Attention, Perception, & Psychophysics (2018) 80:798–812
https://doi.org/10.3758/s13414-017-1460-0

Bayesian active probabilistic classification for psychometric field estimation

Xinyu D. Song[1] · Kiron A. Sukesan[1] · Dennis L. Barbour[1]

Published online: 18 December 2017

Abstract
Psychometric functions are typically estimated by fitting a parametric model to categorical subject responses. Procedures to estimate unidimensional psychometric functions (i.e., psychometric curves) have been subjected to the most research, with modern adaptive methods capable of quickly obtaining accurate estimates. These capabilities have been extended to some multidimensional psychometric functions (i.e., psychometric fields) that are easily parameterizable, but flexible procedures for general psychometric field estimation are lacking. This study introduces a nonparametric Bayesian psychometric field estimator operating on subject queries sequentially selected to improve the estimate in some targeted way. This estimator implements probabilistic classification using Gaussian processes trained by active learning. The accuracy and efficiency of two different actively sampled estimators were compared to two non-actively sampled estimators for simulations of one of the simplest psychometric fields in common use: the pure-tone audiogram. The actively sampled methods achieved estimate accuracy equivalent to the non-actively sampled methods with fewer observations. This trend held for a variety of audiogram phenotypes representative of the range of human auditory perception. Gaussian process classification is a general estimation procedure capable of extending to multiple input variables and response classes. Its success with a two-dimensional psychometric field informed by binary subject responses holds great promise for extension to complex perceptual models currently inaccessible to practical estimation.

Keywords Psychometrics/testing · Psychoacoustics · Audition

Introduction

In psychophysical studies of perception, a psychometric function describes the dependence of an individual's task performance on the physical properties of presented stimuli. Estimation of psychometric functions has a distant origin (Fechner, 1860) with predominant focus in the years since on one-dimensional (1D) input domains, which result in psychometric curves (PCs). PCs are typically modeled as sigmoidal functions with two parameters of interest: threshold α, which describes the stimulus level required to reach a certain detection probability, and spread β, which describes the rate at which performance increases with increasing stimulus level around threshold (Kingdom & Prins, 2016).

A widely applied technique for estimation of PCs is the method of constant stimuli (Fechner, 1860; Treutwein, 1995), in which the psychometric input domain is divided into a fixed number of equally spaced stimuli, each of which is presented in a pseudorandom order with some number of repetitions. The subject's responses to all of these stimuli are used to fit the PC. While effective and accurate, this method is time-consuming in practice. To combat this problem, a number of adaptive techniques (Leek, 2001; Kujala, 2011) have been developed over the years to select input values sequentially for rapid estimation of a psychometric threshold (King-Smith et al., 1994;

Significance Perceptual and cognitive testing have always faced severe constraints in accruing sufficient data from an individual to estimate a psychometric function, which represents a probabilistic prediction of behavior. By embracing, exploiting, and quantifying the interactions among underlying psychological phenomena, we have developed a universal method of efficiently estimating multidimensional psychometric functions. Evaluations using the pure-tone audiogram reveal that this method can achieve its potential to reframe how behavioral queries are used to infer individual perceptual and cognitive processes with limited data.

✉ Dennis L. Barbour
dbarbour@wustl.edu

[1] Laboratory of Sensory Neuroscience and Neuroengineering, Department of Biomedical Engineering, Washington University in St. Louis, 1 Brookings Drive, Box 1097, St. Louis, MO 63130, USA

Levitt, 1971; Pentland, 1980; Taylor & Creelman, 1967) or of an entire PC (Gubner, 2006; Hall, 1981; Kontsevich &Tyler, 1999; Shen & Richards, 2012; Watson & Pelli, 1983).

Although research in psychophysics has primarily focused on the 1D case, many if not most real-world perceptual phenomena are inherently multidimensional and would most effectively be modeled as multidimensional psychometric fields (PFs). One simple example is pure-tone audiometry, a two-dimensional (2D) test in which auditory detection thresholds are assessed using pure tones of various intensities and frequencies. Traditionally, pure-tone audiometry has been reported using a threshold audiogram, in which adaptively sampled thresholds are reported for 6-9 discrete frequencies (Carhart & Jerger, 1959; Hughson & Westlake, 1944). The PF across all frequencies could provide additional useful information about hearing but is rarely if ever estimated in practice, presumably because either poor models exist or accumulating sufficient data to fit a model is prohibitive, or both. For example, children appear to have higher internal noise than adults in acoustic detection and discrimination tasks, though the multiple hours necessary to acquire data in one person hinder research into discerning other conditions under which this observation might be true (Allen & Wightman, 1994; Bargones et al., 1995; Buss et al., 2006, 2009).

Although adaptive sampling techniques are well established for PCs, limited techniques exist for multidimensional cases. In addition to psychometric input variables, in which response probability increases with increasing intensity, PFs often contain one or more non-psychometric input variables, against which response probability does not systematically increase. Therefore, the PFs would not typically be straightforward extensions of PCs, and the nonlinear interactions of additional variables must be modeled differently. A limited number of techniques have demonstrated the ability to perform efficient inference on PFs, including auditory filters (Shen & Richards, 2013), contrast external noise functions (Lesmes et al., 2006), visual fields (Bengtsson et al., 1997), or even more complex visual models (DiMattina, 2015). The PFs in such cases are parameterized and can therefore be estimated in typical fashion by parametric regression. In the case of the pure-tone audiogram, however, the corresponding PF includes a non-psychometric frequency input variable for which any particular parametric justification is weak.

The problem of estimating PFs has recently been addressed using probabilistic classification (Song et al., 2017), which makes use of machine learning techniques for categorical prediction (Hastie et al., 2009) that can deliver probability distributions over sets of discrete outcomes such as a subject's behavioral responses. This previous work demonstrated the ability of particular nonparametric models to estimate audiometric PFs with high accuracy but utilized fixed numbers of deterministic samples to form PF estimates; therefore, their efficiency potential was not fully assessed. Additional previous work has shown that active sampling techniques within the same probabilistic classification framework can rapidly estimate audiometric thresholds in human subjects (Gardner et al., 2015b; Song et al., 2015). The current work seeks to evaluate the accuracy and efficiency of probabilistic classification with active sampling for estimating entire PFs of several audiometric phenotypes.

Mathematical background

Stochastic process models

The goal of PC estimation in psychophysics is to form a model ψ of task performance as a function of scalar input variable x that reflects a physical process. An example would be the probability of detecting a flash of light or burst of sound as a function of stimulus intensity. Typically, $\psi(x)$ would be expressed as a parametric function with parameter values determined through logistic or probit regression (Kingdom & Prins, 2016; Treutwein, 1995). Extending this methodology to more input variables (i.e., PF estimation) would traditionally require explicit parametric representation of each individual input domain, denoted collectively by $\mathbf{x}$.

An alternate formulation for PF estimation is to consider the model $\psi(\mathbf{x})$ to be a stochastic process representing beliefs about task performance as a function of input (Song et al., 2017). For the i^{th} input value $\mathbf{x}_i$, the model output $\psi_i = \psi(\mathbf{x}_i)$ follows a distribution reflecting belief about PF output at $\mathbf{x}_i$. A Gaussian process (GP) is a particular type of stochastic process that assumes output values are drawn from a multivariate Gaussian distribution. As a result, a GP can be effectively described by a mean function $\mu(\mathbf{x})$ and a covariance function or kernel $K(\mathbf{x}, \mathbf{x}')$. Formulating the problem as the estimation of a GP has a great advantage in shifting inference from the model function itself to (potentially parametric) moment functions of a well-characterized stochastic process (Williams & Rasmussen, 1996; Williams, 1998). As will be demonstrated in this manuscript, recasting the traditional inference problem in this way has the flexibility of traditional nonparametric estimation and, with appropriate kernel specification and sampling strategies, the efficiency of traditional parametric estimation. One caveat is that because ψ reflects probability of discrete task outcomes and is therefore bounded, it cannot be Gaussian distributed. The full GP model therefore requires a mapping between intermediate Gaussian variables and the function output.

The following sections offer a relatively succinct overview of GP models. (Song et al., 2017) provides substantially more detail on GPs as they relate to psychophysical inference.

Gaussian process models

Let $f(\mathbf{x})$ be a latent function defined on an arbitrary continuous input space $\mathbf{x} \in \chi$ that is to be estimated. A GP is a convenient mechanism to encode prior knowledge about f, which can be updated given observed data using Bayesian inference. By definition, a GP is a collection of random variables, any finite subset of which jointly form a multivariate Gaussian distribution. Like the multivariate Gaussian distribution, a GP is fully specified by its first two moments: a mean function $\mu(\mathbf{x})$, which describes the central tendency of f, and a positive semidefinite covariance function $K(\mathbf{x}, \mathbf{x}')$, which describes the correlation structure of f about the mean. Given a choice of μ and K, the latent function f can be endowed with a GP prior distribution:

$$p(f) = \mathcal{GP}(\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}')). \tag{1}$$

Particular mean and covariance functions can be selected to capture desired latent function characteristics, such as smoothness, periodicity, linearity, or chaos (Duvenaud, 2014; Rasmussen & Williams, 2006).

