibliographical references (p. [381]-397) and index.

[194] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 18:1–111, 2012.

[195] H S Seung, M Opper, and H Sompolinsky. Query by committee. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 287–294, 1992.

[196] K Seyerlehner, G Widmer, and P Knees. Frame level audio similarity - a codebook approach. *Conference on Digital Audio Effects*, pages 1–8, 2008.

[197] J. Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[198] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[199] E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-Inputs. *Advances in neural information processing*, 2006.

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

[201] C. Song, Z. Qu, N. Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[202] C. Song, Z. Qu, N. Blumm, and Albert-László Barabási. Limits of predictability in human mobility - supplementary material, 2010.

[203] M. Sordo, Òscar Celma, Martin Blech, and Enric Guaus. The quest for musical genres: Do the experts and the wisdom of crowds agree? In *9th International Conference on Music Information Retrieval*, pages 255–260, 2008.

[204] M. Sordo, F. Gouyon, L. Sarmento, and Ò. Celma. Inferring semantic facets of a music folksonomy with wikipedia. *Journal of Web Semantics*, Submitted.

[205] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

[206] S. Stober and N A. Towards User-Adaptive Structuring and Organization of Music Collections. pages 53–65, 2010.

[207] S. Stober and A. Nürnberger. An experimental comparison of similarity adaptation approaches. In *International Society for Music Information Retrieval Conference*, 2011.

[208] S. Stober and A. Nürnberger. Adaptive music retrieval - a state of the art. *Multimedia Tools and Applications*, March 2012.

[209] B.L. Sturm. A survey of evaluation in music genre recognition. *Adaptive Multimedia Retrieval*, 2012.

[210] S. Sundaram and S. Narayanan. Audio retrieval by latent perceptual indexing. *International Conference on Acoustics, Speech and Signal Processing*, pages 49–52, 2008.

[211] R. S. Taylor. The process of asking questions. *American Documentation*, 13(4):391–396, October 1962.

[212] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[213] H. Terasawa, M. Slaney, and J. Berger. A timbre space for speech. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 1729–1732. ISCA, 2005.

[214] William R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935.

[215] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[216] L L Thurstone. A law of comparative judgement. *Psychological Review*, 34, 1927.

[217] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 2001.

[218] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(1):45–66, 2002.

[219] Intelligent Sound Processing Toolbox. `http://kom.aau.dk/project/isound/`, 2012. [Online; accessed 17-12-2012].

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

[221] D. Turnbull, Ruoran Liu, Luke Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. *ISMIR*, 2008.

[222] J. Vanhatalo. *Speeding up the inference in Gaussian process models*. PhD thesis, Alto University, 2010.

[223] J. Vanhatalo, P. Jylänki, and A. Vehtari. Gaussian process regression with Student-t likelihood. In Y Bengio, D Schuurmans, J Lafferty, C K I W.s, and A Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1910–1918. 2009.

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

[225] J. V. Verkuilen. *Regression Models for Paired Comparisons*. PhD thesis, University of Illinois at Urbana-Champaign, 2007.

[226] C. Walder, Kwang K., and B. Schölkopf. Sparse multiscale gaussian process regression. *International Conference on Machine Learning*, pages 1112–1119, 2008.

[227] H. M. Wallach, I. Murray, Ruslan Salakhutdinov, and D. Mimno. Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, (d):1–8, 2009.

[228] Hanna M. Wallach. *Structured topic models for language*. PhD thesis, University of Cambridge, 2008.

[229] J Weston, S Bengio, and P Hamel. Multi-Tasking with Joint Semantic Spaces for Large- Scale Music Annotation and Retrieval Multi-Tasking with Joint Semantic Spaces for Large-Scale Music Annotation and Retrieval. *Journal of New Music Research*, (November 2012):37–41, 2011.

[230] T.D. Wickens. *Elementary signal detection theory*. Oxford University Press, Inc., 2002.

[231] C.I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

[232] A. Wilson and Z. Ghahramani. Copula Processes. *Advances in Neural Information Processing Systems 23*, pages 2460–2468, 2010.

[233] Ole Winther. *Bayesian Mean Field Algorithms for Neural Networks and Gaussian Processes*. PhD thesis, 1998.

[234] D. Wolff and T Weyde. Adapting similarity on the MagnaTagATune database: effects of model and feature choices. *International World Wide Web Conference*, 2012.

[235] D Wolff, T Weyde, S. Stober, and A Nürnberger. A systematic comparison of music similarity adaptation approaches. *13th International Society for Music Information Retrieval Conference*, pages 103–108, 2012.

[236] F. Xia, J. Wang, W. Zhang, T. Liu, and H. Li. Listwise approach to learning to rank - theory and algorithm. *ICML 2008 - Proceedings of the 25th International Conference on Machine Learning*, pages 1192–1199, 2008.

[237] Yi-Hsuan Yang and Homer H Chen. Ranking-based emotion recognition for music organization and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):762–774, 2011.

[238] G. Yao and U. Bockenholt. Bayesian destimation of thurstonian ranking models based on the gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1):79, 1999.

[239] K. Yoshii and M Goto. Continuous pLSI and smoothing techniques for hybrid music recommendation. *International Society for Music Information Retrieval Conference*, pages 339–344, 2009.

[240] K. Yoshii, M Goto, and K Komatani. Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. *International Society for Music Information Retrieval Conference*, 2006.

[241] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. *Proceedings of the international ACM SIGIR conference on Research and development in Information Retrieval*, pages 271–278, 2007.

[242] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine, 2001.

[243] E. Zwicker and H. Fastl. Psychoacoustics: Facts and models. *Springer Series in Information Sciences*, 22(2), 1999.