Conjoint psychometric field estimation for bilateral audiometry

Dennis L. Barbour¹ · James C. DiLorenzo^{1,2} · Kiron A. Sukesan^{1,2} · Xinyu D. Song¹ · Jeff Y. Chen^{1,2} · Eleanor A. Degen¹ · Katherine L. Heisey^{1,3} · Roman Garnett²

Published online: 12 June 2018
© Psychonomic Society, Inc. 2018

Abstract

Behavioral testing in perceptual or cognitive domains requires querying a subject multiple times in order to quantify his or her ability in the corresponding domain. These queries must be conducted sequentially, and any additional testing domains are also typically tested sequentially, such as with distinct tests comprising a test battery. As a result, existing behavioral tests are often lengthy and do not offer comprehensive evaluation. The use of active machine-learning kernel methods for behavioral assessment provides extremely flexible yet efficient estimation tools to more thoroughly investigate perceptual or cognitive processes without incurring the penalty of excessive testing time. Audiometry represents perhaps the simplest test case to demonstrate the utility of these techniques. In pure-tone audiometry, hearing is assessed in the two-dimensional input space of frequency and intensity, and the test is repeated for both ears. Although an individual's ears are not linked physiologically, they share many features in common that lead to correlations suitable for exploitation in testing. The bilateral audiogram estimates hearing thresholds in both ears simultaneously by conjoining their separate input domains into a single search space, which can be evaluated efficiently with modern machine-learning methods. The result is the introduction of the first conjoint psychometric function estimation procedure, which consistently delivers accurate results in significantly less time than sequential disjoint estimators.

Keywords Psychophysics · Perceptual testing · Audiometry · Hearing · Psychometric function

A psychometric curve represents the probabilistic behavior of a subject in response to a unidimensional perceptual or cognitive task. These curves take the form of monotonically increasing probabilities as a function of increasing task ease, indexed by a single independent variable (Fechner, 1860/1966; Kingdom & Prins, 2016). When tasks are represented by multiple independent variables, a psychometric field results. Estimation procedures for unidimensional psychometric curves are many, varied, and widespread and have a long history. Estimation procedures for multidimensional psychometric fields, on the other hands,

are much less advanced. In some ways, psychometric field estimation could be performed like any multiple logistic regression procedure (Hosmer & Lemeshow, 2013). The main issue with this simple formulation, however, is that because data must be accumulated sequentially in these tasks, inefficient inference requiring hours of subject performance is too impractical for any kind of mainstream test. As a result, the relatively few psychometric field estimation procedures tend to have been customized to each application (Bengtsson, Olsson, Heijl, & Rootzén, 1997; Heijl & Krakau, 1975; Lesmes, Jeon, Lu, & Doshier, 2006; Shen & Richards, 2013), with relatively few studies proposing general principles (Cohen, 2003; DiMattina, 2015).

An alternative approach to psychometric field estimation recasts the key inference step from parametric logistic regression to nonparametric or semiparametric probabilistic classification (Song, Garnett, & Barbour, 2017; Song, Sukesan, & Barbour, 2018). This framework naturally scales to multiple input dimensions, provides great flexibility for estimating a wide variety of functions, and sets the stage for other novel theoretical advances. One such advance involves actively learning which stimuli would be most valuable to deliver in order to rapidly converge onto an accurate estimate (Song et al., 2018; Song et al., 2015b). This procedure, referred to as

✉ Dennis L. Barbour
dbarbour@wustl.edu

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

² Department of Computer Science and Engineering, Washington University, 1 Brookings Drive, Box 1045, St. Louis, MO 63130, USA

³ Program in Neuroscience, Washington University School of Medicine, 660 South Euclid Drive, Box 8111, St. Louis, MO 63130, USA

active testing, yields estimation efficiencies at least as high as those of adaptive parametric testing. Another advance allows for multiple stimuli to be delivered simultaneously while retaining the same binary subject responses elicited classically (Gardner, Song, Weinberger, Barbour, & Cunningham, 2015). This procedure, referred to as *multiplexed active testing*, substantially improves test efficiency beyond that of active testing alone by reducing the total number of stimulus presentations.

Another possible advance includes considering multiple nominally decoupled input domains for incorporation into a single task and a single estimator. This extension is possible when the method used to implement the probabilistic classifier is capable of learning nonlinear interactions between the input dimensions of interest. As long as the input domains share some interrelationship, an active method that can learn and exploit this information should produce accurate overall estimates in less time. This procedure is termed *conjoint active testing*.

An example of a simple multidimensional perceptual test that could benefit from conjoint testing is the threshold audiogram, which evaluates the hearing function of each ear typically via a staircase method sequentially applied at discrete frequencies (Carhart & Jerger, 1959; Hughson & Westlake, 1944). This clinical standard is a compromise between the necessary acquisition of important diagnostic information and the length of time required to conduct the test, which is on the order of 15 min for both ears. Estimating full psychometric field audiograms at the same frequency resolution using serial logistic regression would require up to 20 h and is rarely performed for obvious reasons (Allen & Wightman, 1994). An appropriately designed active learning procedure can estimate a full audiogram in less time than is required to estimate a threshold audiogram using conventional methods (Song et al., 2018; Song et al., 2015b).

Like all other perceptual tests to date, estimation of a person's hearing proceeds sequentially, one input domain (i.e., ear) at a time, resulting in two independently measured unilateral audiograms. The information about one ear is not used to infer information about the other ear during testing, though such information could be used to speed estimation of the other ear in two ways. In the first strategy, once the estimation of one ear's hearing is complete, that ear's audiogram could be used as a Bayesian prior to initiate testing of the contralateral ear. If the two ears share common features above and beyond what is shared among human ears generally, the contralateral ear could in that case be accurately estimated in less time.

A more compelling strategy, however, would be to use information from each ear to infer the audiograms of both ears in real time as the test is being conducted. This mutually conjoint testing strategy is referred to as the *bilateral audiogram*, which should be equivalent to the two separately estimated unilateral audiograms, but it should be determined in less time because of shared information between the ears. The purpose of this study was to develop the mathematical theory of conjoint psychometric field estimation, apply it to the bilateral audiogram, and evaluate the

efficiency and accuracy of this novel method for determining the hearing thresholds of a subject's two ears.

Theory

Gaussian processes (GPs) can be used to model probabilistic inference about some function of interest $f(x)$. That is, instead of simply producing a pointwise estimate $\hat{f}(x)$, a GP returns a probability distribution $p(f)$. A user may also encode domain-specific knowledge of f through a prior distribution. The GP can then be conditioned on observed data $D = \{x_i, y_i\}_{i=1}^n$ to form a posterior distribution $p(f|D)$. Formally, a GP is a collection of random variables such that the joint distribution of any finite subset of these variables is a multivariate Gaussian distribution (Rasmussen & Williams, 2006). It is more conceptually straightforward to think of GPs as distributions over functions. Just as a variable drawn from a Gaussian distribution is specified by the distribution's mean and variance—that is, $p(x) \sim N(\mu, \sigma^2)$ —a function drawn from a GP distribution is specified by that GP's mean and kernel functions—that is, $p(f) \sim GP(\mu(x), K(x, x'))$. The mean function encodes the central tendency of functions drawn from the GP, whereas the kernel function encodes information about the shapes these functions may take around the mean. Kernel functions can vary widely in construction and have a large impact on the posterior distribution of the GP. Typically, kernel functions are designed to express the belief that “similar inputs should produce similar outputs” (Duvenaud, 2014). The GP model can be used in both classification and regression settings and enables the conditioning of prior beliefs after observing data in order to produce a new posterior belief about the function values via Bayes's theorem:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}$$

The GP model for audiogram estimation yields probabilistic estimates for the likelihood of tone detection, which is inherently a classification task. To properly construct a framework for GP classification, however, it is convenient to first examine GP regression.