A GP prior defines a belief about f over a continuous domain $\mathbf{x}$. The GP is typically evaluated only on a finite set of inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, which will take the form of an n-dimensional multivariate Gaussian distribution for this case. As observations are made, this prior distribution can be used to compute a posterior distribution incorporating the new information, typically used to perform prediction about f at a new input point $\mathbf{x}^*$.

Binary classification using GPs

In traditional regression problems, the i^{th} continuous-valued output observation $y_i = y(\mathbf{x}_i)$ is considered to be a reflection of the underlying latent function value f corrupted by additive observation noise, such that $y_i = f(\mathbf{x}_i) + \varepsilon_i$. In classification problems, however, the latent function relates to class membership, and observations reflect discrete outcomes in which an observed output y_i can take only a finite set of N categorical values $C_1, C_2, \ldots C_N$. One common type of classification problem particularly relevant for psychometric theory is binary classification, in which $y_i \in \{0, 1\}$. Many psychometric task outcomes (e.g., "detected vs. undetected" or "correct vs. incorrect" or "same" vs. "different") represent binary classifications. The terminology used here will be "success" and "failure."

Because probabilities can only take on values in the range [0,1], $f \in \mathbb{R}$ must therefore be transformed using a monotonic link function. The probability of success $p(y = 1|f)$ is generated by transforming the latent function f using a sigmoidal likelihood function Φ such that the resulting values lie in the interval [0,1]. A convenient choice of Φ consistent with standard psychometric theory (Kingdom and Prins, 2016) is the cumulative Gaussian or probit function.

The advantages of the GP classification framework begin to become apparent when contemplating how to refine the model with observations, developed here for multiple independent inputs and a single binary output. Consider a set of observed or "training" data $\{\mathbf{X}, \mathbf{y}\}$, where each element $y_i \in \{0, 1\}$ in $\mathbf{y}$ is an observed value at the corresponding multidimensional input location $\mathbf{x}_i$ in $\mathbf{X}$. With a GP prior, Bayes's theorem can be used to compute the posterior distribution of latent function values $\mathbf{f}$ at input values $\mathbf{X}$:

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}}. \tag{2}$$

Suppose an estimate of success probability at an unknown test input $\mathbf{x}^*$ is desired. The predictive distribution for the corresponding latent variable f^* is computed by marginalizing over the training set latent function values $\mathbf{f}$:

$$p(f^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*) = \int p(f^*|\mathbf{X}, \mathbf{f}, \mathbf{x}^*)p(\mathbf{f}|\mathbf{X}, \mathbf{y})d\mathbf{f} \tag{3}$$

From here, the success probability can be computed by integrating over the unknown test latent function variable:

$$\begin{aligned} p(y^* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}^*) &= \int p(y^* = 1|f^*)p(f^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)df^* \\ &= \int \Phi(f^*)p(f^*|\mathbf{X}, \mathbf{y}, \mathbf{x}^*)df^* \end{aligned} \tag{4}$$

Because of the sigmoidal nonlinearity inherent in the likelihood, the integrals in Eqs. 2–4 are analytically intractable. The posterior distribution can be estimated, however, either by using a Gaussian approximation or by sampling techniques such as Markov chain Monte Carlo (MCMC) (Rasmussen & Williams, 2006). Methods for Gaussian approximation of the posterior distribution include expectation propagation (Minka, 2001) or Laplace's method (Williams & Barber, 1998). Laplace approximation attempts to approximate the posterior distribution by fitting a Gaussian distribution to a second-order Taylor expansion of the posterior around its mean. Expectation propagation attempts to approximate the posterior distribution by matching the first and second moments—the mean and variance—of the posterior distribution to those of a Gaussian. Expectation propagation is used in this study because of its compatibility with iterative queries and generally superior performance to Laplace's method. The result of this estimation procedure is a posterior belief about the probability of success given some prior beliefs and some observations. In the binary classification case considered here, success probability is Bernoulli distributed, $\psi(\mathbf{x}) \sim p(y|f, \mathbf{x}) = \text{Bernoulli}(\Phi(f(\mathbf{x})))$, while the latent function is defined by a Gaussian process, $f(\mathbf{x}) \sim p(f|\mathbf{x}) = \mathcal{GP}(\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}'))$. The estimate in this case is a stochastic process, consistent with Bayesian inference and the GP model.

Note that this GP formulation distinctly parallels previous work in psychometric estimation that has decomposed a PC model into "core" and "link" functions. The core function describes how underlying processes vary with input, and the link function is a sigmoidal function that transforms the core function into the probabilistic range [0, 1] (Kuss et al., 2005; Wichmann & Hill, 2001b). For the GP, the core function equivalent is the latent function f, which is defined using the GP prior, and the link function equivalent is the likelihood function Φ, which for classification can be one of a number of monotonic sigmoidal functions. Consistent with previous work, both components can be manipulated independently to encode different psychometric properties.

Hyperparameters

The choice of form for the GP mean and covariance functions critically determines the space of possible functions that can be inferred. These mean and covariance functions may have associated parameters that can also impact behavior of the predictive model in important ways. As parameters of the GP mean and covariance functions and not of the latent function itself, these entities are called hyperparameters and are denoted as $\boldsymbol{\theta}$.

Hyperparameter values can sometimes be chosen *a priori*, which assumes perfect or compelling knowledge of latent function behavior as it relates to those hyperparameters. If uninformative prior beliefs are assigned to the hyperparameters, as is done in this study, a set of hyperparameters $\hat{\boldsymbol{\theta}}$ can be learned after observing some data by maximizing the marginal likelihood of the observations $(\mathbf{X}, \mathbf{y})$ given the hyperparameters: $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta})\ p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta})\ d\mathbf{f}$. This maximization is often analytically intractable and is typically approximated using numerical techniques such as gradient descent (Bishop, 2006; Rasmussen & Williams, 2006). The resulting hyperparameters are a set that best fits the observed data, an application of the second level of hierarchical Bayesian inference (Jefferys & Berger, 1992; Xiang & Fackler, 2015).

Active sampling

PC and PF estimation can make use of deterministic sampling, where a predetermined set of stimuli is presented to the subject. Two examples of deterministic sampling schemes are the method of constant stimuli, which uses m repetitions at n different input values (Fechner, 1860) and Halton sampling, which delivers a set of well-spaced values across the input domain (Halton, 1964; Song et al., 2017). The stimuli in a fixed set are generally not equally informative, however, and reducing the number of uninformative observations can dramatically reduce the time needed to obtain a reasonable estimate.

Active sampling (typically called "active learning" in the machine learning literature) describes the process of sequentially selecting the most informative samples to probe a function space efficiently (Gardner et al., 2015b; Settles, 2009; Song et al., 2015). Borrowing terminology from Bayesian optimization (Brochu et al., 2010; Guestrin et al., 2005; Osborne et al., 2009), active sampling can broadly be conceptualized as employing an acquisition function A that quantifies how desirable any particular sample will be based on a chosen metric. At each iteration, the input point $\hat{\mathbf{x}}^*$ corresponding to the highest value in the acquisition function $A(\mathbf{x}^*)$ is selected as the subsequent sample. Acquisition functions are typically derived using Bayesian decision theory acting on a utility function U. In such cases the acquisition function reflects the expected utility over the observations for the set of observations at each iteration, such that $\hat{\mathbf{x}}^* = \operatorname{argmax}_{\mathbf{x}^* \in \mathbf{X}^*} A(\mathbf{x}^*) = \operatorname{argmax}_{\mathbf{x}^* \in \mathbf{X}^*} U(\mathbf{x}^*|\mathbf{X}, \mathbf{y})$ (Chaloner & Verdinelli, 1995; Park, 2013). One standard example of a query framework is uncertainty sampling, where the acquisition function describes model uncertainty (Lewis & Catlett, 1994; Lewis & Gale, 1994). A number of established active sampling techniques exist (Settles, 2009), and the acquisition function can be specifically chosen or tuned depending on desired behavior.

This paper distinguishes active sampling methods from more traditional adaptive sampling methods in psychometric estimation, which include rule-based procedures such as the method of limits (Levitt, 1971; Treutwein, 1995). A variant of the method of limits is used prominently in hearing loss diagnosis (American National Standards Institute, 2004; Carhart & Jerger, 1959; Hughson & Westlake, 1944). Although iterative, adaptive sampling techniques follow rule-based heuristics dependent upon observed responses rather than quantifying utility across possible inputs in the sample space. Several advanced psychometric estimation procedures traditionally also referred to as "adaptive sampling" methods do select points that maximize or minimize a metric over the input domain, such as expected variance or expected information gain (King-Smith et al., 1994; Kontsevich & Tyler, 1999; Leek, 2001; Shen & Richards, 2013). It will be useful in the present context to adopt the phrases "adaptive sampling" for simpler heuristic methods and "active sampling" for methods with formally defined optimization steps in order to distinguish these two types of procedures.

Methods

The performance of the active probabilistic classification algorithm described above was evaluated in a particular physical domain. Following the approach of (Song et al., 2017), the

domain of interest was selected to be audiometry, or an individual's ability to hear pure tones as a function of tone frequency and intensity. This choice of modality was made for several reasons: (1) hearing ability measurement is very common, with basic hearing measurement taking up an estimated 1.9 million hours each year across all audiological professionals (Margolis & Morgan, 2008); and (2) audiometric functions exhibit nonlinear interactions and can vary widely as a function of frequency, making them difficult to parameterize effectively (Dubno et al., 2013; Özdamar et al., 1990).

