Contrast Response Function Estimation with Nonparametric Bayesian Active Learning

Dom CP Marticorena[1], Quinn Wai Wong[1], Jake Browning[2], Ken Wilbur[2], Samyukta Jayakumar[3], Pinakin Davey[4], Aaron R. Seitz[5], Jacob R. Gardner[6], Dennis L. Barbour[1]

[1]Department of Biomedical Engineering, Washington University, 1 Brookings Drive, St. Louis, MO 63130

[2]Department of Computer Science and Engineering, Washington University, 1 Brookings Drive, St. Louis, MO 63130

[3]Department of Psychology, University of California, Riverside, 900 University Ave., Riverside, CA 92521

[4]College of Optometry, Western University of Health Sciences, 701 E Second St., Pomona, CA 91766

[5]Department of Psychology, 125 Forsyth Street, Northeastern University, Boston, MA 02115

[6]Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut St., Philadelphia, PA 19104

Abstract

Multidimensional psychometric functions can typically be estimated nonparametrically for greater accuracy or parametrically for greater efficiency. By recasting the estimation problem from regression to classification, however, powerful machine learning tools can be leveraged to improve both accuracy and efficiency simultaneously. Contrast Sensitivity Functions (CSFs) are behaviorally estimated curves that provide insight into both peripheral and central visual function. They are impractically long to be used in many clinical workflows without compromises of some sort, however, such as sampling only a few spatial frequencies or making strong assumptions about the shape of the function. This paper describes the development of the Machine Learning Contrast Response Function (MLCRF) estimator, which quantifies the expected probability of success in performing a contrast detection or discrimination task. A machine learning CSF can then be derived from the MLCRF. Using simulated eyes created from canonical CSF curves and actual human contrast response data, the accuracy and efficiency of the MLCSF was evaluated in order to determine its potential utility for research and clinical applications. With stimuli selected randomly, the MLCSF estimator converged toward ground truth. With optimal stimulus selection via Bayesian active learning, convergence was about an order of magnitude faster, requiring only tens of stimuli to achieve reasonable estimates. Inclusion of an informative prior provided no discernible advantage to the estimator as configured. The MLCSF exhibits performance characteristics on par with state-of-the-art CSF estimators and therefore should be explored further to uncover its full potential.

Precis

Machine learning classifiers enable accurate and efficient contrast sensitivity function estimation with item-level prediction for individual eyes.

Keywords

Contrast sensitivity, psychophysics, machine learning, Bayesian modeling, active learning

Introduction

Visual contrast sensitivity reflects both peripheral and central visual processing ability. As such, it is useful for diagnosing a variety of visual disorders. The simplest, easiest, cheapest and most portable way to quantify this ability is by querying directly—delivering appropriate visual stimuli and recording the resulting behavioral responses. Even with steady advances in quantifying physiological biomarkers of the retina and brain (Calderone et al., 2013; Yarmohammadi et al., 2016), no alternate diagnostic pathway is foreseen that will completely supplant psychophysical testing for evaluating visual system function.

As with all psychophysical tests, however, estimating contrast sensitivity functions (CSFs) requires serial behavioral data acquisition to estimate latent variables, leading to impractically long acquisition times due to high variance in the underlying physiological processes. While full CSFs can therefore have significant clinical value, rapid psychophysical screenings that sacrifice quantitative precision are often more desirable for practical reasons.

Methods have been proposed for accelerating CSF testing while preserving sufficient accuracy for clinical decision making (Gu et al., 2016; Lesmes et al., 2010; Wang et al., 2016). Because of intrinsic noisiness, speeding up the estimates typically requires making assumptions—incompletely justified in most cases—about some model parameters in order to reduce the degrees of freedom of the model to be learned (Treutwein & Strasburger, 1999). This process results in less flexible models that might work well on average but fail to capture nuances of some individual curves, especially local nuances such as notches (Tahir et al., 2009; Woods et al., 1996).

For example, a common modern assumption about the shape of the CSF exploited in some of the most advanced CSF estimators is a truncated parabola (Zhao et al., 2021). While this shape visually reflects common CSF trends and has been shown to lead to more accurate estimates compared to alternate parametric forms (Watson & Ahumada, 2005), the parameters of the model do not provide mechanistic insight into the underlying physiological construct. Reductionist modeling in this case, therefore, appears to primarily be useful for achieving sufficient test accuracy with fewer data samples, allowing for more rapid tests. Additionally, reducing the resulting model to a small number of parameters does provide a useful shorthand for researchers and clinicians to conceptualize and communicate general CSF shape.

The continued expansion of machine learning success over the past two decades can be at least partially attributed to improved ability to constrain highly flexible models by discerning complex patterns in data (Basha & Rajput, 2019). While some of the most prominent successes involve

extremely large data sets (Zhou et al., 2017), similar advances can also be applied to much smaller data sets of various types (Kokol et al., 2021). If more flexible CSF models can be trained in reasonable amounts of time, then more appropriate estimates can be achieved for unusual phenotypes not fit well by a traditional, assumption-laden functional form.

Exactly this capability has been achieved with the threshold audiogram, an auditory analog of the CSF (Cox & de Vries, 2021; Schlittenlacher et al., 2018; Song et al., 2015), as well as threshold perimetry tests (Chesley & Barbour, 2020). The current study summarizes the application of similar principles to construct a CSF estimator, along with a preliminary assessment of estimator performance. The result is a highly flexible nonparametric estimator that has the potential to capture a wide variety of individual CSF curves in practical amounts of time. Perhaps most importantly, this single estimator can be tuned toward either efficiency or accuracy, depending on the application or testing time available.

Modeling Framework

Background

The contrast sensitivity function is defined as the performance threshold of a behavioral task in response to manipulations of the spatial frequency and visual contrast of a visual stimulus (Ginsburg, 2003). Other stimulus parameters such as mean luminance, size and visual eccentricity are typically held constant or manipulated in stepwise fashion, resulting in a series of CSF curves in the latter case (Kolb et al., 1995). The CSF curve reflects a particular success probability contour on the underlying multidimensional psychometric function, or psychometric field. Considering for a moment only a single spatial frequency, signal detection theory predicts that the ability to detect the presence of a visual pattern is a monotonically increasing function of contrast (Green & Swets, 1966; Kingdom & Prins, 2010). The resulting one-dimensional psychometric function $\psi(\kappa)$ is modeled as a sigmoid in contrast κ.