In a typical multidimensional regression problem, the observed inputs $\mathbf{X}$ and observed outputs $\mathbf{y}$ take on real values and are related through some function f , to which we have access only via noisy observations. For convenience, this example assumes that the noise is drawn independently and identically from a Gaussian distribution with mean 0 and standard deviation s :

$$\begin{aligned} \mathbf{x}_i &\in \mathbb{R}^d \\ y_i &\in \mathbb{R} \\ y_i &= f(\mathbf{x}_i) + \varepsilon \\ \varepsilon &\sim N(0, s^2) \\ D &= \{\mathbf{x}_i, y_i\}_{i=1}^n = \{\mathbf{X}, \mathbf{y}\} \end{aligned}$$

The GP by definition implies a joint distribution on the function values of any set of input points:

$$p(f|\mathbf{X}) = N(\mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X}))$$

More importantly, GPs allow us to condition the predictive distribution over unseen points $\mathbf{X}_*$ on (possibly noisy) observations of f . Let $\mathbf{y}=f(\mathbf{X})+\varepsilon$ be noisy observations of f at training inputs $\mathbf{X}$, and let $\mathbf{f}_*=f(\mathbf{X}_*)$ be the test outputs of interest. Then, the joint distribution implied by the GP is

$$p\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix}\right) = N\left(\begin{bmatrix} \mu(\mathbf{X}) \\ \mu(\mathbf{X}_*) \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + s^2\mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right)$$

An application of Bayes's theorem yields

$$p(\mathbf{f}_*|\mathbf{X}_*, D) = N(\mu_{f|D}(\mathbf{X}_*), K_{f|D}(\mathbf{X}_*, \mathbf{X}_*))$$

where

$$\begin{aligned} \mu_{f|D}(\mathbf{X}_*) &= \mu(\mathbf{X}) + K(\mathbf{X}_*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + s^2\mathbf{I})^{-1}(\mathbf{y} - \mu(\mathbf{X})) \\ K_{f|D}(\mathbf{X}_*, \mathbf{X}_*) &= K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + s^2\mathbf{I})^{-1}K(\mathbf{X}, \mathbf{X}_*) \end{aligned}$$

(Rasmussen & Williams, 2006). The posterior mean and covariance functions reflect both the prior assumptions and the information contained in the observations.

In classification problems, the target function shifts from producing real-valued outputs to a discrete space where y_i can take on only a fixed number of classes $C_1, C_2, \dots, C_m$. Of particular interest here is the special case of binary classification, in which outputs can take on one of two classes: $y_i \in \{0, 1\}$. Linear classification methods assume that the class-conditional probability of belonging to the “positive” class is a nonlinear transformation of an underlying function known as the *latent function*, which applies the following transformation to the likelihood:

$$p(y(\mathbf{x}_i) = 1|f(\mathbf{x}_i)) = \Phi(f(\mathbf{x}_i)) = p(y_i|f_i)$$

The observation function Φ can be any sigmoidal function. Common choices of sigmoidal functions include the logistic function $\Phi(w) = \frac{\exp(w)}{1+\exp(w)}$ and the cumulative Gaussian $\Phi(w) = \int_{-\infty}^w \frac{\exp(-z^2)}{\sqrt{2\pi}} dz$. One further complication to the GP classification problem must be taken into account. Under the assumption that the observations are conditionally independent given the latent function values, Bayes's theorem gives the posterior distribution as

$$p(\mathbf{f}|D) = \frac{1}{Z} p(\mathbf{f}|\mathbf{X}) p(\mathbf{y}|\mathbf{f}) = \frac{1}{Z} N(\mu(\mathbf{X}), \sigma(\mathbf{X})) \prod_i p(y_i|f_i)$$

where Z is a normalization factor that is approximated in the

schemes discussed below. In the regression setting, the posterior distribution is easy to work with directly, because it is the product of a Gaussian prior and a Gaussian likelihood. Likelihood is sigmoidal in the classification setting, however, and the product of a Gaussian distribution with a sigmoidal function does not produce a tractable posterior distribution. The model must instead approximate the posterior with a Gaussian distribution in order to exploit the computational advantages of the GP estimation framework. Common approximation schemes include Laplace approximation and expectation propagation (Rasmussen & Williams, 2006). *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 (Williams & Barber, 1998). *Expectation propagation* attempts to approximate the posterior distribution by matching the first and second moments—the mean and variance—of the posterior distribution (Minka, 2001).

As mentioned previously, kernel functions encode information about the shape and smoothness of the functions drawn from a GP. Although the GP itself is a nonparametric model, many kernel functions themselves have parameters Θ , referred to as *hyperparameters*. The adjustment of hyperparameters exerts considerable influence over the predictive distribution of the GP. For instance, the popular squared exponential kernel is parameterized by its length scale ℓ and output variance σ (Rasmussen & Williams, 2006):

$$K(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\ell}\right)$$

Again, the model belief about the hyperparameters can be computed via Bayes's theorem:

$$p(\Theta|D, H) = p(y|\mathbf{x}, \Theta)p(\Theta|H)$$

where $p(\Theta|H)$ is the hyperparameter prior, which can be used to encode domain knowledge about the settings of hyperparameters or may be left uninformative (Rasmussen & Williams, 2006). Determining the posterior distribution is often computationally intractable, and thus settings of the hyperparameters may be chosen through optimization algorithms such as gradient descent.

One notable advantage of the GP model is that its probabilistic predictions enable a set of techniques collectively known as *active learning*. Active learning, sometimes called “optimal experimental design,” allows a machine-learning model to select the data it samples in order to perform better with less training (Settles, 2009). To contrast with adaptive techniques, queries via active learning are chosen in such a way as to minimize some loss function. For example, an active learning procedure may select a query designed to minimize the expected error of the model against the latent function. In general, the

application of active learning proceeds as follows: First, use the existing model to classify unobserved data; next, find the best next point to query on the basis of some objective function, and query the data via an oracle (e.g., a human expert); finally, retrain the classifier, and repeat these steps until satisfied.

The most common form of active learning is uncertainty sampling (Lewis & Gale, 1994; Settles, 2009). Models employing uncertainty sampling will query regions in the input domain about which the model is most uncertain. In the case of probabilistic classification, uncertainty sampling corresponds to querying the instances for which the probability of being either adjacent class is closest to 0.5. This method can rapidly identify a class boundary for a target function of interest, but because uncertainty sampling attempts to query exactly where $p(y = 1|\mathbf{x}) = 0.5$ (in the binary case), the model underexplores the input space. In the context of psychometric fields, the transition from one class to another (i.e., the psychometric spread) is not as readily estimated in this case (Song et al., 2018).