Current clinical hearing measurement primarily uses the modified Hughson-Westlake (HW) method (Carhart & Jerger, 1959; Hughson & Westlake, 1944), a rule-based adaptive technique that proceeds iteratively across discrete frequencies and returns the 70.7 % threshold value on the 1D PC for a tested frequency (Levitt, 1971). Because only six to nine frequencies are typically tested, however, the method cannot form a continuous estimate of threshold, and no estimate of psychometric spread is achieved. Therefore, the full audiometric function is never inferred by this method.

Although the experiments described here are centered on a pure-tone audiometric detection task, the technique itself is general and applicable to numerous uni- and multidimensional psychometric estimation tasks.

Ground-truth models

Simulated audiometric functions were used for evaluation of the estimation algorithm. Construction of the ground truth model audiometric function $\psi_{true}(\mathbf{x})$ was done following the technique described in (Song et al., 2017). Threshold curves as a function of frequency were simulated by estimation of 1 of 4 human audiometric phenotypes (Dubno et al., 2013) via spline interpolation and linear extrapolation (Fig. 1). These phenotypes reflect human audiogram measures that best match particular etiologies of hearing loss confirmed in animal models. The ability to accurately estimate each audiogram phenotype can therefore influence the accuracy of hearing loss diagnoses, particularly when combined with demographic, genetic and/or physiological data.

At each frequency, a 1D sigmoidal psychometric curve was constructed using a standard cumulative Gaussian equation of the following form (Kingdom & Prins, 2016):

$$p(x) = \frac{1}{\sqrt{2\pi}\beta} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\alpha}{\beta}\right)^2} dt \tag{5}$$

where α defines the point of 50 % detection probability and β defines the psychometric spread. The audiogram threshold value at that frequency corresponded to the 70.7 % detection probability point along the PF (Levitt, 1971). Because little psychophysical data exist on tone-detection spread in humans (estimation of spread is not achieved in the typical Hughson-Westlake method), six different values for β between 0.2 and 10 dB/% were used for each ground truth audiometric function. The overall 2D PF $\psi_{true}(\mathbf{x}) = \psi_{true}(\omega, \iota)$, where ω denotes frequency and ι denotes intensity, combines the audiogram shape across frequency with the sigmoidal 1D PC in intensity, describing the detection probability $p(\mathbf{x}_i)$ for an input frequency/intensity pair $\mathbf{x}_i = (\omega_i, \iota_i)$, where i indexes the input tone.

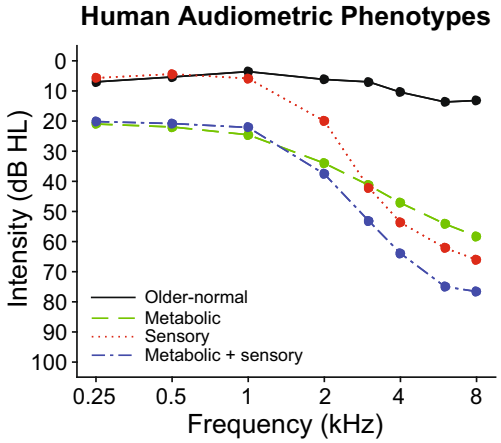

Fig. 1 The four audiogram phenotypes reported in (Dubno et al., 2013) used to construct full ground truth psychometric fields using spline interpolation and linear extrapolation

For a particular frequency/intensity input $\mathbf{x}_i$, the PF returns a detection probability $p(\mathbf{x}_i)$ corresponding to that queried stimulus. The binary response y_i at that point is generated by a single draw from a Bernoulli distribution with probability $p(\mathbf{x}_i)$, with detected and non-detected responses represented as values of 1 and 0, respectively (Treutwein, 1995).

Gaussian process setup

The GP framework was constructed following the procedure described in (Song et al., 2017). For the 2D inference problem, an estimate of a subject's detection probability as a function of the frequency and intensity $\mathbf{x} = (\omega, \iota)$ of pure-tone stimuli is desired. The quantity of interest is $p(y = 1|\mathbf{x})$, and a GP prior is placed on the latent function: $f \sim \mathcal{GP}\left(\mu(\mathbf{x}), K\left(\mathbf{x}, \mathbf{x}'\right)\right)$, which is transformed using the sigmoidal observation model:

$$p(y = 1|f) = \int_{-\infty}^{f} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \tag{6}$$

The choice of covariance function can be informed with substantial prior information. Consistent with standard psychometric models, response probability is modeled as monotonically increasing with increasing tone intensity. A linear covariance function $K\left(\iota, \iota'\right) = s_1^2\left(\iota, \iota'\right)$ in intensity was selected for the intensity domain. This selection constrains the latent function f to only linear functions in ι, which takes the

form of a cumulative Gaussian function when combined with the observation model in Eq. 6 (Song et al., 2017).

Unlike for intensity, however, the dependence of detection probability on frequency is not explicit and can vary widely depending on the audiogram shape. A reasonable observation is that the PF as a function of frequency is continuous and generally smooth according to some length scale. That is, observations nearby in frequency are more likely to be correlated than distant points (Brant & Fozard, 1990; Carhart & Jerger, 1959; Kiang et al., 1965; Leek, 2001; Levitt, 1971; Liberman, 1978; Geisler et al., 1985; Von Békésy, 1960). Therefore, the common squared exponential (SE) covariance function $K(\omega,\omega') = s_2^2 \exp\left[\frac{-(\omega-\omega')^2}{2\ell^2}\right]$ was applied in the frequency dimension. The full covariance function combines the linear covariance function in intensity and the SE covariance function in frequency:

$$K(\mathbf{x},\mathbf{x}') = K\left((\omega,\iota),(\omega',\iota')\right) = s_1^2(\iota\cdot\iota') + s_2^2\exp\left[\frac{-(\omega-\omega')}{2\ell^2}\right] \quad (7)$$

In Eq. 7, s_1 and s_2 are scaling factors and ℓ is a length scale that determines how much correlation drops off with distance in frequency. A constant mean function $\mu(\mathbf{x}) = c$ was selected because any deviation from the mean function can be captured efficiently in the covariance function.

The posterior distribution $p(\mathbf{f}^*|\mathbf{X},\mathbf{y},\mathbf{X}^*)$ was computed using a dense grid of test points $\mathbf{X}^*$ across frequency/intensity space: 0.125 to 16 kHz in semitone increments for frequency, and −20 to 120 dB in 1-dB increments for intensity.

Hyperparameter learning

The current work utilized uninformative prior information regarding hyperparameters, which makes no assumptions about the values of the hyperparameters before data are observed. Formally, this scenario can be viewed as a uniform prior across all values on each hyperparameter, where all values of hyperparameters are considered equally likely before any data is observed. This choice of prior was intentional in order to demonstrate that the active sampling method could converge to accurate estimates even when launched without the benefit of known audiogram details.

Rather than explicit specification of hyperparameter values *a priori*, the best-fitting set of hyperparameters $\boldsymbol{\theta} = \{c, s_1, s_2, \ell\}$ was computed on each iteration by maximizing the log marginal likelihood with respect to the observed samples $(\mathbf{X},\mathbf{y})$ (Rasmussen & Williams, 2006; Song et al., 2017). In all cases, gradient descent was used to numerically estimate the best-fitting hyperparameters. Following computation of the hyperparameters, the posterior distribution $p(\mathbf{f}^*|\mathbf{X},\mathbf{y},\mathbf{X}^*,\boldsymbol{\theta})$ was computed by expectation propagation.

The gradient descent procedure can be somewhat sensitive to its starting location, particularly for non-convex surfaces. To improve the stability of the hyperparameter learning technique, gradient descent was performed twice on each iteration from two distinct starting locations: (1) the best-fitting set of hyperparameters from the previous iteration, and (2) a random set of hyperparameters drawn from a Gaussian distribution centered at $\mathbf{0}$. Of the two gradient descent runs, the better-fitting result, i.e., the set of hyperparameters with higher marginal likelihood $p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$, was selected as the set of learned hyperparameters for that iteration.

Sampling methods

The efficacy of several sampling techniques within the GP classification framework was evaluated. In total, four techniques were evaluated: one random sampling scheme, one deterministic sampling scheme, and two active sampling schemes. Each is described in more detail below.

Random sampling On each iteration, the sample point selected is determined at random from the set of all possible grid points. This technique reflects a useful efficiency baseline to which each other sampling technique can be compared.

Halton sampling On each iteration, the sample point chosen is drawn from the corresponding ordered value in a Halton sequence, which provides a well-spaced deterministic set of draws on some predefined interval (Halton, 1964). Previous work has utilized this approach to effectively estimate PFs (Song et al., 2017).

Bayesian active learning by variance (BALV) In the context of uncertainty sampling, one convenient (if greedy) measure of uncertainty is the model variance (Settles, 2009). Within the GP framework in this case, the acquisition function is the posterior variance σ_y^2, i.e. $A(\mathbf{x}^*) = \sigma_y^2(\mathbf{x}^*)$, which is inversely related to model confidence and maximized at points where the label probability is closest to 0.5. Previous work has demonstrated that BALV can produce estimates of audiogram thresholds consistent with those produced by the HW approach (Song et al., 2015), and the current work extends this technique to inference of the entire PF.