Because the entire psychometric function $\psi(\kappa)$ is rarely of interest in clinical applications, adaptive staircase methods have long been adopted to estimate only the threshold of this sigmoid (King-Smith, 1984; Levitt, 1971; Treutwein, 1995). Such methods exploit assumptions made about the spread (or equivalently, slope) of the sigmoid in order to set appropriate step sizes that probe either side of the threshold. While fast, these nonparametric methods are inherently unscalable to higher dimensions (i.e., adding independent variables contributing to a psychometric field) because they offer no efficiency gain with increasing dimensionality. In other words, they have little capacity to incorporate knowledge gained at one combination of independent variables to improve the threshold estimate at

another. The result is a highly flexible but relatively inefficient CSF estimation method. Such methods have, however, been shown to exhibit superior estimation accuracy relative to parametric methods (Watson & Ahumada, 2005).

Incorporating the psychometric spread into a psychometric curve estimator as a free parameter is possible (Leek, 2001), though this typically requires about an order of magnitude more data than when estimating threshold alone (Kingdom & Prins, 2010; King-Smith & Rose, 1997; Kontsevich & Tyler, 1999). These methods are scalable to multiple dimensions (DiMattina, 2015; Lesmes et al., 2006), but the gains to efficiency by sharing inference across all the input variables is generally negated by the additional data requirements to estimate the full function. Therefore, estimating a full psychometric field of contrast responses by combining a parametric estimate of psychometric spread with a low-order parameterization across spatial frequency would result in a relatively inflexible estimator suffering from poor accuracy and efficiency relative to alternatives. Likely for these reasons, no estimator of this type is known to have been developed for CSF estimation.

Because the CSF is a threshold curve, parameterizing it directly and making appropriate assumptions about psychometric spread represents a reasonable approach to gain sufficient accuracy and practical efficiency. Indeed, multiple parametric forms of the CSF have been evaluated for accuracy and efficiency (Rohaly & Owsley, 1993; Watson & Ahumada, 2005). A balance between model complexity and ability to map individual CSFs, along with some interpretability of model parameters, provides the truncated parabola with an advantage over other model forms. These advantages were exploited to develop the efficient quickCSF method (Lesmes et al., 2010), as well as modern extensions with even more desirable characteristics (Zhao et al., 2021).

A practical and effective modern computational CSF estimation therefore exists that appears to balance efficiency and flexibility, but there may still be room for improvement. For example, while demonstrably better than other models, the truncated parabola is unable to adequately capture the natural detail of some individual CSFs (Chung & Legge, 2016; Rohaly & Owsley, 1993). Further, despite the interpretability of the truncated parabola parameters regarding the shape of the CSF curve itself, any relationship to underlying physiological variables has yet to be established, compromising a mechanistic interpretation of the resulting CSF models.

Nonparametric estimators historically exhibit great flexibility but poor efficiency, as described above. The machine learning field, however, has expended enormous efforts over the past 2 decades developing highly flexible estimators for a wide variety of applications. Exploiting such advances may pay dividends for CSF estimation. Indeed, by recasting the psychometric field estimation problem from one of multiple regression to one of probabilistic classification, new machine learning

tools can be brought to bear on this longstanding problem. These methods exhibit little apparent advantage for one-dimensional psychometric function estimation, but considerable advantages for multidimensional estimation (Song et al., 2017), particularly when accompanied by optimal task item selection in the form of Bayesian active learning (Song et al., 2018).

This study involves developing a probabilistic machine learning classifier for estimating a contrast response function, then extracting a CSF estimate from the resulting model. Canonical and real CSF curves are used to create generative models, which then provide simulated behavioral responses representing ground truth values. The ability of the new estimator to fit these known functions is evaluated for different stimulus presentation sequences and bias conditions. The performance of the algorithm is evaluated in terms of accuracy and efficiency.

Gaussian Process Classification

A full development of the following can be found in (Song et al., 2017). Briefly, define $f(\mathbf{x})$ to be a latent function defined on a continuous multidimensional space $\mathbf{x} \in \mathcal{X}$. A Gaussian process (GP) with its inherent mean and covariance functions represents a convenient means to encode prior knowledge about the latent function:

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

This knowledge then can be updated following new data collection according to Bayesian principles. We evaluate the GP over a finite sample of independent values $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x}_n\}$. In binary classification tasks the dependent variable can take on one of two values indicating failure or success: $y_i \in \{0,1\}$. The probability of success $p(y = 1|f)$ is modeled by a sigmoidal link function $\psi(\mathbf{x})$. This function is distributed according to a Bernoulli likelihood:

$$\psi(\mathbf{x}) \sim p(y = 1|f, \mathbf{x}) = \mathrm{Bernoulli}(\psi(f(\mathbf{x}))). \quad (2)$$

The general sigmoidal link function can be given by

$$\psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)\phi(x; \alpha, \beta), \quad (3)$$

where ϕ is a logistic function parameterized by threshold α and spread β, γ is a guess rate, and λ is a lapse rate (Wichmann & Hill, 2001). The threshold as a function of spatial frequency becomes the CSF. The guess rate for a two-alternative forced-choice task, for example, would be set to 0.5 and for a detection task might be set to 0. The current study simulated only detection tasks and fixed $\lambda = \gamma =$

0.01 to allow for a small proportion of lucky guesses and lapses of attention. The threshold and spread parameters α and β were learned by the algorithm for each person and spatial frequency.

The resulting machine learning contrast response function (MLCRF) estimator implements a multidimensional nonparametric probabilistic classifier subdividing the input domain of the psychometric field into subdomains of "task success" and "task failure." Recasting the modeling problem in terms of classification allows advances in powerful machine learning classification to be exploited. The CSF reflects a single fixed-probability contour on the full MLCRF, giving rise to the MLCSF as our functional estimate.

Bayesian Active Learning

A full development of the following can be found in (Song et al., 2018). Briefly, Bayes' rule is applied to compute updated posteriors upon observation of data $\{\mathbf{y}, \mathbf{X}\}$. Because of the nonlinear link function and use of the Bernoulli likelihood, the posteriors cannot be computed in closed form. They are thus estimated via approximate inference techniques, specifically variational inference (Hensman et al., 2015; Titsias, 2009).