Bayesian active learning by disagreement (BALD) attempts to circumvent this problem via an information-theoretic approach (Houlsby, Huszár, Ghahramani, & Lengyel, 2011). Information-theoretic optimization has been successful at implementing efficient parametric perceptual modeling, first for unidimensional (Kontsevich & Tyler, 1999) and then for multidimensional (DiMattina, 2015; Kujala & Lukka, 2006) psychometric functions. The implementation of the BALD method used here (Garnett, Osborne, & Hennig, 2013) assumes the existence of some hyperparameters Θ that control the relationship between the inputs and outputs $p(y|\mathbf{x}, \Theta)$. When performing GP regression with a squared exponential kernel, for example, Θ would be the length scale and output variance hyperparameters. Under the Bayesian framework, it is possible to infer a posterior distribution over the hyperparameters $p(\Theta|D)$. Each possible setting of Θ represents a distinct hypothesis about the relationship between the inputs and outputs. The goal of the BALD method is to reduce the number of viable hypotheses as quickly as possible by minimizing the entropy of the posterior distribution of Θ . To that end, BALD queries the point $\mathbf{x}$ that maximizes the decrease in expected entropy:

$$\operatorname{argmax}_{\mathbf{x}^*} H[\Theta|D] - E_{y^* \sim p(y^*|\mathbf{x}^*, D)} [H[\Theta|y^*, \mathbf{x}^*, D]]$$

where $H[\Theta|D]$ is Shannon's entropy of Θ given D . This expression can be difficult to compute directly because the latent parameters often exist in high-dimensional space, but they can be rewritten in terms of entropies in the one-dimensional output space (Kujala & Lukka, 2006):

$$\operatorname{argmax}_{\mathbf{x}^*} H[y^*|\mathbf{x}^*, D] - E_{\Theta \sim p(\Theta|D)} [H[y^*|\mathbf{x}^*, \Theta]]$$

This expression can be computed in linear time, making it easy to work with in practice. BALD selects the $\mathbf{x}$ for which the entire model is most uncertain about y (i.e., high $H[y|\mathbf{x}]$) but for which the individual predictions given a setting of the hyperparameters are very confident. This can be interpreted as “seeking the $\mathbf{x}$ for which the [hyper]parameters under the posterior disagree about the outcome the most” (Houlsby et al., 2011).

Materials and methods

Simulated subjects

Simulated subjects were assigned distinct ground-truth audiograms for each ear. These audiograms defined the probability of stimulus detection over a two-dimensional input domain consisting of sound frequency and intensity. The audiogram shapes were defined by one of four canonical human audiogram phenotypes: older-normal, sensory, metabolic, and metabolic + sensory (in order of severity of hearing loss). These phenotypic categories are evident in the population data of human audiograms and are informed by etiologies determined via physiological study in animal models (Dubno, Eckert, Lee, Matthews, & Schmiedt, 2013). The average audiogram shape for these categories therefore spans diagnostic categories from normal (older-normal) through the most common pathologic categories that theoretically could affect each ear separately (metabolic, sensory, metabolic + sensory). Three different pairings of ground-truth audiograms (normal–normal, normal–pathologic, and pathologic–pathologic) therefore reflect conditions with varying putative estimation benefit from considering both ears conjointly. These canonical audiogram phenotypes have been used previously to evaluate the accuracy (Song et al., 2017) and efficiency (Song et al., 2018) of disjoint machine-learning audiometry.

In the context of this work, the *threshold* is defined as an input point $x = (\omega, I)$ such that $p(y = 1|\mathbf{x}) = 0.5$. These standard phenotypes provide threshold estimates at octave frequencies, as would typically be observed by the Hughson–Westlake procedure (American National Standards Institute, 1978; Carhart & Jerger, 1959; Hughson & Westlake, 1944; International Organization for Standardization, 2010). Spline interpolation and linear extrapolation were used to generate a continuous ground-truth threshold across frequencies. At each semitone of frequency, a cumulative Gaussian was used to generate a sigmoidal psychometric curve representing the probabilities of tone detection above and below threshold (Song et al., 2017). The cumulative Gaussian was parameterized by the intensity and threshold (I, t) :

$$p(y = 1|I, t) = \int_{-\infty}^I \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t-t')^2}{2}\right) dt'$$

Threshold is a function of frequency. The subject response for any particular tone was simulated as a Bernoulli random variable with success probability given by the cumulative Gaussian.

Bilateral audiogram

Traditional pure-tone audiometry involves delivering tones in a two-dimensional continuous input domain indexed by frequency and intensity. The input domain for the bilateral audiogram is augmented to include a third discrete “ear” dimension, yielding $x_i = (\omega_i, I_i, e_i)$. In querying a simulated subject’s audiogram, the conjoint estimator determines in which ear to deliver the tone as well as the frequency and intensity of tone delivered. Binary responses of “heard” or “unheard” are recorded as described previously for each simulated tone delivery.

The conjoint estimator uses a constant mean function, $\mu(\mathbf{x}) = c \forall \mathbf{x} \in X$. This mean function is not representative of any particular audiogram phenotype, and deviation from the mean is captured in the posterior distribution of the GP classification model. The GP kernel function was derived from prior knowledge about the behavior of audiograms. Knowing that a subject’s psychometric curve for any frequency is sigmoidal allows us to place a linear kernel in the intensity dimension:

$$K_I(\mathbf{x}, \mathbf{x}') = I \cdot I'$$

A linear kernel incorporates the prior knowledge that higher intensities are more likely to be detected than lower intensities. The precise shape of this detectability can be modeled with the likelihood function, described below.

Additionally, the model leverages the continuity of audiogram thresholds by placing an isotropic squared exponential kernel with unit magnitude over the frequency domain:

$$K_\omega(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\omega - \omega')^2}{2\ell}\right)$$

where ℓ is the length scale.

For quantification of the covariation between the ears, the conjoint estimator uses a discrete covariance function that directly parameterizes relationships between each pair of points in the discrete space.

$$K_e(\mathbf{x}, \mathbf{x}') = \begin{cases} s_{11} & \text{if } \mathbf{x}, \mathbf{x}' \in e_1 \\ s_{12} & \text{if } e \neq e' \\ s_{22} & \text{if } \mathbf{x}, \mathbf{x}' \in e_2 \end{cases}$$

This model can explicitly define the covariance between ears without having to relate them via some functional form. Computationally, this is done by modeling the discrete covariance as the Cholesky decomposition of a 2×2 matrix, $K = \Lambda \Lambda^T$. The model combines the individual covariance

functions into a single kernel as follows:

$$K(\mathbf{x}, \mathbf{x}') = K_e(\mathbf{x}, \mathbf{x}') (K_\omega(\mathbf{x}, \mathbf{x}') + K_I(\mathbf{x}, \mathbf{x}'))$$

Finally, the model uses the cumulative Gaussian likelihood function for binary classification, which is both standard for GP classification and effectively captures the sigmoidal behavior of psychometric functions as described above.

The exact form of the posterior requires computing the product of the likelihood and prior distributions. In the case of GP classification, the product of a Gaussian distribution with a sigmoidal function does not produce a tractable posterior distribution. The model must instead approximate the posterior. The estimator evaluated here uses expectation propagation, which approximates each of the sigmoidal likelihoods with moment-matching Gaussian distributions to derive a Gaussian posterior distribution (Gelman et al., 2014).

Simulations

The GP estimator is fully characterized by its mean and covariance functions. The GP used here has a constant mean function with one hyperparameter, and the kernel function has four hyperparameters: one for the length of the squared exponential kernel, and three for the discrete kernel. Because the posterior distribution of the hyperparameters $p(\Theta|D)$ may be multimodal, standard gradient descent approaches run the risk of becoming trapped in a local extremum. To circumvent this issue, the estimator performs gradient descent on two sets of hyperparameters after each observation. The first set comes from the most recent results of the model (or the hyperparameter prior in absence of any data). The second set is drawn from a multivariate Gaussian distribution whose mean is the hyperparameter prior derived below. Gradient descent is performed on both settings of the hyperparameters, and the setting with higher likelihood $p(D|\Theta)$ is retained for the next iteration.