Bayesian active learning by disagreement (BALD) BALD is another uncertainty sampling approach that seeks the sample point producing the largest decrease in expected posterior entropy; this approach has been shown to yield the sample for which individual configurations of an underlying latent function most disagree about the outcome (Houlsby et al., 2011).

The acquisition function in this case is the decrease in posterior entropy given a set of training observations $(\mathbf{X}, \mathbf{y})$ and a new observation $(\mathbf{x}^*, y^*)$:

$$A(\mathbf{x}^*) = H[f \mid \mathbf{X}, \mathbf{y}] - \mathbb{E}_{y^*}\left[H\left[f \mid \mathbf{X}, \mathbf{y}, \mathbf{x}^*, y^*\right]\right] \quad (8)$$

also known as the mutual information between y^* and the latent function f (Cover & Thomas, 2012). The BALD technique has been used previously to infer multidimensional neural receptive field and fMRI structure (Park et al., 2011; Park, 2013) as well as audiogram thresholds (Gardner et al., 2015b), which is extended here to infer the entire PF. Furthermore, posterior entropy minimization techniques have been used to estimate thresholds and spreads for parametric uni- and multidimensional psychometric functions (DiMattina, 2015; Doire et al., 2017; Kontsevich & Tyler, 1999; Watson, 2017). The current work applies this sampling framework for estimation of nonparametric PFs.

Heuristics Following computation, all values in the acquisition function were normalized to the range 0 and 1. For both BALV and BALD, zero-mean Gaussian noise with a small acquisition variance σ_h^2 was added to each value in the normalized acquisition function. This heuristic generally improved sampling results by introducing small perturbations to the acquisition maximization procedure, which reduced repeated sampling within small, confined regions of the input domain yet still allowed acquisition function magnitude to dominate for sample selection. Because of the acquisition function normalization, the value for noise variance effectively represents a percentage of the full acquisition function range. An acquisition variance value of $\sigma_h^2 = 0.2$, reflecting zero-mean noise with variance equal to 20 % of the full acquisition range, was effective at preventing excessively concentrated samples for active sampling and was used for all BALV and BALD runs. Examples of excessive clustering at extremes of frequency can be observed in (Song et al., 2015). To fully resolve this issue, future acquisition function design should formally compensate for finite input domains.

Evaluation Performance was evaluated using several model and sampling parameters:

- *Audiogram shapes*: Older-normal, sensory, metabolic, and sensory + metabolic audiogram phenotypes (Dubno et al., 2013) were used to generate α values across frequency.
- *Spread value:* β values of 0.2, 0.5, 1, 2, 5, and 10 dB/percent, assumed isotropic across all frequencies, were used to determine spread.
- *Sampling method:* Random sampling, Halton sampling, BALV, and BALD were chosen as distinct sampling strategies within the GP framework.
- *Sample number:* To identify the effect of number of observed responses on model performance, sampling was conducted iteratively up to 100 observations, and the performance of each sample count (1–100) was evaluated.
- *Simulation repetition:* For each distinct parameter set, ten independent repetitions of GP inference were conducted.

The combination of all model and sampling configurations resulted in 96,000 total simulations.

Model prediction was assessed in several ways. At each frequency on a fine grid (0.25–8 kHz in semitone increments), the α and β value of a 1D PC can be derived by finding the x-intercept and inverse slope of the latent function f and can be compared to the α and β values of the known true PC at the corresponding frequency. Edge frequencies (0.125–0.25 and 8–16 kHz) were used to train the GP but not to evaluate prediction due to known edge effects of PF estimation (Song et al., 2015, 2017). Accuracy was evaluated using mean deviation of parameter estimates from the true value, and reliability was evaluated using the variance of model estimates across all repetitions with the same parameter values and sampling schemes. Furthermore, the model's estimate for the overall PF can be computed by passing f through the observation model $p(y|f)$ and can then be numerically compared to the true PF $\psi_{true}(\mathbf{x})$.

Goodness of fit of the model predictions to the observed data was evaluated using the Pearson χ^2 statistic: $\chi^2 = \sum_i \frac{N_i[p(\mathbf{x}_i) - P_i]^2}{P_i(1-P_i)}$. For any frequency/intensity pair $\mathbf{x}_i = (\omega_i, \iota_i)$, $p(\mathbf{x}_i)$ is the proportion correct for the data, P_i is the proportion correct for the model, and N_i is the number of observations at that input $\mathbf{x}_i$ (Klein, 2001; Wichmann & Hill, 2001a). This statistic can be interpreted as a weighted sum of squared residuals, with larger values representing poorer fits. Significance was evaluated by comparing the χ^2 statistic for any frequency/intensity pair $\mathbf{x}_i = (\omega_i, \iota_i)$ to the chi-squared distribution with J degrees of freedom, where J is the number of distinct frequency/intensity pairs sampled.

Results

In 174 simulations out of 96,000 (0.18 %), a numerical issue with the algorithm prevented it from executing properly. These failed simulations were therefore removed, with the remaining 95,826 successful simulations used in all following analysis. Any observation that would have been provided by a failed simulation was replaced with a query at a random input point. Furthermore, a small number of simulations (1625 of 95,826) generated outliers due to poor gradient descent convergence. Because of their disproportionate influence on the mean trends, these simulations were omitted from the

averaged data by thresholding at the 98th percentile. Similarly, simulations containing fewer than 15 observed sample points were not displayed in graphic summaries due to generally poor performance, particularly for random and Halton sampling.

A representative run of GP inference after 100 BALD samples can be seen in Fig. 2. In this example, ground truth was a sensory phenotype with β = 2 dB/%. Figure 2A shows the distribution of samples collected in frequency-intensity space for this particular run. Note that the samples are largely concentrated within a band around putative threshold where they would be particularly useful at refining the PF estimate. The posterior mean after 100 iterations is shown in the background. Figure 2B shows the estimated α curve across frequency compared to the true α curve. After 100 samples, the BALD procedure produces a continuous estimate of threshold as a function of frequency that closely matches ground truth. Figure 2C shows a slice of the model PF (the 1D PC) at ω = 1 kHz compared to the ground truth slice at that frequency. The close match of these two curves indicates that the BALD estimate of β matches ground truth as well as α. The agreement in spread extends across frequency, as well.

The BALD sampling method produces desired estimator behavior by selecting tones near detection threshold. Its relative performance can be seen in Fig. 3, which shows representative single GP estimation runs (posterior mean and selected samples) for each sampling method after 100 samples. The ground truth PF is a metabolic + sensory phenotype with β = 2 dB/%, shown in the inset. BALD (Fig. 3A) and BALV (Fig. 3B) sampling schemes show similar sample selection around putative threshold, with the estimated spread values for BALD more closely resembling ground truth. Halton sampling (Fig. 3C) selects relatively well-spaced draws spanning the entire input space, and samples are therefore not densely populated along the threshold line compared to BALD and BALV. Random sampling (Fig. 3D) shows a predicted threshold curve that noticeably diverges from the ground truth PF for these 100 samples.

The relative tendency of any point in frequency/intensity space to be queried for a particular sampling scheme can be visualized using mean acquisition maps, which are shown in Fig. 4 for the same ground truth PF as in Fig. 3. The acquisition map is constructed by averaging the acquisition function values across all simulations for each sampling method, with higher values corresponding to input locations more likely to be queried. BALD and BALV (Figs. 4A and 4B) sampled densely around threshold, with BALV appearing somewhat more tightly distributed near the threshold curve. Because Halton sampling (Fig. 4C) is deterministic, only a small number of points have high acquisition function values. Random sampling (Fig. 4D) results in a predictably random acquisition map.

Previous tests of BALV in humans revealed unbiased estimates of the 0.707 threshold point relative to the Hughson-Westlake procedure (Song et al., 2015). The most parsimonious explanation for this result is simultaneously unbiased 0.5 threshold estimates and psychometric spread estimates. The average signed error (estimated – actual) of each sampling method and each measure (i.e., alpha, beta and posterior mean) in the current study is summarized in Table 1. All sampling methods have mean signed errors near 0 for α and posterior mean and alpha when averaged across all experiments. At smaller spread values, β estimates likewise exhibit little or no bias. At higher spread values, however, all methods on average tended to underestimate the spread with 100