Variational inference is a technique that finds the best approximation of the true posterior distribution from a family of simpler distributions by minimizing the Kullback-Liebler divergence between the approximate and true posteriors (Gardner et al., 2018). We can efficiently estimate the posterior distribution of the Gaussian process (GP) model, which when trained with all existing data can compute model updates for any new sample $\mathbf{x}^* \in \mathbf{X}^*$ defined over spatial frequency and visual contrast. Therefore, the new sample $\mathbf{x}^*$ that, upon observation, maximizes some utility function $U(\mathbf{x}^*)$ is optimal. We define an acquisition function for obtaining this sample as

$$A(\mathbf{x}^*) = \operatorname{argmax}_{\mathbf{x}^* \in \mathbf{X}^*} U(\mathbf{x}^* | \mathbf{X}, \mathbf{y}), \tag{4}$$

where $U(\cdot)$ is a utility function reflecting model quality. We implement uncertainty sampling by defining the utility function as the differential entropy calculated using the predictive mean μ and variance σ^2 of the model observed through the likelihood. The differential entropy H_{diff} can be given as:

$$H_{diff}(\mathbf{x}^*) = H\left[\Phi\left(\frac{\mu(\mathbf{x}^*)}{\sigma(\mathbf{x}^*)}\right)\right] - \frac{C}{\sqrt{\sigma^2(\mathbf{x}^*)+C^2}} \exp\left(\frac{-\mu^2(\mathbf{x}^*)}{2\sigma(\mathbf{x}^*)+C^2}\right), \tag{5}$$

where $H(\cdot)$ is the binary entropy function $H(x) = -x\log_2(x) - (1-x)x\log_2(1-x)$, $\Phi(\cdot)$ is the CDF of a standard normal, and C is a normalizing factor $C = \sqrt{\frac{\pi ln(2)}{2}}$, which affords the approximation of the second term in closed form (Houlsby et al., 2011).

This acquisition function finds the next best sample point $\mathbf{x}^*$ that maximizes the differential entropy, which is a proxy for information gain.

Methods

Simulations

Ground truth models for four canonical CSF phenotypes were constructed from textbook threshold curves, as depicted in **Figure 1** (Kalloniatis & Luu, 1995). These curves are maximally smooth in shape and occupy extremes in the domain of likely CSFs. Similarly, ground truth CSF models for seven neurotypical individuals and twelve individuals with schizophrenia were constructed from threshold values extracted during a CSF training regimen at discrete spatial frequencies (**Figure 2**).

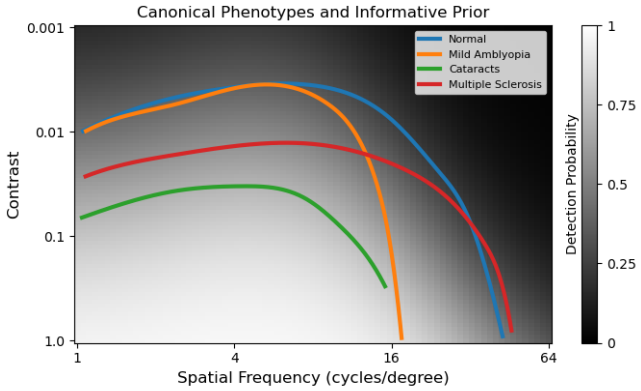

Figure 1: Canonical CSF curves representing typical contrast detection thresholds for four different disease phenotypes. These phenotypes illustrate the general shapes of CSFs as well as variations in loss of visual function across spatial frequency for different clinical conditions. Prior belief used for all informative prior conditions is shown in grayscale.

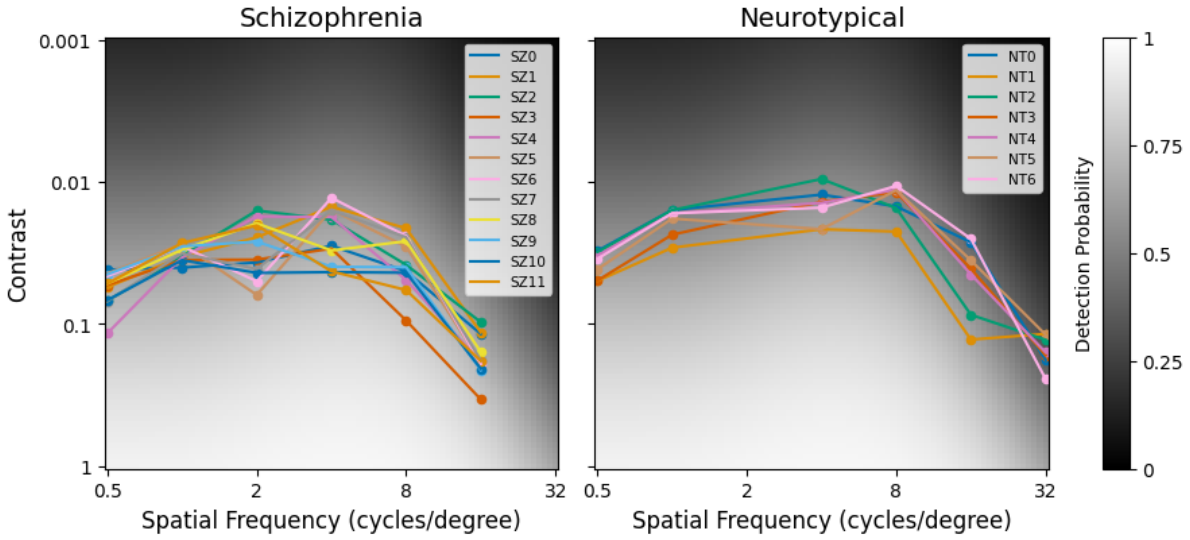

Figure 2: CSF curves estimated from an intermediate stage of a contrast training study for schizophrenic individuals (left) and participants with no known neurological disorder (right). These CSFs have more local variations than the overly smooth canonical curves. Prior belief used for all informative prior conditions is shown in grayscale.