The first iterations of the estimator originally performed poorly from inefficient early sampling. Resolving this issue involved learning reasonable hyperparameter priors to serve as a starting point for sample selection. Each of the four common human phenotypes discussed earlier has at least one optimal setting to its hyperparameters to minimize estimation error. Because the kernel function is symmetric, ten unique pairs of audiogram profiles can be derived from the four phenotypes. Data were collected for each of the ten audiogram profile pairs far in excess of what would be collected in a clinical or experimental setting. First, 400 stimuli were delivered across both ears using Halton sampling (Halton, 1964; Song et al., 2018). Then, an additional 100 stimuli were queried via BALD to gain additional sampling density around the threshold (Houlsby et al., 2011; Song et al., 2018). Hyperparameters were learned using modified gradient descent. The same concept was repeated with varying numbers

of Halton and BALD queries, though the hyperparameters converged to within 2% of each after about 300 samples. The final setting of the hyperparameter priors was computed by taking an average of the hyperparameters learned for each of the ten audiogram pairs, weighted by the prevalence of those phenotypes in human populations (Dubno et al., 2013).

Three different sampling and estimation procedures were conducted and compared against each other:

- *Estimator 1: Unconstrained mutually conjoint GP audiogram estimation (unconstrained conjoint).* This method performs inference using the conjoint audiogram estimation extension of GPs described above, giving the model complete choice over which ear to query, as well as which frequency/intensity pair to deliver. This procedure occasionally results in multiple stimuli being delivered sequentially to the same ear, particularly in cases in which the model is more unsure of the audiogram in one ear than the other.
- *Estimator 2: Alternating mutually conjoint GP audiogram estimation (alternating conjoint).* This method performs inference using the conjoint audiogram estimation extension of GPs described above but is artificially constrained to alternate samples between the left and right ears. Odd samples refer to the first ear and even samples to the second ear.
- *Estimator 3: Disjoint GP audiogram estimation (disjoint).* This method performs inference using two separate models of the existing GP audiogram framework (Song et al., 2017; Song et al., 2018). Information in this approach is not shared between ears. Tone delivery is alternated between left and right ears so that odd samples refer to the first ear and even samples to the second ear.

Three particular left/right ground truth audiogram configurations were modeled in order to demonstrate the utility of the conjoint audiogram estimation framework:

- *Case 1: Older normal.* This case was defined as having the older-normal phenotype in both ears
- *Case 2: Asymmetric hearing loss.* This case was defined as having the older-normal phenotype in one ear, and the metabolic + sensory phenotype in the other ear. This situation reflects more severe asymmetric hearing loss than is typical in human populations.
- *Case 3: Symmetric hearing loss.* This case was defined as having metabolic + sensory hearing loss in one ear, and sensory hearing loss in the other. This case is more typical of hearing loss in human subjects. The sensory and metabolic + sensory phenotypes have slightly different thresholds. Distinct phenotypes were selected for this case in order to more accurately reflect presentations of hearing loss in human subjects, in which generally the left and right audiograms are not exactly identical.

One hundred tests of each estimator–case pair were run, for a total of 900 tests. For each test, 100 tones were delivered sequentially to the simulated subjects and responses from the underlying ground truth were recorded. To avoid unstable hyperparameter learning and uninformative early querying, the first 15 tones were delivered via a modified Halton sampling algorithm. The modified Halton sampling algorithm constrained initial tone deliveries to be below 60 dB HL, which is a safeguard for human testing. Subsequent tones were sampled via BALD, with constraints for the disjoint and alternating conjoint cases as discussed above. Hyperparameters were learned via a modified gradient descent algorithm on every iteration starting with Iteration 16. Hyperparameter learning was off for the first 15 iterations of each test to prevent model instability. For each tone delivery, the model posterior distribution across the entire input space was recorded. From here, the abscissa intercept of the latent function was determined to calculate the 50% threshold over semitone frequencies from 0.125 to 16 kHz. The accuracy for a single test was evaluated using the mean absolute error between the estimated threshold and the true threshold at each semitone frequency. The results were then averaged at each iteration across all 100 tests to obtain the mean threshold error per iteration for each of the three estimators in each of the three cases. In addition to comparing mean threshold error per iteration, the mean number of iterations required to achieve less than 5 dB threshold error in each ear and then in both ears was also determined. This value was chosen as a measure of “convergence” because it is the minimum step size in the Hughson–Westlake procedure and is close to the empirical test–retest reliability of that test (Mahomed, Eikelboom, & Soer, 2013).

To quantify the similarity between audiogram thresholds for pairs of ears, a measure of intraclass correlation is necessary (Fisher, 1925). The perils of using a measure of interclass correlation in this case, such as the Pearson correlation, are exemplified for audiometry by debates on the appropriate interpretation of shared variation between the ears (Coren, 1989; Coren & Hakstian, 1990; Divenyi & Haupt, 1992). In this study, an unbiased Lin’s concordance correlation coefficient (CCC) was used to quantify similarity between the audiogram thresholds in left and right ears (l and r , respectively):

$$CCC = \frac{2S_{lr}}{S_l^2 + S_r^2 + (\bar{l} - \bar{r})^2}$$

$$S_{lr} = \frac{1}{N-1} \sum_{n=1}^N (l - \bar{l})(r - \bar{r})$$

$$S_l^2 = \frac{1}{N-1} \sum_{n=1}^N (l - \bar{l})^2$$

$$S_r^2 = \frac{1}{N-1} \sum_{n=1}^N (r - \bar{r})^2$$

$$\bar{l} = \frac{1}{N} \sum_{n=1}^N l$$

$$\bar{r} = \frac{1}{N} \sum_{n=1}^N r$$

This coefficient measures the agreement between two variables while taking into account the difference in their means (Lin, 1989).

Results

GP classification was used to simultaneously estimate the right-ear and left-ear audiograms of simulated subjects. The GP framework produces continuous audiogram estimates across the entire input domain, and tones were actively sampled in order to reduce the number of stimuli required to achieve acceptable error thresholds. The goals of this experiment were (1) to achieve with a single bilateral audiogram estimation procedure accuracy in estimated threshold comparable to conventional unilateral methods and (2) to achieve these results across both ears more quickly than would otherwise be possible with consecutive unilateral audiogram estimations.

To determine how much baseline similarity might be expected within and between the hearing functions of a human's ears, a large database of audiograms was first analyzed. Over the course of a decade, the National Institute for Occupational Safety and Health (NIOSH) collected discrete Hughson–Westlake audiogram threshold curves in roughly 1.1 million individuals to provide the scientific basis for recommending occupational noise exposure limits (Masterson et al., 2013). Hearing threshold levels at 500, 1000, 2000, 3000, 4000, 6000, and 8000 Hz for both the left and right ears were determined per worker and recorded within the NIOSH database.

A CCC value was first computed between thresholds of adjacent frequencies for both ears of each individual in the database, resulting in approximately 1.1 million values. The histogram of these values can be seen in Fig. 1A, revealing considerable average concordance, with a median of 0.299. To determine how much information this concordance represents

relative to the overall threshold distribution in the population, the same number of randomly selected threshold comparisons were made from shuffled individuals, ears, and frequencies. The histogram of these approximately 1.1 million CCC values can be seen in Fig. 1B. The median of this distribution is $-.007$, indicating very little, if any, baseline information across random frequencies.