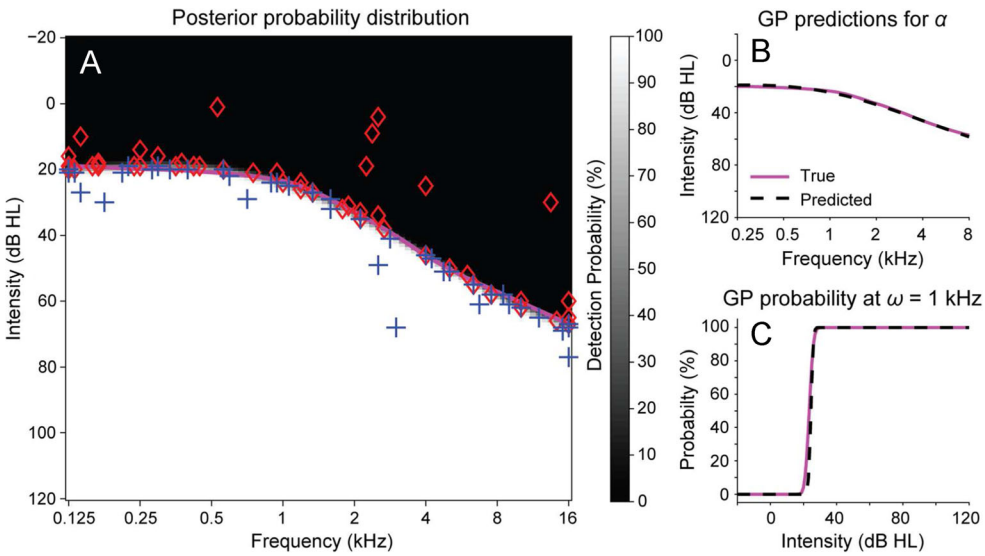

Fig. 2 Representative example of GP inference on a metabolic phenotype with β = 1 dB/%. (A) Posterior mean overlaid with samples and ground truth α threshold curve. (B) The GP estimate of α (dashed line) as compared to the true values of α (solid line). (C) A slice of the estimated psychometric field at ω = 1 kHz (dashed line), compared with a slice of the ground truth psychometric field at that frequency (solid line). BALD was used as the sampling method. Blue plus symbols denote detected stimuli; red diamond symbols denote undetected stimuli

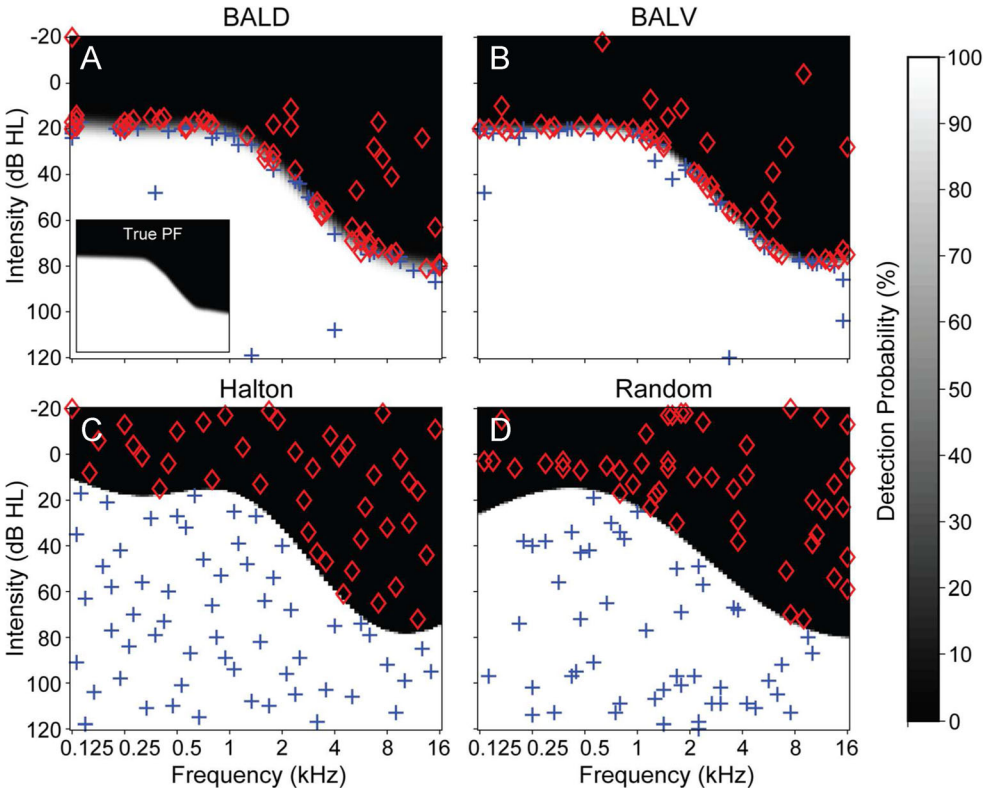

Fig. 3 Representative posterior mean and samples after 100 iterations for a metabolic + sensory phenotype with $\beta = 2$ dB/% and for each sampling method: (A) BALD sampling; (B) BALV sampling; (C) Halton sampling; and (D) random sampling. Inset in (A) shows the ground truth psychometric field used to generate all responses. Blue plus symbols denote detected stimuli; red diamond symbols denote undetected stimuli

samples. BALD generally appeared to have the lowest signed errors of the four sampling methods, and all methods showed less bias in β with more samples. Spread values for human auditory detection are likely to fall toward the lower end of the values tested here (i.e., $\beta = 0.5$–1), which could explain why the offset-threshold of 0.707 was able to be estimated so accurately in humans.

The mean unsigned error of each method was evaluated at each number of samples. Figure 5 shows the overall performance of each sampling method as a function of sample number averaged across all phenotypes, true β values, and repetitions. Metrics evaluated were unsigned error in threshold α (Fig. 5A), unsigned error in spread β (Fig. 5B), and mean pointwise absolute difference between predicted and ground truth probability surfaces (Fig. 5C), a nonparametric comparison. All methods on average improved their mean estimate accuracy with more sample iterations. Active methods strongly outperformed non-active methods for α estimation and mean pointwise difference, with trends in β being more consistent between all 4 sampling types. For α prediction, active sampling methods were within clinical reliability criterion (±5 dB) across all frequencies within approximately 20 tones; non-active methods required approximately 50–60 tones to achieve the same criterion. Performance of β estimation for BALV appeared to saturate with fewer samples than BALD. Perhaps this result occurs because BALV tends to sample near the 50 % probability boundary, thereby providing little information about spread in the probabilistic class boundary.

The influence of different audiogram phenotypes on estimator performance is evaluated in Fig. 6, which shows the prediction error of each sampling method for each audiogram profile separately, averaged across true β values and repetitions. Performance is generally similar for the different phenotypes, with the active sampling methods consistently outperforming the non-active sampling methods. In some cases, such as older-normal, active sampling can achieve clinically relevant threshold accuracy of 5 dB in fewer than 15 samples. Hearing loss phenotypes appear to require more iterations for accurate threshold estimates, but are still able to achieve 5 dB accuracy in fewer than 30 samples using active methods.

Higher values of spread β introduced increased uncertainty in the PF transition zones. The effect of β value on estimator performance was evaluated in Fig. 7, which shows the unsigned prediction error for each sampling method for different β values averaged across different phenotypes and repetitions. The results demonstrate broadly decreasing accuracy for all metrics with increasing β. Active sampling methods generally

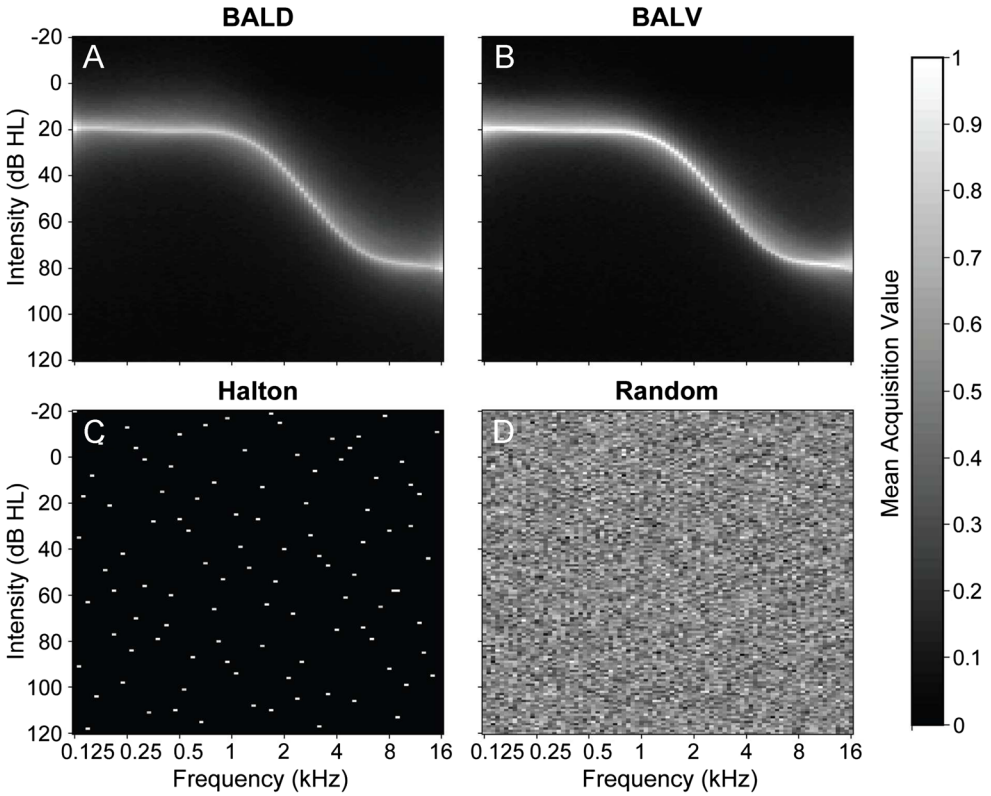

Fig. 4 Mean acquisition maps for each sampling technique on a metabolic + sensory phenotype with β = 2 dB/%, where each plot shows the normalized acquisition function map averaged across all iterations and repetitions for a single sampling method: (A) BALD sampling; (B) BALV sampling; (C) Halton sampling; and (D) random sampling

outperform non-active methods, with the potential exception of β inference using BALV for the highest spread value (Fig. 7B6). As described earlier, BALV tends to generate samples that are more effective at estimating threshold than spread.