In all cases, threshold curves as a function of spatial frequency were simulated by splines. The resulting threshold curves were used to create generative ground truth CRF models representing a visual pattern detection task as a function of spatial frequency and visual contrast. At each spatial frequency a one-dimensional sigmoidal psychometric curve was constructed using the four-parameter model given above:

$$\psi(\kappa; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)\phi(\kappa; \alpha, \beta), \tag{6}$$

including the standard logistic function

$$\phi(\kappa) = 1/(1 + \exp(-(\kappa - \alpha)/\beta). \tag{7}$$

In these equations κ is the contrast value, $\psi(\kappa) = p(y = 1|\kappa)$ is the probability of detection, α is the $\phi = 0.5$ detection threshold, and β is the psychometric spread. A plausible spread of 0.5 was used across all spatial frequencies (Zhao et al., 2021). The impact of spread on the behavior of this type of estimator across 2 orders of magnitude has been explored previously, and its precise value is unlikely to influence the initial conclusions drawn about estimator performance (Song et al., 2017, 2018). At any indicated contrast level, binary observations (i.e., detected or not detected) were generated by sampling from a Bernoulli distribution with success probability given by ψ. A total of 23 generative models formed the collection of simulated eyes evaluated in this study.

Psychometric Field for Contrast Response

The psychometric field defining the CRF over the domains of spatial frequency and contrast is given by $\psi(\omega, \kappa)$, while the underlying latent function is defined by $f(\omega, \kappa)$. With the setup described above, the entire estimation procedure for the CRF is reduced to learning a GP over the latent function. A GP is completely determined by its mean and covariance functions (Rasmussen & Williams, 2006). By selecting closed forms for the mean and covariance functions, estimation further reduces to updating the corresponding parameters, which are then combined with observed data to generate a posterior belief about the CRF. Because the parameters in question reflect the GP rather than the latent function, they represent hyperparameters of the overall model, yielding a formally semiparametric estimator for contrast response in this case.

A wide variety of functional forms can be selected for the mean and covariance functions to incorporate a set of assumptions about the form of the latent function. These form looser constraints than familiar parametric model forms, retaining flexibility to fit a wide variety of functional shapes while potentially fitting those shapes with fewer data than other estimator classes.

The GP mean function is assumed to be constant:

$$\mu(\mathbf{x}) = \mu((\omega, \kappa)) = c. \tag{9}$$

This may seem to be a counterintuitive choice because the shape of the CSF is known not to be flat. It is important to remember, however, that this function parameterizes the GP and not the latent function. It is essentially capturing the overall contrast sensitivity across all spatial frequencies. As such, this hyperparameter alone could be useful, for example, at distinguishing cataracts from normal vision.

Covariation along the contrast dimension is assumed to be linear. When combined with the logistic link function, the result is a sigmoid that is shifted and scaled to reflect the psychometric properties of contrast detection. Covariation in the spatial frequency dimension is assumed to be continuous and smooth. Parametric estimators cannot easily represent such broad constraints without incorporating many parameters, typically increasing data requirements to fit all of them. The form of the GP covariance function, on the other hand, can be compactly formulated to represent both of these assumptions:

$$K(\mathbf{x}, \mathbf{x}') = K((\omega, \kappa), (\omega', \kappa')) = {s_1}^2(\kappa\kappa') + {s_2}^2 \exp[-(\omega - \omega')/2l^2]. \tag{10}$$

The hyperparameters s_1 and s_2 represent scaling factors and l represents a length constant along the contrast dimension. Other kernel designs are possible, but this design has been particularly successful at estimating behavioral functions similar to the CSF (Song et al., 2017).

Implementation

All simulation, machine learning, and evaluation software was written in Python using major libraries PyTorch (*PyTorch 1.13*, n.d.) and GPyTorch (*GPyTorch 1.9.1*, n.d.).

Seventy-one logarithmically spaced spatial frequencies were used for the estimation such that $\omega \in [1\ 64]$ cycles per degree (i.e., 7 octaves) for experiment 1 and $\omega \in [0.5\ 32]$ cycles per degree for experiment 2. Contrasts used for the estimation included $\kappa \in [0.001\ 1]$ (i.e., 3 orders of magnitude) with 81 logarithmically spaced values. Therefore, the grid of input values used for modeling, prediction and quantification was discretized to have just under 6000 values. This resolution is arbitrary and could be modified as needed.

Samples were generated either via randomly drawing uniformly from all possible combinations of frequency and contrast on the evaluation grid, or from actively learning the most informative next sample for estimation. In both cases, two types of priors were selected to combine with newly observed data to form the posterior. One was uninformative, including no information about CSFs, and the other was informative, including canonical CSF shape information.

For the uninformative prior condition, the mean function of the GP was initialized with $c = 0$. Zero on the latent function maps to a probability of success of 0.5, implying that without any data, the estimator assumes maximum uncertainty about the shape of the CRF. The covariance function was initialized such that $s_1 = s_2 = l = 1$. The intention of this prior was to allow the sampled data to speak for itself in order to deliver a final estimate with few assumptions.

For the informative prior condition, 1000 uniformly random samples across spatial frequency and contrast were divided equally among the four canonical phenotypes observed in the range of [1 64] cycles per degree and labeled by the corresponding generative models. A single GP was fit to this entire set of observations. The scaling factor of the linear kernel was then multiplied by 0.4 in order to expand the transition between stimulus response regions. This manipulation “flattened the prior” to weaken the bias it injected into the model while still staying informative. The posterior mean of this GP was then used to initialize the mean function of all later GPs. This exact prior was used for experiment 1. For experiment 2, the GP was extrapolated down to 0.5 cycles per degree and used as a prior on [0.5 32] cycles per degree. The relationship between this prior and ground truth CSFs can be

seen in Figures 1 and 2, revealing a wide transition between behavioral response regions overlapping the threshold boundaries.

Two separate experiments were conducted to test the new estimator under different conditions. For experiment 1, four ground truth generative models were created from the four canonical phenotypes depicted in Figure 1 via high-density sampling and cubic spline fitting. Psychometric spreads were fixed at 0.5. These phenotypes were selected to demonstrate estimator performance under extremes of phenotypic variation. Four combinations of sampling methods (random, active) and prior selection (uninformative, informative) were used to acquire data from the generative models. After each new data point, the CRF was updated as a posterior defined over the entire input domain. The $\phi = 0.5$ contour of the predictive posterior mean of the CRF became the CSF estimate because this value forms the equiprobable boundary between the two response classes. At the octaves of spatial frequency from 1 cycle/degree, the root mean square error (RMSE) between the ground truth CSF and the estimated CSF was quantified. The estimated CSF was discretized to the nearest contrast grid value for this calculation. Each phenotype was evaluated separately for 10 repetitions and the average behavior summarized.