The high concordance for adjacent thresholds is presumably a reflection of similar physiology for nearby points on the cochlea. In other words, it is a quantification of the smoothness of the hearing threshold functions. Traditional audiometry makes very little use of such prior knowledge, but disjoint GP audiogram estimation takes advantage of this information across the population, initially by incorporating appropriate prior beliefs, and also, as tests progress, by learning each individual's frequency concordance. The concordance learned by the GP manifests as a nonlinear interaction between frequency and intensity and results in threshold curves that are diagnostic for hearing ability.

A CCC value was then computed between the hearing thresholds of the two ears for each individual in the database (“paired”), again resulting in approximately 1.1 million values. The histogram of these values can be seen in Fig. 2A. The median of this distribution is $.515$, indicating even more information between the two ears than between adjacent frequencies on the same ear. As a means of comparison, CCC values were also computed between the left ear of a randomly selected individual and the right ear of a different randomly selected individual (“unpaired”) from the database. A second set of approximately 1.1 million values was generated in this way, as can be seen in Fig. 2B. The histogram for unpaired ears was nearly symmetric and centered slightly positive of zero, with a median of $.0757$. Knowing the distribution of paired hearing thresholds across the population at large appears to provide some information about the likely threshold

Fig. 1 (A) Audiogram thresholds for all the adjacent frequencies in the same individual (“adjacent”) yield high positive concordance. (B) Audiogram thresholds from arbitrary frequencies throughout the population show no average concordance.

Fig. 2 (A) Pairs of audiogram thresholds derived from ears in the same individual (“paired”) yield high positive concordance. (B) Pairs of audiogram thresholds deriving from ears in different individuals

of any particular individual. The more significant finding, however, is that knowing the hearing function of one ear provides a considerable amount of information about its contralateral counterpart. It is this shared variability between the ears that is exploited by conjoint GP audiogram estimation in order to accelerate testing.

Although the NIOSH database is large, and therefore informative about the underlying population from which its samples are drawn, that underlying population only reflects American working-age individuals at some risk for damaging their hearing on the job. Other populations, including retirees, children, and non-Americans, might be expected to demonstrate other distributions of shared variability in hearing functions between the ears. A method capable of learning this shared variability for each individual would be able to accommodate such differences. Individualization of this sort reflects the design intentions of machine-learning audiometry.

A representative run of the bilateral audiogram estimation algorithm can be seen in Fig. 3. The ground truth for this figure was an asymmetric hearing loss case identified by older-normal hearing in Ear 1 and metabolic + sensory hearing loss in Ear 2. The first 15 tones were selected via Halton sampling in order to improve the stability of the GP model. Subsequent tones were sampled using BALD. The samples selected via BALD tend to cluster around the predicted threshold, where they are more informative about the true audiogram threshold. Note that after 14 tone deliveries, the conjoint GP model has not yet identified the microstructure in the older-normal ear and is not particularly confident about the threshold location in the metabolic + sensory ear. After 98 tone deliveries, the model has correctly identified the microstructure in the older-normal ear and has both accurately and confidently identified the threshold in the metabolic + sensory ear. The conjoint estimation procedure therefore appears to yield credible estimates of canonical hearing functions.

(“unpaired”) yield concordance much closer to 0, though still with a tendency toward positive values.

Figure 4 shows the mean threshold error per iteration for Case 1, which was defined as having the older-normal phenotype in both ears. Note that both conjoint approaches outperform the disjoint approach for Ear 2, particularly in the early iterations. This example for Ear 1 shows that the disjoint approach is capable of occasionally obtaining exactly the right single observation to substantially reduce its error, putting it ahead of the other sampling strategies for the subsequent few samples. In general, however, the disjoint strategy lags behind the conjoint strategies, though given sufficient samples, they all converge to relatively low threshold errors.

The mean number of iterations required for each approach to achieve a 5-dB mean threshold error separately in each ear of the bilateral normal-hearing case ($n = 100$) is summarized in Table 1. Ear 2 was always sampled second, and with this phenotype, never generated prediction error greater than 5 dB in the unconstrained conjoint condition (i.e., error was less than 5 dB after the first tone for all 100 simulations). Conjoint approaches required significantly fewer samples to achieve the 5-dB threshold than disjoint approaches ($p < 10^{-5}$, uncorrected permutation tests with 10,000 resamples), with the exception of alternating conjoint versus disjoint for Ear 1 ($p = .58$, uncorrected permutation test with 10,000 resamples). The alternating conjoint strategy appears to perform no worse than disjoint in this case, which may be due to the prior mean including a structure more similar to the older-normal phenotype than to the other phenotypes. Both conjoint approaches in this case tend to exhibit higher standard deviations than the disjoint approach. This outcome probably results because initial differences in the Halton sampling algorithm reinforce the constant-threshold belief more in some iterations than others. The constant-threshold belief becomes stronger with more evidence, as would be the case if samples were shared among ears.

Fig. 3 Example of the posterior mean, representing the full audiogram, for simulated asymmetric hearing loss, as estimated by the unconstrained conjoint GP audiogram estimation model. In all images, the predicted probability of tone detection is shown in grayscale. Blue plusses are

heard stimuli, and red diamonds are unheard stimuli. The true threshold from the simulation is shown in purple. (Top) Posterior means after 14 samples. (Bottom) Posterior means after 98 samples.

Figure 5 shows the mean threshold error per iteration for Case 2, which was defined as having the older-normal phenotype in one ear and the metabolic + sensory phenotype in the other ear. Note that both conjoint approaches outperform the disjoint approach, particularly in the metabolic + sensory ear. The unconstrained conjoint method occasionally chooses to sacrifice some early performance in the older-normal ear in

exchange for faster convergence in the metabolic + sensory ear. The unconstrained approach is able to make this choice because the model uncertainty is typically higher in early iterations on the metabolic + sensory phenotype than it is on the older-normal phenotype (Fig. 3).

The mean number of iterations per ear required for each model to achieve a 5-dB mean error in the asymmetric hearing

Fig. 4 (A, B) Mean threshold error per iteration with Case 1, no hearing loss. Insets show the threshold audiogram shapes.

Table 1 Mean numbers of iterations (with standard deviations) required to achieve the 5-dB threshold error for each ear, Case 1

	Ear 1: Older-Normal	Ear 2: Older-Normal
Unconstrained Conjoint	11.1 ± 4.3*	1 ± 0*
Alternating Conjoint	16.7 ± 3.5	1.07 ± 0.70*
Disjoint	16.9 ± 2.5	19.7 ± 3.0

No older-normal unconstrained conjoint configuration tested for Ear 2 exceeded the 5-dB tolerance after the first tone. Test results significantly different from the disjoint condition for the same ear at $p < 10^{-5}$ are indicated by *.

loss case is summarized in Table 2. Conjoint approaches required significantly fewer samples to achieve the 5-dB threshold than disjoint approaches ($p < 10^{-5}$, uncorrected permutation tests with 10,000 resamples). Although the unconstrained conjoint method tends to sacrifice some early older-normal performance for faster metabolic + sensory convergence, the overall number of tones required to achieve convergence in both ears is lower for the unconstrained conjoint approach than for the alternating conjoint approach. The limiting factor for all three methods was identifying the metabolic + sensory phenotype in the second ear because it differed more from the expectations of the model than did the older-normal phenotype. The conjoint configurations still performed better than disjoint, with unconstrained conjoint requiring on average only about 90% of the tone samples needed by alternating conjoint, and only about 60% of the tone samples required by disjoint, in order to achieve convergence.