Across all 95,826 valid simulations, the median χ^2 statistic was 5.03×10^{-5}. After using the chi-square distribution with the appropriate degrees of freedom to compute probability values for each simulation, nine of 95,826 simulations, or 0.0094 %, were detected to be statistically poor fits at a significance level of $p < 0.05$, Bonferroni corrected for multiple comparisons.

Discussion

This paper describes a method of accurately and efficiently estimating multidimensional PFs by combining Gaussian process probabilistic classification and active sampling. The current work extends previous efforts employing this estimation framework with deterministic sampling techniques (Song et al., 2017). Active sampling produces a marked efficiency increase for the estimation of complex PFs. Notably, active sampling methods required around 15–30 samples to reach the clinical reliability criterion of ±5 dB in audiometric

Table 1 Mean ± standard deviation signed errors (estimate – actual) of 2D GP posterior probability measures, in units of probability %, across all experiments for each sampling method

Measure being estimated	BALD	BALV	Halton	Random
Posterior Mean	–0.00997 ± 0.0172	0.0172 ± 0.509	0.123 ± 1.39	0.199 ± 1.61
α	0.0147 ± 0.719	0.0491 ± 1.01	–0.222 ± 1.98	–0.283 ± 2.28
β=0.2	0.0195 ± 0.159	–0.0119 ± 0.176	0.440 ± 1.18	0.134 ± 0.510
β=0.5	–0.115 ± 0.203	–0.149 ± 0.248	–0.118 ± 0.803	0.158 ± 0.755
β=1	0.0407 ± 0.431	–0.139 ± 0.511	–0.394 ± 0.953	–0.412 ± 1.15
β=2	–0.441 ± 0.596	–0.399 ± 0.826	–1.15 ± 1.25	–0.993 ± 1.42
β=5	–0.571 ± 1.57	–1.68 ± 2.15	–2.11 ± 2.52	–2.49 ± 2.57
β=10	–2.25 ± 3.86	–2.42 ± 5.02	–2.00 ± 3.78	–2.25 ± 4.17

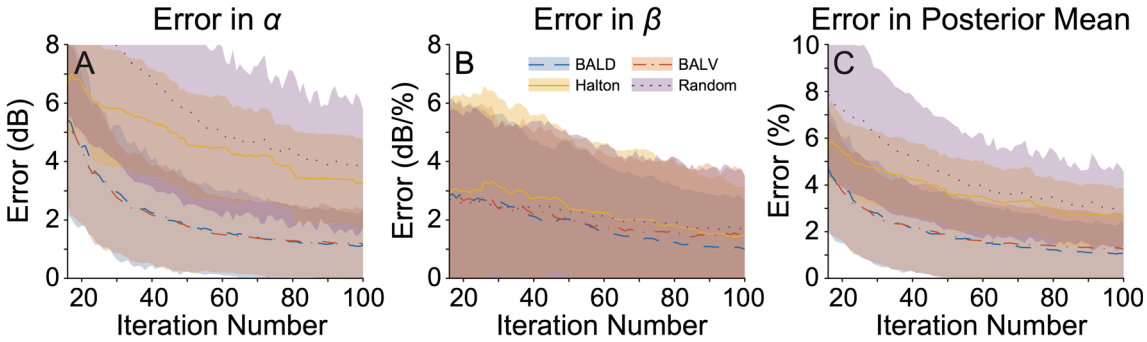

Fig. 5 Summary of GP estimator performance for each sampling method across all conditions and repetitions. Colored lines denote the mean, and shaded areas denote 1 standard deviation above and below the mean. Across iteration numbers (A) and (B) show unsigned differences (errors) between model predictions and ground truth for α and β, respectively, and (C) shows the mean unsigned difference in probability percent between the GP posterior mean and ground truth

threshold estimation (Fausti et al., 1990; Swanepoel et al., 2010), which is consistent with data from human subjects (Song et al., 2015). This sample number improves considerably upon the approximately 50–60 samples needed to reach the ±5 dB criterion for random and Halton sampling (Song et al., 2017), as well as upon the number of samples typically required for a complete Hughson-Westlake run, which often utilizes 100 or more samples per ear (Carhart & Jerger; Katz et al., 2009; Mahomed *et al.* 2013; Song et al., 2015).

Overall, non-active techniques (random and Halton) performed comparatively more poorly than active techniques, with random sampling generally performing the worse of the two. The two active sampling techniques (BALV and BALD) exhibited similar performance to one another, with both methods selecting samples near the putative threshold as expected. BALV demonstrated some difficulty with accurate β estimation for higher spread values (Fig. 7B6) compared to the other sampling techniques, however. This result can be

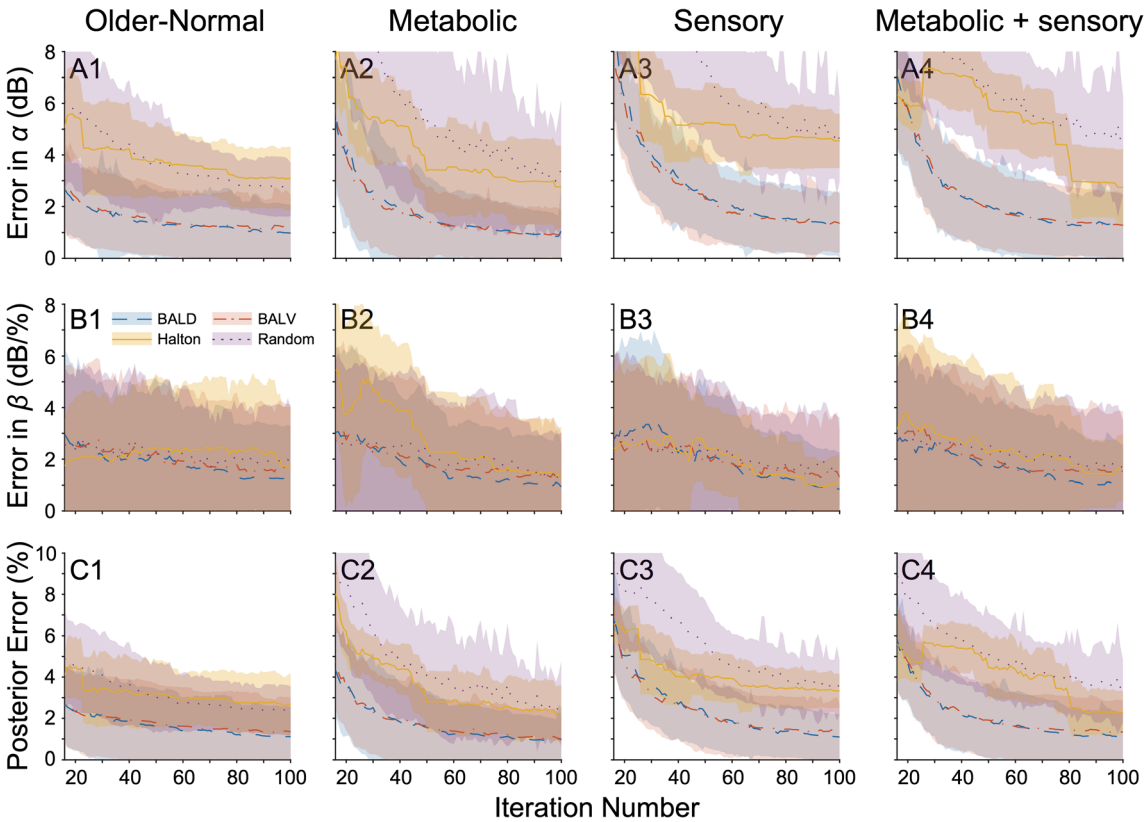

Fig. 6 Summary of GP estimator performance for each sampling mode separated by audiogram phenotype. Colored lines denote the mean and shaded areas denote 1 standard deviation above and below the mean. Unsigned differences in α and β between model predictions and ground truth for each phenotype are shown in (A1–A4) and (B1–B4), respectively; (C1–C4) shows the mean unsigned difference in probability percent between the GP posterior mean and ground truth