Because the canonical examples are overly smooth, experiment 2 made use of generative ground truth models taken from a cohort of neurotypicals and a cohort of schizophrenic individuals performing a contrast detection task as part of a previous study (Yaghoubi et al., 2022). The study was conducted across three sites: Weill Cornell Medicine (WCM), Nathan S. Kline Institute for Psychiatric Research (NKI), and the University of California, Riverside (UCR). Seven neurotypical (NT) participants and 12 patients with schizophrenia (SZ) were recruited for the study. The total number of participants at each site was as follows: WCM: 6 (4 males; age: mean = 33.5 yrs, SD = 8.48); NKI: 6 (3 males; age: mean = 45.6 yrs, SD = 9.54); and UCR: 7 (2 males; age: mean = 19.93 yrs, SD = 2.15). All subjects reported normal or corrected to normal vision.

The gamified training paradigm used in this experiment derived from (Deveau et al., 2014). The task was administered using an Apple iPad Pro 12.9 inch screen (second generation) at a luminance of 600 cd/m^2, resolution of 2732 × 2048 pixels and pixel density of 264 pixels per inch. The viewing distance of all participants from the screen was 20 inches. The iPads used at all sites were calibrated similarly to reduce the variance between each site where the study was conducted. The stimulus set consisted of Gabor patches (targets) at 6 spatial frequencies. An initial test was performed for each group of participants to approximate the maximum spatial frequency that could be perceiced. The spatial frequencies used for NT were 0.5, 1, 4, 8, 16 and 32 cycles per degree and for SZ were 0.5, 1, 2, 4, 8 and 16 cycles per degree, respectively. The stimuli were also presented in 8 orientations (0°, 22.5°,

45°, 67.5°, 90°, 112.5°, 135°, 157.5°). Gaussian windows of Gabors varied with σ between 0.25° and 1° and with phases (0°, 45°, 90°, 135°).

Each group of participants underwent a slightly different training procedure, i.e., SZ performed training for up to 40 sessions (1 session per day), with each session lasting approximately 30 minutes. NT performed 40 training sessions in 20 days, i.e., two sessions per day. Each session consisted of different blocks where Gabor patches at all 6 spatial frequencies were presented.

Each block lasted for 120 s where an array of targets with randomly selected orientation and increasing spatial frequency appeared all at once scattered across the screen (**Figure 3**). The contrast of the target was adaptively determined using a 3 down 1 up staircase. Contrast was decreased whenever 80% of the targets were selected and increased when fewer than 40% of the targets were selected with a 2.5 s per target time limit. Staircases were independently run on each spatial frequency across blocks of training. Spatially varying auditory feedback was given to the participants, i.e., low frequency tones corresponded with targets on the bottom of the screen whereas high frequency tones corresponded to stimuli at the top of the screen. Thus, the horizontal and vertical locations on the screen each corresponded to a unique tone. The sounds provided an important cue to the location of the visual stimuli and were included to boost learning as has been found in studies of multisensory facilitation (Shams & Seitz, 2008).

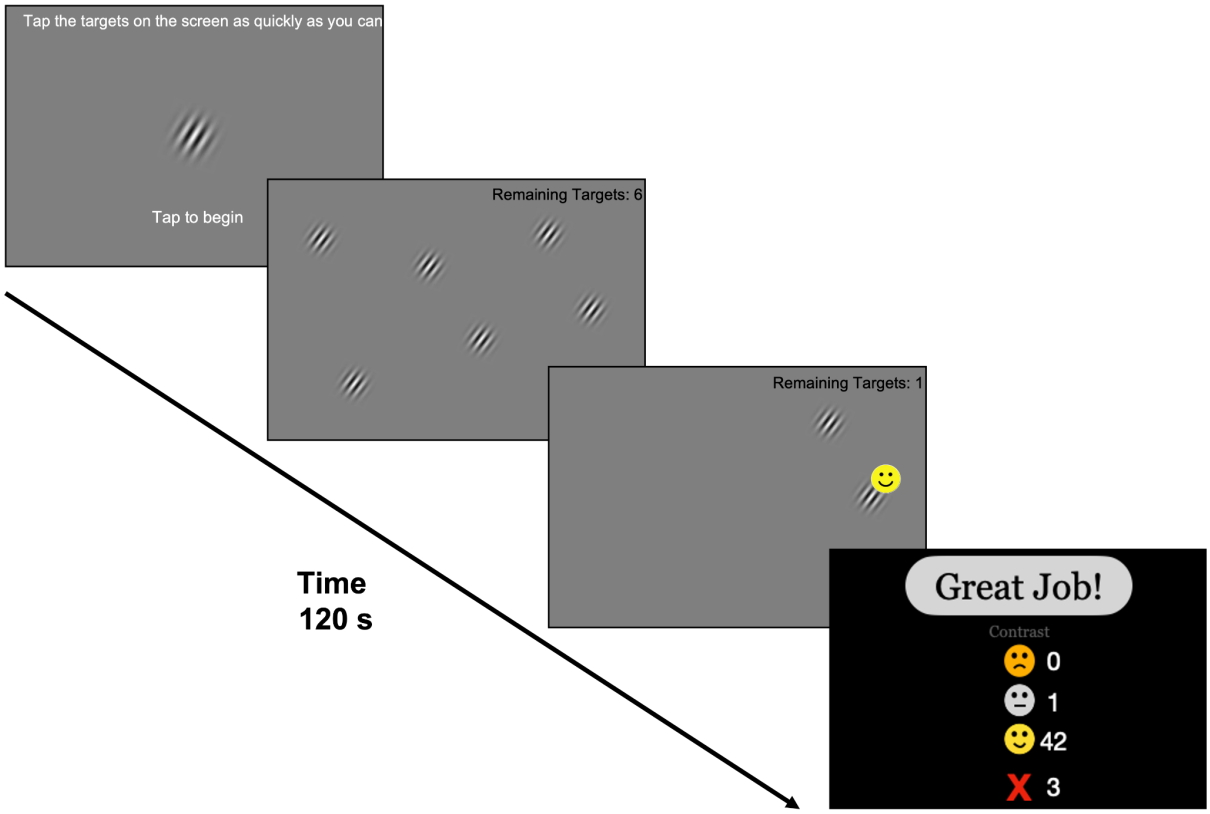

Figure 3: Task design used in experiment 2. Participants were initially presented with a sample target image and were also shown the remaining number of targets that needed to be identified on the screen. At the end of each block, they were provided a summary of their performance. Training blocks varied in terms of the orientation and spatial frequency of the targets.