Figure 6 shows the mean threshold error per iteration for Case 3, which was defined as having the metabolic + sensory phenotype in one ear and the sensory phenotype in the other. Once again, both conjoint approaches outperform the disjoint approach, particularly in the metabolic + sensory ear. In this case, the unconstrained conjoint approach can leverage its ability to

choose in which ear to deliver stimuli and achieves faster convergence than both other approaches. All three models again require more samples to identify pathological phenotypes. This result suggests that further efficiency improvement may be achieved by using a more informative prior mean.

The mean number of iterations per ear required for each model to achieve the 5-dB average threshold error in the symmetric hearing loss case is summarized in Table 3. Conjoint approaches required significantly fewer samples to achieve the 5-dB threshold than disjoint approaches ($p < 10^{-5}$, uncorrected permutation tests with 10,000 resamples). Unlike in Case 2, the unconstrained conjoint approach can leverage its knowledge of interear correlation to substantially improve the time to convergence in both ears by changing the distribution of ear samples. As a result, the unconstrained conjoint method is able to converge in both ears faster than the alternating conjoint approach can converge in either ear. Once again, the unconstrained conjoint approach converges in both ears using fewer than 60% of the samples required in the disjoint approach.

The mean number of tones required for each of the approaches to achieve the 5-dB threshold error in both ears for each phenotype can be seen in Table 4. Numbers here were selected as the largest tone count at which the mean threshold error per iteration crossed below 5 dB in both ears. It is possible for the conjoint GP approach to start with a “lucky guess” of the true audiogram and exhibit a mean threshold error below 5 dB for early iterations but to have the threshold error increase in later iterations. Presenting an average of the final crossing below convergence threshold provides for a fairer comparison. To summarize the results of the three presented cases, both conjoint methods outperform the disjoint approach on average for every configuration. Furthermore, the constant-mean assumption performs substantially better on older-normal phenotypes than it does on any of the pathological phenotypes.

Fig. 5 (A, B) Mean threshold error per iteration with Case 2, asymmetric hearing loss. Insets show the threshold audiogram shapes.

Table 2 Mean numbers of iterations (with standard deviations) required to achieve the 5-dB threshold error for each ear, Case 2

	Ear 1: Older Normal	Ear 2: Metabolic Sensory
Unconstrained Conjoint	9.72 ± 5.3*	28.0 ± 5.8*
Alternating Conjoint	9.63 ± 7.7*	31.4 ± 6.6*
Disjoint	16.6 ± 2.5	46.0 ± 9.3

Test results significantly different from the disjoint condition for the same ear at $p < 10^{-5}$ are indicated by *.

Also included in Table 4 are time-to-criterion values in relative terms. Regardless of the phenotype pairings, the unconstrained conjoint approach requires on average approximately 60% of the samples required by the disjoint approach. One interesting observation made clear from this analysis is that the performance of the alternating conjoint method is closest to that of the unconstrained conjoint method in the case of asymmetric hearing loss. This outcome occurs because the conjoint model learns that there is relatively little correlation between the two ears and that to learn a good audiogram estimation, it must split its samples more evenly between the two ears. The unconstrained conjoint and alternating conjoint approaches therefore exhibit similar querying strategies in the asymmetric case, whereas in the older-normal and symmetric cases their querying strategies are quite different. This difference illustrates the key advantage of conjoint estimation, in which multidimensional queries are optimized for the bilateral phenotype of each individual as the test progresses.

As a feasibility check, unconstrained conjoint hearing threshold estimation was performed bilaterally on a normal-hearing human listener, and the results were compared to those from disjoint estimation. The procedures for disjoint estimation were exactly as described previously (Song et al., 2015b). The procedures for conjoint estimation matched the sound delivery and

response characteristics of the disjoint procedure, but tone selection and hearing threshold estimation were performed according to the principles developed for the conjoint simulations above. The results of this human test can be seen in Fig. 7. Threshold values were still converging in each ear after 20 tones in the disjoint case but were very near their final values after 20 tones in the conjoint case. This finding is consistent with the simulation results and reflects high feasibility that the projected gains in efficiency provided by conjoint audiometric testing will be realized in humans.

Discussion

Behavioral assessments of perceptual and cognitive processes represent organ-level system-identification procedures for the brain. Constructing models for individual brains faces the fundamental limitation whereby data to fit the model must be accumulated sequentially. This limitation severely constrains the sophistication of models that can be estimated in reasonable amounts of time. Procedures to improve the efficiency of this process by optimally selecting stimuli in real time have been developed repeatedly in multiple domains, often with substantial improvements in speed of model convergence (Bengtsson et al., 1997; DiMattina, 2015; Kontsevich & Tyler, 1999; Lesmes et al., 2006; Myung & Pitt, 2009; Shen & Richards, 2013; Watson, 2017).

Physiological assessments of neural function represents tissue-level system-identification procedures for the brain, and these procedures have evolved largely independently of behavioral methods. Often the total time of data collection is more constrained for neurophysiology than for behavior, and many advanced system identification procedures for efficiently modeling neuronal function in real time have been developed (Benda, Gollisch, Machens, & Herz, 2007; DiMattina &

Fig. 6 (A, B) Mean threshold error per iteration with Case 3, symmetric hearing loss. Insets show the threshold audiogram shapes.

Table 3 Mean numbers of iterations (with standard deviations) required to achieve the 5-dB threshold error for each ear, Case 3

	Ear 1: Metabolic Sensory	Ear 2: Sensory
Unconstrained Conjoint	30.5 ± 6.1 *	29.2 ± 5.0 *
Alternating Conjoint	31.3 ± 5.9 *	33.4 ± 5.3 *
Disjoint	42.5 ± 8.6	48.3 ± 9.3

Test results significantly different from the disjoint condition for the same ear at $p < 10^{-5}$ are indicated by *.

Zhang, 2013; Lewi, Butera, & Paninski, 2009; Paninski, Pillow, & Lewi, 2007; Pillow & Park, 2016), including the use of GP methods (Park, Horwitz, & Pillow, 2011; Pillow & Park, 2016; Rad & Paninski, 2010; Song et al., 2015a). One of the key advantages GPs provide is the ability to titrate prior beliefs finely under the same estimation architecture, leading to procedures that can be both flexible and efficient (Song et al., 2017; Song et al., 2018; Song et al., 2015b). Applying such tools to behavioral assessment provides similar benefits.

The conventional audiogram represents a relatively simple multidimensional behavioral test typically evaluated with weak prior beliefs. As a result, considerable audiometric information is usually discarded that could be used to improve audiogram test accuracy and speed. The large amounts of paired and unpaired ear data in the NIOSH database indicate that information exists from the contralateral ear that could be quite useful for incorporating into measures of the ipsilateral ear. The Hughson–Westlake procedure provides a very limited mechanism to do so, since the only real flexibility in the test available to the clinician or experimenter is the starting sound level. Machine-learning audiometry, on the other hand, can exploit various degrees of prior information by design. Active GP estimation is able to determine correlations between variables in real time as data are accumulated. Although a person's two ears are not themselves physiologically linked, they do share many things in common, including the same or similar age, genetics, blood supply, downstream neural processes, accumulated acoustic exposure, accumulated toxin exposure, and so forth. GP inference therefore

represents an excellent method for exploiting these correlations for improving test accuracy and efficiency.