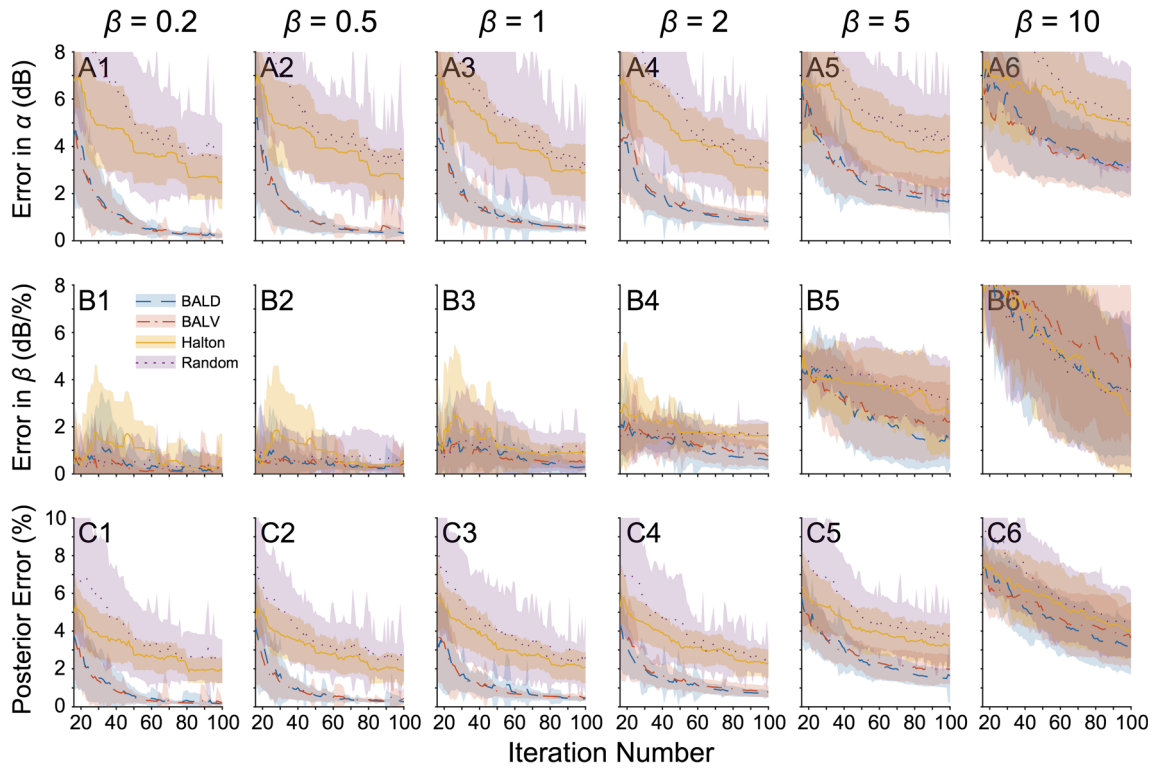

Fig. 7 Summary of GP estimator performance for each sampling mode separated by β (spread) value. Colored lines denote the mean and shaded areas denote 1 standard deviation above and below. Unsigned differences in α and β between model predictions and ground truth for each β value are shown in (A1–A6) and (B1–B6), respectively; (C1–C6) shows the mean unsigned difference in probability percent between the GP posterior mean and ground truth

attributed to the fact that BALV selects samples closest to a probability of detection of 0.5 and may not span a sufficient intensity range around threshold to adequately estimate larger β values. BALD incorporates an inherent tradeoff between exploration and exploitation (Houlsby et al., 2011), which helps mitigate this issue.

For the current work, hyperparameters for the mean and covariance functions were learned on each iteration with uninformative assumptions about the range of possible values they might take (i.e., a uniform hyperprior across all hyperparameter values). In some cases, however, large outliers or systematic misestimates for parameters (particularly for large spreads) occurred that may be attributed to hyperparameter values. For instance, observed uncertainty around the threshold for PFs may be captured using a smaller length constant rather than an increased β. In future work, specific hyperprior distributions on the hyperparameter values may be chosen given sufficient prior knowledge about a particular hyperparameter to refine the space of possible hyperparameter values within realistic ranges. Additionally, because the hyperparameter space is not necessarily convex, numerical techniques more robust to non-convex surfaces than gradient descent, such as stochastic gradient descent (Bishop, 2006) or simulated annealing (Kirkpatrick et al., 1983), can instead be used to find the point of maximum log marginal likelihood. Finally, it is possible to omit hyperparameter learning from certain iterations, such as early iterations where the limited number of samples may not best reflect the underlying probability detection surface.

Notably, the heuristic of adding zero-mean Gaussian noise with variance equal to 20 % of the acquisition function range improved the accuracy of the sampling techniques BALD and BALV (data not shown). Although modest improvements were observed for BALV, performance of BALD improved substantially with the implementation of this heuristic. Unmodified, BALD tended to select samples in clusters, particularly along the edges in the frequency dimension. The additive Gaussian noise allowed for selection of points that were not the absolute maximum of the acquisition function, mitigating the previously observed clustering problem.

Currently, the assigned GP prior makes few assumptions about the shape of the audiometric function, except that it is continuously smooth in the frequency dimension and sigmoidally increasing in the intensity dimension. More audiometric data could be used to better inform the GP prior construction. If there is a widely accepted

parametric form for a PF in a particular dimension, a mean and covariance function can be chosen to reflect this parameterization. In the case of predictable shapes for particular disease states, such as noise-induced hearing loss, Bayesian active model selection can be used to rapidly distinguish between different disorders (Gardner et al., 2015a). If this framework is applied across multiple testing sessions, for example, data from the patient's previous session can be incorporated into the new instance rather than starting from a relatively uninformative prior. This tradeoff between flexibility and specificity is a natural feature of GP estimation that allows researchers or clinicians to incorporate prior information as they see fit.

The current work demonstrates that the GP framework can be effective with several off-the-shelf sampling methods, both non-active and active, with active generally outperforming non-active methods. Active sampling methods are not constrained by the GP framework, however. Other adaptive or active sampling techniques could be used in conjunction with a GP model, if desired. Abundant literature describes 1-step lookahead active sampling techniques commonly used for PC estimation (King-Smith et al., 1994; Kontsevich & Tyler, 1999; Leek, 2001; Lesmes et al., 2006; Shen & Richards, 2013), and even clinical adaptive techniques such as the modified Hughson-Westlake procedure can be substituted. Regardless of which sampling procedure is deployed, the observations selected by that method can be passed into the GP framework to form a posterior estimate of the PF.

The choice of a linear covariance function through a sigmoidal observation model results in a PF that naturally spans the full range $[0, 1]$ (Kuss et al., 2005; Song et al., 2017), which implicitly assumes lapse and guess rates of 0. Although for detection tasks these rates are typically close to 0, failing to incorporate these rates into the model can introduce model bias when lapses truly exist (Wichmann & Hill, 2001a). The GP framework can account for arbitrary lapse rates by modification of the observation model, though requiring additional hyperparameters and complexity. Such a generalization would also allow for ready extension of the GP framework to other psychometric tasks, such as two-alternative forced choice, where the guess rate is explicitly non-zero.

Because of the linear term in the covariance function $s_1^2(\iota \cdot \iota')$, the current model assumes that the value of β does not vary as a function of frequency. While scant data exist to indicate the utility of modeling variable audiometric spread as a function of frequency, it is certainly possible to do so with a straightforward extension of the current framework. To model a PF whose β values vary with frequency, the additive linear term in the covariance function can be modified by a multiplicative frequency-dependent term. Note that because a SE covariance function places constraints on covariances rather than function shapes, no similar limitation exists for α: the overall function smoothness need only be constrained to be uniform across the entire domain. The SE covariance function is a straightforward and flexible covariance function for this application, but given sufficient prior knowledge for the shape of a PF, other covariance functions can be selected. Covariance functions that are periodic or whose length scales change as a function of some dimension may be suitable for particular function shapes.

Conclusion

This paper describes a method for estimating psychometric fields with actively selected queries, extending previous work using this technique with non-active samples. The method makes use of Gaussian process classification, a nonparametric Bayesian inference framework that allows for estimating complex PFs with limited categorical observations. Results show that active sampling techniques generally outperform Halton and random sampling for a 2D PF, reaching clinical accuracy in 20–30 samples. The flexibility of this technique allows for straightforward extension to other psychophysical domains, integration with other active sampling strategies, and incorporation of more informative prior beliefs into the model. This technique therefore represents an efficient, flexible method for estimating arbitrary PFs.

Acknowledgements Funding for this project was provided by the Center for Integration of Medicine and Innovative Technology (CIMIT) and National Center for Advancing Translational Sciences (NCATS) grant UL1 TR002345.

References

Allen, P., and Wightman, F. (1994). "Psychometric functions for children's detection of tones in noise," J Speech Lang Hear Res 37, 205-215.

American National Standards Institute (2004). "Methods for manual pure-tone threshold audiometry," ANSI 3, 21.

Bargones, J. Y., Werner, L. A., and Marean, G. C. (1995). "Infant psychometric functions for detection: Mechanisms of immature sensitivity," J Acoust Soc Am 98, 99-111.

Bengtsson, B., Olsson, J., Heijl, A., and Rootzén, H. (1997). "A new generation of algorithms for computerized threshold perimetry, SITA," Acta ophthalmol 75, 368-375.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Springer, New York).

Brant, L. J., and Fozard, J. L. (1990). "Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging," J Acoust Soc Am 88, 813-820.

Brochu, E., Cora, V. M., and De Freitas, N. (2010). "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," arXiv preprint arXiv:1012.2599.

Buss, E., Hall III, J. W., and Grose, J. H. (2006). "Development and the role of internal noise in detection and discrimination thresholds with narrow band stimuli," J Acoust Soc Am 120, 2777-2788.

Buss, E., Hall, J. W., 3rd, and Grose, J. H. (2009). "Psychometric functions for pure tone intensity discrimination: Slope differences in school-aged children and adults," J Acoust Soc Am 125, 1050-1058.

Carhart, R., and Jerger, J. (1959). "Preferred method for clinical determination of pure-tone thresholds," J Speech Hear Disord 24, 330-345.

Chaloner, K., and Verdinelli, I. (1995). "Bayesian experimental design: A review," Stat Sci 10, 273-304.

Cover, T. M., and Thomas, J. A. (2012). Elements of information theory (John Wiley & Sons, New York).