Like most CSF tests, the CSF values from this study were computed at a small number of discrete spatial frequencies. Spline interpolation was again used to create smooth ground truth CSFs, this time of individual participants' CSF curves. A value of 0.5 was again used as the psychometric spread to produce simulated contrast response values from the generative models. Newly generated raw data in this fashion were used to train a multidimensional GP probabilistic classifier, as in experiment 1. Estimated CSF values were again compared to ground truth CSF values at octave frequencies with RMSE. Instead of multiple repeats of the same model, however, in this case performance was averaged across all 19 individuals in the data set.

Evaluation

In all cases the predictive posterior mean over spatial frequency and contrast is taken as the output of the MLCRF estimator. Because of the nature of Gaussian processes, this is equivalent to a maximum a posteriori (MAP) estimate. It should be kept in mind, however, that this method is fully Bayesian, meaning that every point estimate in the displayed predictive posteriors is actually just a single point in a complete distribution.

MLCSF estimates were extracted from MLCRFs as the 0.5 probability counter as a function of spatial frequency. This definition has been shown empirically in both simulations and in human experiments to represent an unbiased estimator of the underlying threshold (Song et al., 2015, 2017). Deviation from ground truth was quantified by RMSE.

Results

Experiment 1

The first experiment involved estimating several canonical CSF phenotypes. **Figure 4** depicts for each canonical phenotype the results of one of these estimation runs with 100 randomly sampled combinations of stimulus spatial frequency and contrast, beginning with an uninformative prior belief. The 0.5 probability contours of the MLCRF predictive posterior mean functions visually correspond to the ground truth CSF functions to some degree. The capability of this estimator to infer smooth functions indicative of standard disease variants is apparent.

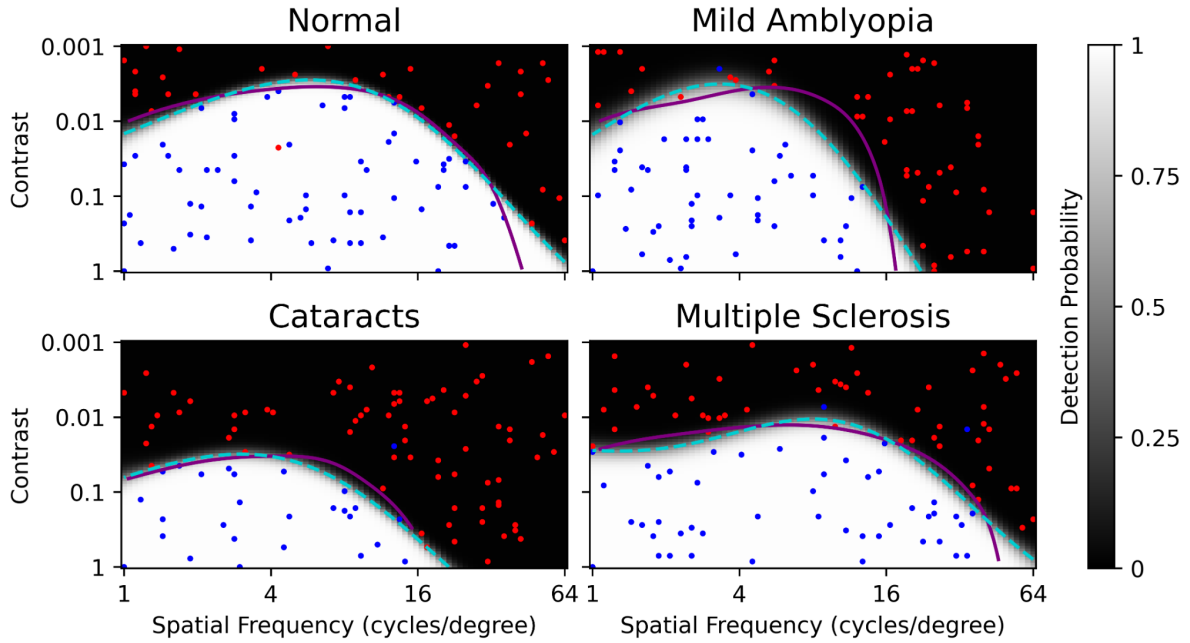

Figure 4: Four canonical CSF phenotypes were converted to generative models in order to create simulated binary detection data. Ground truth CSF curves are shown in purple. Successful detections of a visual pattern in the stimulus are indicated by blue plot symbols, while failures are indicated in red. One hundred samples are randomly distributed in spatial frequency and contrast. Grayscales indicate the learned predictive posterior means from the MLCRF estimator. Dashed turquoise lines represent 0.5 detection probability thresholds determined from the predictive posteriors, indicating that the CSF functions were learned reasonably well. Deviations from ground truth clearly arise from a low density of sampling in the vicinity, which is a limitation of random sampling and fixed-location sampling (i.e., grid search or method of constant stimuli, not shown). RMSE values for these MLCSF curves are as follows: Normal, 0.128; Mild Amblyopia, 0.303; Cataracts, 0.114; Multiple Sclerosis, 0.108.

The limitations of random sampling are also on view, as there are multiple regions where, by chance, no samples were taken near threshold. Samples near threshold are intrinsically more informative about the location of that threshold than samples farther away. One of the substantive contributions of modern machine learning has been the development of numerical algorithms for implementing optimization over a wide variety of functional definitions. Even though MLCRF estimates a nonparametric function, for example, the most informative data to collect in order to refine that estimate can readily be computed within this framework, whereas that was not feasible with classical methods.