The present series of experiments demonstrated that incorporating subject responses to tones delivered to either ear dynamically during a test can reduce the testing time by up to 80% with an estimator designed to take advantage of this information. All previous audiograms have been designed to evaluate either each ear sequentially or the better ear through free-field sound delivery. The bilateral audiogram delivers clinically diagnostic data for each ear separately with nearly the efficiency of single-ear screening methods. Stimuli are delivered to each ear individually, but inference is drawn for both ears simultaneously. This procedure extends the roving-frequency procedure of machine-learning audiometry, already shown to deliver equivalent thresholds to the fixed-frequency Hughson–Westlake procedure (Song et al., 2015b), into a roving-ear procedure. The stimuli themselves are still pure tones and the instructions to participants are unchanged: Respond every time you hear a tone as soon as you hear it. Therefore, this new testing procedure is completely backward-compatible with current measurement methods. Because it also returns tone detection threshold estimates, the test results and interpretations are also completely backward-compatible. The true value of the technique, however, does not derive simply because it can deliver a faster threshold audiogram.

As has been documented previously, machine-learning audiometry delivers continuous threshold estimates as a function of frequency (Song et al., 2015b), and psychometric spread estimates in addition to threshold (Song et al., 2017; Song et al., 2018). The theoretical extension described here of conjoining two stimulus input domains has the distinct advantage of speeding a full estimation procedure considerably. To obtain even a few psychometric threshold and spread estimates from both ears using the best extant estimators requires hours of data collection (Allen & Wightman, 1994; Buss, Hall, & Grose, 2006, 2009). The bilateral machine-learning audiogram obtains these estimates in minutes by defining a middle ground between estimating two sequential 2-D psychometric fields and performing a full 4-D psychometric field estimation. Both of these approaches would take longer because of shared variability among the input dimensions (i.e.,

Table 4 Mean numbers of tones and percentages of tones, relative to disjoint, required across both ears for each of the three models to achieve better than the 5-dB mean threshold error

		Older Normal	Asymmetric	Symmetric
Unconstrained Conjoint	Tone count	11.5 ± 4.5	27.9 ± 3.9	32.1 ± 4.7
	Relative to disjoint	60.7% ± 22.6%	58.4% ± 8.5%	57.8% ± 8.5%
Alternating Conjoint	Tone count	17.3 ± 4.9	32.9 ± 5.2	36.1 ± 4.8
	Relative to disjoint	71.5% ± 24.6%	65.6% ± 11.3%	86.9% ± 8.7%
Disjoint	Tone count	19.9 ± 2.4	46.0 ± 6.1	55.0 ± 8.2

Fig. 7 (A, B) Thresholds estimated after 10, 20, and 40 tones actively sampled to independently evaluate the left and right ears in a human subject. The threshold curves are evolving throughout tone delivery. (C, D) Thresholds estimated after 10, 20, and 40 actively sampled tones selected to evaluate a joint function of the left and right ears under the assumption that they are not correlated. The threshold curves have all achieved nearly their final values after 20 tones per ear.

interactions among the model predictors) that a conjoint GP can exploit. Sparseness in the data distributions for models of higher dimensionalities may lead to practical limits in estimating such models, although the dynamic dimensionality reduction options available to GPs, such as conjoint estimation, appear to be able to extend practical estimation to relatively high dimensionality.

Future advantages will derive from conjoining additional input domains. In audiometry, for example, if appropriate transducers are connected to a patient, conjoint air conduction and bone conduction audiometry could be performed simultaneously, thereby speeding both tests. Even more significant, by conjoining tone stimuli with contralateral masking stimuli, every audiogram can become a masked audiogram, automatically accommodating asymmetric hearing loss in real time as data are collected. The testing time would not be noticeably increased from an unmasked audiogram for those individuals with symmetric hearing without need of masking. Conjoint ipsilateral masking makes an even more compelling case, such that hearing capability could potentially be assessed dynamically in suboptimal acoustic environments. Because conjoint methods scale modularly, adding other auditory tests such as DPOAEs and speech intelligibility to standard pure-tone audiometry can be done automatically, as well. Professional human expertise will be critical for ensuring that valid data are collected, but the complexities of high-dimensional diagnostic input domains can be navigated efficiently and effectively by augmenting clinician or experimenter expertise with advanced machine-learning algorithms.

The focus of this study has been on improving psychometric field estimation speed over conventional methods. The nature of the GP estimator at the core of machine-learning audiometry, however, can accommodate other advances yet to be realized. For example, nonstationarities in the underlying function, such as attention lapses or criterion drift, can introduce estimation errors. A standard approach to account for these conditions in parametric estimators involves introducing additional parameters to model them directly (Prins, 2012; Wichmann & Hill, 2001). This method is available to the GP estimator if the likelihood function is modified to accommodate nonstationarities. Other options exist for the GP, however, such as replacing the linear kernel with a squared exponential kernel. This change retains all other features of GP estimation but results in psychometric functions resembling nonparametric estimates (Miller & Ulrich, 2001; Zychaluk & Foster, 2009). Such modularity allows users a tremendous range of options across which to design an effective estimator. In addition, the malleability of the GP for incorporating strong or weak prior beliefs further enables speed to be traded off as desired against robustness, all within the same estimation framework. Together, this modularity and malleability are what enable GP estimation—a semiparametric method as presented here—to achieve both the efficiency of a parametric estimator and the flexibility of a nonparametric estimator. These native capabilities, along with mathematical extensions such as multiplexed and conjoint estimation, yield substantial advantages to estimation procedures for a wide array of behavioral processes.

Conclusion

Bilateral audiometry, the first implementation of conjoint psychometric testing, can achieve its potential to return accurate results in significantly less time than sequential disjoint tests. The time required to acquire audiograms in both ears conjointly of a simulated participant was as little as 60% of the time (i.e., an 80% speedup) required to achieve the same degree of accuracy with disjointly acquired data. The sound stimuli and participant instructions are identical to those in current testing procedures, as are the results returned, so test delivery and interpretation are unchanged. Conjoint machine-learning audiometry also has great potential to incorporate additional testing procedures directly into the methodology with alternate kernel design, eventually leading to unified tests that actively customize stimulus delivery and diagnostic inference for each subject. The future of audiologic testing involves working up patients dynamically via multiple subjective and objective testing modalities that mutually reinforce one another to construct a thorough assessment of a patient's hearing. Furthermore, the general principles of conjoint psychometric testing can be applied more broadly in the behavioral sciences

in order to unify the individual components of perceptual or cognitive test batteries into a single efficient test.

Author note Funding for this project was provided by the Center for Integration of Medicine and Innovative Technology (CIMIT) and the National Center for Advancing Translational Sciences (NCATS) Grant UL1-TR002345. D.L.B. has a patent pending on the technology described in this article.