DiMattina, C. (2015). "Fast adaptive estimation of multidimensional psychometric functions," J Vis 15, 5.

Doire, C. S., Brookes, M., and Naylor, P. A. (2017). "Robust and efficient Bayesian adaptive psychometric function estimation," J Acoust Soc Am 141, 2501-2512.

Dubno, J. R., Eckert, M. A., Lee, F.-S., Matthews, L. J., and Schmiedt, R. A. (2013). "Classifying human audiometric phenotypes of age-related hearing loss from animal models," J Assoc Res Otolaryngol 14, 687-701.

Duvenaud, D. (2014). Automatic Model Construction with Gaussian Processes (University of Cambridge, Cambridge, England), pp. 1-132.

Fausti, S. A., Frey, R., Henry, J., Knutsen, J., and Olson, D. (1990). "Reliability and validity of high-frequency (8–20 kHz) thresholds obtained on a computer-based audiometer as compared to a documented laboratory system," J Am Acad Audiol 1, 162-170.

Fechner, G. T. (1860). Elements of Psychophysics (Holt, Rhinehart & Winston, New York), pp. 1-286.

Gardner, J., Malkomes, G., Garnett, R., Weinberger, K. Q., Barbour, D., and Cunningham, J. P. (2015a). "Bayesian active model selection with an application to automated audiometry," Adv Neural Inf Process Syst, (Morgan Kaufmann Publishers Inc., San Francisco, CA), 2377-2385.

Gardner, J. M., Song, X. D., Cunningham, J. P., Barbour, D. L., and Weinberger, K. Q. (2015b). "Psychophysical testing with Bayesian active learning," Uncertain Artif Intell, (Morgan Kaufmann Publishers Inc., San Francisco, CA), 286-295.

Geisler, C. D., Deng, L., and Greenberg, S. R. (1985). "Thresholds for primary auditory fibers using statistically defined criteria," Journal of the Acoustical Society of America 77, 1102-1109.

Gubner, J. A. (2006). Probability and random processes for electrical and computer engineers (Cambridge University Press, Cambridge, UK).

Guestrin, C., Krause, A., and Singh, A. P. (2005). "Near-optimal sensor placements in gaussian processes," Proceedings of the 22nd international conference on Machine learning, (Association for Computing Machinery), 265-272.

Hall, J. L. (1981). "Hybrid adaptive procedure for estimation of psychometric functions," The Journal of the Acoustical Society of America 69, 1763-1769.

Halton, J. H. (1964). "Algorithm 247: Radical-inverse quasi-random point sequence," Commun ACM 7, 701-702.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning* (Springer).

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). "Bayesian active learning for classification and preference learning," arXiv preprint arXiv:1112.5745.

Hughson, W., and Westlake, H. (1944). "Manual for program outline for rehabilitation of aural casualties both military and civilian," Trans Am Acad Ophthalmol Otolaryngol 48, 1-15.

Jefferys, W. H., and Berger, J. O. (1992). "Ockham's razor and Bayesian analysis," Am Sci 80, 64-72.

Katz, J., Medwetsky, L., Burkhard, R., and Hood, L. (2009). Handbook *of Clinical Audiology* (Lippincott Williams & Wilkins).

Kiang, N. Y. S., Watanabe, T., Thomas, E. C., and Clark, L. F. (1965). Discharge Patterns of Single Fibers in the Cat's Auditory Nerve (The MIT Press, Cambridge, MA), pp. 1-154.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., and Supowit, A. (1994). "Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation," Vision Res 34, 885-912.

Kingdom, F. A. A., and Prins, N. (2016). Psychophysics: A Practical Introduction (Academic Press, London), pp. 1-307.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). "Optimization by simulated annealing," Science 220, 671-680.

Klein, S. A. (2001). "Measuring, estimating, and understanding the psychometric function: a commentary," Percept Psychophys 63, 1421-1455.

Kontsevich, L. L., and Tyler, C. W. (1999). "Bayesian adaptive estimation of psychometric slope and threshold," Vision Res 39, 2729-2737.

Kujala, J. V. (2011). "Bayesian adaptive estimation: a theoretical review," in Descriptive and Normative Approaches to Human Behavior, edited by E. Dzhafarov, and L. Perry (World Scientific Publishing Company, Singapore), pp. 123-159.

Kuss, M., Jäkel, F., and Wichmann, F. A. (2005). "Bayesian inference for psychometric functions," J Vis 5, 8.

Leek, M. R. (2001). "Adaptive procedures in psychophysical research," Percept Psychophys 63, 1279-1292.

Lesmes, L. L., Jeon, S. T., Lu, Z. L., and Dosher, B. A. (2006). "Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method," Vision Res 46, 3160-3176.

Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," J Acoust Soc Am 49, 467-477.

Lewis, D. D., and Catlett, J. (1994). "Heterogeneous uncertainty sampling for supervised learning," Proceedings of the Eleventh International Conference on Machine Learning, 148-156.

Lewis, D. D., and Gale, W. A. (1994). "A sequential algorithm for training text classifiers," Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, (Springer-Verlag New York, Inc.), 3-12.

Liberman, M. C. (1978). "Auditory-nerve response from cats raised in a low-noise chamber," J Acoust Soc Am 63, 442-455.

Mahomed, F., Eikelboom, R. H., and Soer, M. (2013). "Validity of automated threshold audiometry: A systematic review and meta-analysis," Ear Hearing 34, 745-752.

Margolis, R. H., and Morgan, D. E. (2008). "Automated pure-tone audiometry: an analysis of capacity, need, and benefit," Am J Audiol 17, 109-113.

Minka, T. P. (2001). "Expectation propagation for approximate Bayesian inference," Uncertain Artif Intell 17, (Morgan Kaufmann Publishers Inc., San Francisco, CA), 362-369.

Osborne, M. A., Garnett, R., and Roberts, S. J. (2009). "Gaussian processes for global optimization," 3rd international conference on learning and intelligent optimization (LION3), 1-15.

Özdamar, Ö., Eilers, R. E., Miskiel, E., and Widen, J. (1990). "Classification of audiograms by sequential testing using a dynamic Bayesian procedure," The Journal of the Acoustical Society of America 88, 2171-2179.

Park, M., Horwitz, G., and Pillow, J. W. (2011). "Active learning of neural response functions with Gaussian processes," Adv Neural Inf Process Syst, (Curran Associates, Manila, Philippines), 2043-2051.

Park, M. J. (2013). Bayesian learning methods for neural coding (The University of Texas at Austin, Austin, Texas), pp. 157.

Pentland, A. (1980). "Maximum likelihood estimation: The best PEST," Perception & Psychophysics 28, 377-379.

Rasmussen, C. E., and Williams, C. K. I. (2006). Gaussian Processes for Machine Learning (The MIT Press, Cambridge, MA), pp. 1-248.

Settles, B. (2009). "Active learning literature survey," in *Computer Sciences Technical Report 1648* (University of Wisconsin, Madison).

Shen, Y., and Richards, V. M. (2012). "A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention," J Acoust Soc Am 132, 957-967.

Shen, Y., and Richards, V. M. (2013). "Bayesian adaptive estimation of the auditory filter," J Acoust Soc Am 134, 1134-1145.

Song, X. D., Garnett, R., and Barbour, D. L. (2017). "Psychometric function estimation by probabilistic classification," J Acoust Soc Am 141, 2513-2525.

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., and Barbour, D. L. (2015). "Fast, continuous audiogram estimation using machine learning," Ear Hearing 36 e326-e335.

Swanepoel, D. W., Mngemane, S., Molemong, S., Mkwanazi, H., and Tutshini, S. (2010). "Hearing assessment—reliability, accuracy, and efficiency of automated audiometry," Telemed J E Health 16, 557-563.

Taylor, M. M., and Creelman, C. D. (1967). "PEST: Efficient estimates on probability functions," The Journal of the Acoustical Society of America 41, 782-787.

Treutwein, B. (1995). "Adaptive psychophysical procedures," Vision Res 35, 2503-2522.

Von Békésy, G. (1960). Experiments in Hearing (McGraw-Hill, New York), pp. 1-745.

Watson, A. B. (2017). "QUEST+: A general multidimensional Bayesian adaptive psychometric method," J Vis 17, 10.

Watson, A. B., and Pelli, D. G. (1983). "QUEST: A Bayesian adaptive psychometric method," Perception & Psychophysics 33, 113-120.

Wichmann, F. A., and Hill, N. J. (2001a). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept Psychophys 63, 1293-1313.

Wichmann, F. A., and Hill, N. J. (2001b). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," Percept Psychophys 63, 1314-1329.

Williams, C. K., and Barber, D. (1998). "Bayesian classification with Gaussian processes," IEEE Trans Pattern Anal Mach Intell 20, 1342-1351.

Williams, C. K., and Rasmussen, C. E. (1996). "Gaussian processes for regression," in Adv Neural Inf Process Syst 8 (NIPS '95), edited by D. Touretzky, M. Mozer, and M. Hasselmo (MIT Press, Cambridge, MA).

Williams, C. K. I. (1998). "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," in Learning in Graphical Models, edited by M. I. Jordan (Springer Netherlands, Dordrecht), pp. 599-621.

Xiang, N., and Fackler, C. (2015). "Objective Bayesian analysis in acoustics," Acoust Today 11, 54-61.