Active learning is the machine learning principle that seeks to select new samples (i.e., stimulus parameter combinations) such that their observation provides the most information about the latent model. This procedure is analogous to optimal adaptive Bayesian estimation for parametric models.

An example with 100 samples of active learning and uninformative prior for the canonical multiple sclerosis phenotype is shown in **Figure 5**. The spatial frequency and contrast of the first eight samples were deterministically determined according to a Halton set (Halton, 1964) and the resulting simulated responses combined with an uninformative prior in order to prepare for active learning.

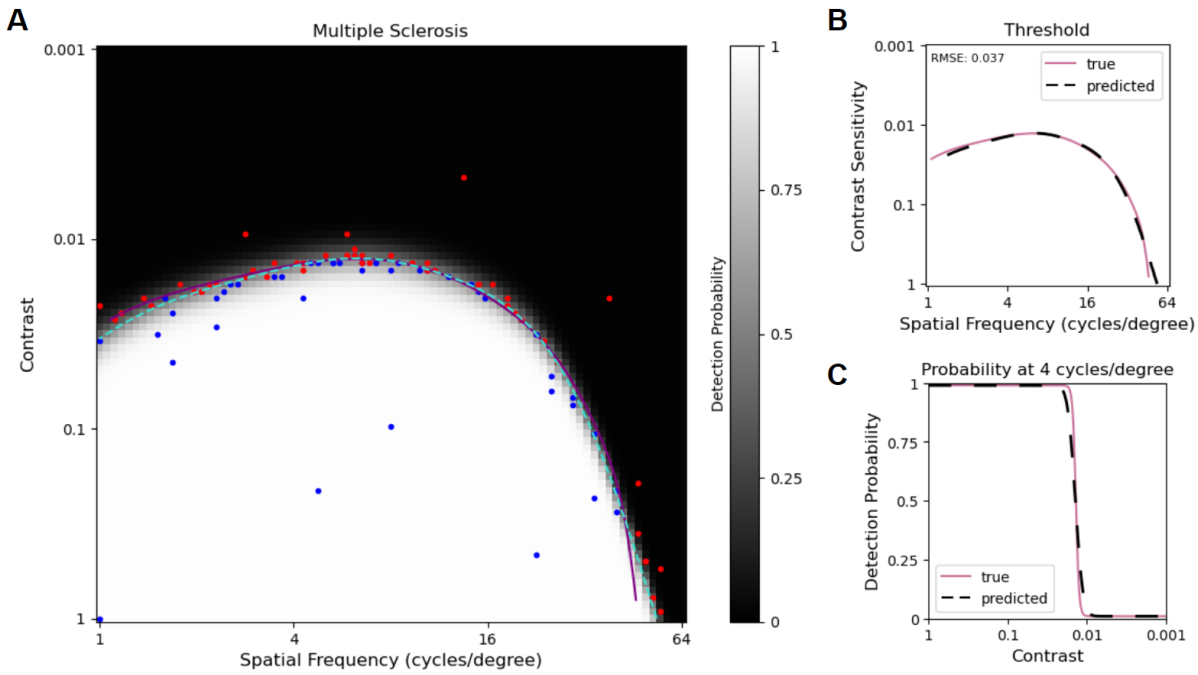

Figure 5: The MLCRF estimated with 100 samples of active learning for one canonical phenotype. **A**. Simulated behavioral responses (blue=success, red=failure), ground truth multiple sclerosis CSF (purple), learned MLCSF (dashed turquoise) and predictive posterior mean (grayscale) plotted according to the conventions of Figure 4. **B**. MLCSF curve (dashed) replotted along with the ground truth CSF curve (purple). Ordinate units indicate stimulus contrast. **C**. MLCRF psychometric curve at 4 cycles per degree plotted along with the ground truth psychometric function curve.

Active learning involves optimally selecting the next stimulus to deliver based on how informative that new data point would be for updating the model. After training model M_i on the initial i sample points during active learning, the most informative next sample point is determined, that stimulus is delivered, the simulated participant response is observed, and a new model M_{i+1} is learned with i+1 points and model M_i as the prior. This procedure continues until data from 100 samples are accumulated.

As can be seen in **Figure 5A**, most of the samples in the active learning condition are distributed near the eventual threshold curve, which would be expected for an acquisition function seeking to maximize informativeness. The ability to rapidly converge on the most informative samples is a key contributing factor of real-time optimization enabling more efficient testing procedures. The

predictive posterior MLCRF mean function can be seen in grayscale, indicating that at every combination of spatial frequency and contrast within the domain of support, the MLCRF estimator returns a probability of stimulus detection. This ability to deliver item-level predictions makes the MLCRF a fully predictive model—in other words, a complete psychometric field.

Direct comparisons between estimates and ground truths for this example are visualized in the side panels. The MLCSF estimate and ground truth threshold as a function of spatial frequency align visually in **Figure 5B**. Accurate estimation of the CSF is the goal of this research because of its established clinical utility. Just as in the case of models employing a truncated parabola, the MLCSF is continuous. Its lack of rigid shape constraints, however, is a reflection of its flexibility, indicated here by subtle curvature details.

The MLCSF represents a contour within the MLCRF, made in the current experiments by slicing the predictive posterior mean function parallel to the spatial frequency / contrast plane at a value of 0.5. One can also slice the MLCRF perpendicular to this plane at a single spatial frequency in order to reveal its psychometric shape. The resulting sigmoid for this example at 4 cycles per degree is depicted in **Figure 5C**, showing a learned close correspondence to the generative ground truth model. Estimating a single psychometric curve accurately with conventional methods is generally considered to require hundreds of samples, yet only 100 samples total have been accumulated for the entire MLCRF model here. This rapid convergence to a fully predictive model reveals the efficiency of the estimator.

Recall that the MLCRF is a probabilistic classifier, so its predictive posterior reflects a subdivision of feature space into separate regions likely to give rise to one observation or the other (i.e., success or failure). Greater uncertainty of the class boundary leads to wider transitions between regions, which translates into a larger psychometric spread in plots such as those of Figure 5C. Conversely, greater certainty leads to narrower transitions. In all cases this transition would be expected to be wider than the ground truth spread because of the additional uncertainty added by the likelihood in order to achieve item-level prediction. This behavior is expected and desired. If an application calls for unbiased estimates of psychometric spread, the use of the posterior mean function of a probabilistic classifier observed through a sigmoidal link function has been shown to be effective (Song et al., 2017, 2018). Because of this earlier work and the yet-to-be-established utility of spread for clinical applications of contrast detection, no additional analysis of spread estimation is performed for the current study.