References

- Allen, P., & Wightman, F. (1994). Psychometric functions for children's detection of tones in noise. *Journal of Speech, Language, and Hearing Research*, 37, 205–215.
- American National Standards Institute. (1978). Methods for manual pure-tone threshold audiometry (Standard No. ANSI S3.21-1978). Washington, DC: Author.
- Benda, J., Gollisch, T., Machens, C. K., & Herz, A. V. (2007). From response to stimulus: Adaptive sampling in sensory physiology. *Current Opinion in Neurobiology*, 17, 430–436.
- Bengtsson, B., Olsson, J., Heijl, A., & Rootzén, H. (1997). A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmologica*, 75, 368–375.
- Buss, E., Hall, J. W., III, & Grose, J. H. (2006). Development and the role of internal noise in detection and discrimination thresholds with narrow band stimuli. *Journal of the Acoustical Society of America*, 120, 2777–2788.
- Buss, E., Hall, J. W., III, & Grose, J. H. (2009). Psychometric functions for pure tone intensity discrimination: Slope differences in school-aged children and adults. *Journal of the Acoustical Society of America*, 125, 1050–1058.
- Carhart, R., & Jerger, J. (1959). Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech and Hearing Disorders*, 24, 330–345.
- Cohen, D. J. (2003). Direct estimation of multidimensional perceptual distributions: assessing hue and form. *Perception & Psychophysics*, 65, 1145–1160.
- Coren, S. (1989). Summarizing pure-tone hearing thresholds—The equipollence of components of the audiogram. *Bulletin of the Psychonomic Society*, 27, 42–44.
- Coren, S., & Hakstian, A. R. (1990). Methodological implications of interaural correlation: Count heads not ears. *Perception & Psychophysics*, 48, 291–294.
- DiMattina, C. (2015). Fast adaptive estimation of multidimensional psychometric functions. *Journal of Vision*, 15(9), 5. <https://doi.org/10.1167/15.9.5>
- DiMattina, C., & Zhang, K. (2013). Adaptive stimulus optimization for sensory systems neuroscience. *Frontiers in Neural Circuits*, 7, 101. <https://doi.org/10.3389/fncir.2013.00101>
- Divenyi, P. L., & Haupt, K. M. (1992). In defense of the right and left audiograms: A reply to Coren (1989) and Coren and Hakstian (1990). *Perception & Psychophysics*, 52, 107–110.
- Dubno, J. R., Eckert, M. A., Lee, F.-S., Matthews, L. J., & Schmiedt, R. A. (2013). Classifying human audiometric phenotypes of age-related hearing loss from animal models. *Journal of the Association for Research in Otolaryngology*, 14, 687–701.
- Duvenaud, D. (2014). *Automatic model construction with gaussian processes* (Doctoral dissertation). University of Cambridge, Cambridge, UK.
- Fechner, G. T. (1966). Elements of psychophysics (H. E. Adler, Trans.; D. H. Howes & E. C. Boring Eds.). New York, NY: Holt, Rinehart & Winston. (Original work published 1860).
- Fisher, R. A. (1925). Intraclass correlations and the analysis of variance. In *Statistical methods for research workers* (pp. 177–207). Edinburgh, UK: Oliver & Boyd.
- Gardner, J. R., Song, X., Weinberger, K. Q., Barbour, D., & Cunningham, J. P. (2015). Psychophysical detection testing with Bayesian active learning. In *Uncertainty and artificial intelligence: Proceedings of the thirty-first conference* (pp. 286–295). Corvallis, OR: AUAI Press.
- Garnett, R., Osborne, M. A., & Hennig, P. (2013). *Active learning of linear embeddings for Gaussian processes*. arXiv:1012.2599
- Gelman, A., Vehtari, A., Jylänki, P., Robert, C., Chopin, N., & Cunningham, J. P. (2014). *Expectation propagation as a way of life*. arXiv:1412.4869v2
- Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7, 701–702.
- Heijl, A., & Krakau, C. (1975). An automatic static perimeter, design and pilot study. *Acta Ophthalmologica*, 53, 293–310.
- Hosmer, D. W., & Lemeshow, S. (2013). The multiple logistic regression model. In *Applied logistic regression* (3rd ed., pp. 35–48). Hoboken, NJ: Wiley.
- Houlsby, N., Huszar, F., Ghahramani, Z., & Lengyel, M. (2011). *Bayesian active learning for classification and preference learning*. arXiv:1112.5745
- Hughson, W., & Westlake, H. (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. *Transactions of the American Academy of Ophthalmology and Otolaryngology*, 48, 1–15.
- International Organization for Standardization. (2010). ISO 8253-1:2010: Acoustics—Audiometric test methods—Part 1: Pure-tone air and bone conduction audiometry. Geneva, Switzerland: ISO.
- Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics: A practical introduction* (2nd ed.). London, UK: Elsevier Academic Press.
- Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39, 2729–2737.
- Kujala, J. V., & Lukka, T. J. (2006). Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, 50, 369–389.
- Lesmes, L. L., Jeon, S. T., Lu, Z. L., & Doshier, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. *Vision Research*, 46, 3160–3176.
- Lewi, J., Butera, R., & Paninski, L. (2009). Sequential optimal design of neurophysiology experiments. *Neural Computation*, 21, 619–687.
- Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3–12). New York, NY: Springer.
- Lin, L. I. (1989). A concordance correlation-coefficient to evaluate reproducibility. *Biometrics*, 45, 255–268. <https://doi.org/10.2307/2532051>
- Mahomed, F., Eikelboom, R. H., & Soer, M. (2013). Validity of automated threshold audiometry: A systematic review and meta-analysis. *Ear and Hearing*, 34, 745–752.
- Masterson, E. A., Tak, S., Themann, C. L., Wall, D. K., Groenewold, M. R., Daddens, J. A., & Calvert, G. M. (2013). Prevalence of hearing loss in the United States by industry. *American Journal of Industrial Medicine*, 56, 670–681.
- Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman–Kärber method. *Perception & Psychophysics*, 63, 1399–1420.
- Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In J. Breese & D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence* (pp. 362–369). Corvallis, OR: AUAI Press. arXiv:1301.2294

- Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116, 499–518. <https://doi.org/10.1037/a0016104>
- Paninski, L., Pillow, J., & Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in Brain Research*, 165, 493–507.
- Park, M., Horwitz, G., & Pillow, J. W. (2011). Active learning of neural response functions with Gaussian processes. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *NIPS '11: Proceedings of the 24th International Conference on Neural Information Processing Systems* (pp. 2043–2051). New York, NY: Curran Associates.
- Pillow, J. W., & Park, M. J. (2016). Adaptive Bayesian methods for closed-loop neurophysiology. In A. El Hady (Ed.), *Closed loop neuroscience* (pp. 3–18). Amsterdam, The Netherlands: Elsevier.
- Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, 12(6), 25. <https://doi.org/10.1167/12.6.25>
- Rad, K. R., & Paninski, L. (2010). Efficient, adaptive estimation of two-dimensional firing rate surfaces via Gaussian process methods. *Network*, 21, 142–168.
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
- Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report 1648). Madison, WI: University of Wisconsin–Madison. Retrieved from burrsettles.com/publications
- Shen, Y., & Richards, V. M. (2013). Bayesian adaptive estimation of the auditory filter. *Journal of the Acoustical Society of America*, 134, 1134–1145.
- Song, X. D., Garnett, R., & Barbour, D. L. (2017). Psychometric function estimation by probabilistic classification. *Journal of the Acoustical Society of America*, 141, 2513–2525.
- Song, X. D., Sukesan, K. A., & Barbour, D. L. (2018). Bayesian active probabilistic classification for psychometric field estimation. *Attention, Perception, & Psychophysics*, 80, 798–812. <https://doi.org/10.3758/s13414-017-1460-0>
- Song, X. D., Sun, W., & Barbour, D. L. (2015a). *Rapid estimation of neuronal frequency response area using Gaussian process regression*. Article presented at the annual conference of the Society for Neuroscience, Chicago, IL.
- Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., & Barbour, D. L. (2015b). Fast, continuous audiogram estimation using machine learning. *Ear and Hearing*, 36, e326–e335.
- Watson, A. B. (2017). QUEST+: A general multidimensional Bayesian adaptive psychometric method. *Journal of Vision*, 17(3), 10:1–27. [10.1167/17.3.10](https://doi.org/10.1167/17.3.10)
- Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63, 1293–1313. <https://doi.org/10.3758/BF03194544>
- Williams, C. K., & Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 1342–1351.
- Zychaluk, K., & Foster, D. H. (2009). Model-free estimation of the psychometric function. *Attention, Perception, & Psychophysics*, 71, 1414–1425. <https://doi.org/10.3758/APP.71.6.1414>