The previous analysis provided proof-of-concept evidence in support of the accuracy and efficiency of the MLCRF estimator and, by extension, the MLCSF estimator. By updating a MLCSF model each

time a new data point is collected and comparing to CSF ground truth, the average accuracy of the estimator can be evaluated as a function of the amount of data collected (i.e., the sample count or number of behavioral trials).

In repeat simulations for each canonical phenotype, two independent estimator configurations were evaluated in a 2×2 arrangement: sampling strategy and selection of prior. The two choices for sampling were random, as depicted in Figure 4, and active, as depicted in Figure 5. Random sampling represents a way to quantify the performance of a machine learning estimator independent of its internal optimization algorithm, which can perform poorly for complex functions by failing to converge or becoming trapped at local optima. Assuming the optimizer is functioning well, active sampling is expected to produce systematically better performance than random sampling.

The two choices for prior were uninformative and canonical phenotype mixture. An informative prior incorporates knowledge into the model that originates from beyond the immediately collected data. A prior represents a bias, and a bias in the right direction is expected to improve the efficiency of the estimator, i.e., allowing it to systematically converge to a final estimate faster than with the use of an unbiased, uninformative prior. A poorly chosen bias can, on the other hand, push learning in the wrong direction and decrease efficiency.

The four combinations of these estimator configurations were evaluated for each canonical phenotype 10 times and the resulting average MLCSF accuracy compared against ground truth CSFs plotted as a function of sample count in **Figure 6**. For all canonical phenotypes and estimator configurations, the MLCSF converged toward an accurate representation of ground truth CSF values, as signified by the trend toward low root-mean-square errors. For a visual indication of estimator quality at a given RMSE value, refer to Figures 4 and 5.

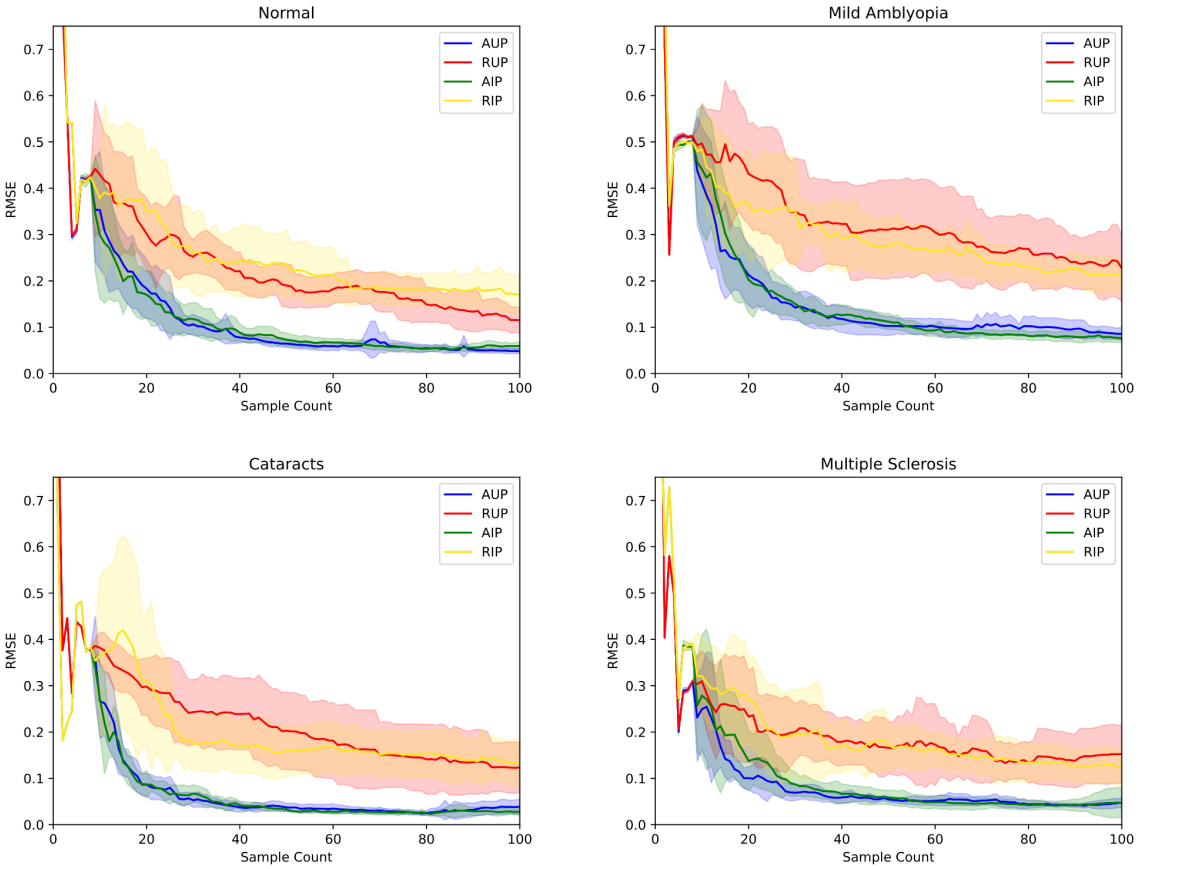

Figure 6: Mean +/– standard deviation root-mean-square error (RMSE) values of all four canonical phenotypes averaged from 10 repeat experiments using up to 100 samples of random with uninformative priors (RUP), random with informative priors (RIP), active with uninformative priors (AUP) and active with informative priors (AIP).

As expected, active learning, with sampling intentionally targeting threshold values, accelerated convergence toward accurate CSF representations. Average active sampling performance achieved equivalence to 100-sample average random sampling performance in less than 20 samples in all cases. Given that the first 8 samples were deterministic for every run, approximately 10 active samples turned out to be about as effective as approximately 100 random samples. This expected behavior confirms that the optimization component of MLCRF is functioning as designed.

A more complex scenario can be seen, however, when comparing the relative contributions of different priors. A larger impact of prior on random sampling is apparent compared to active sampling. Even for random sampling, however, it does not systematically improve convergence. For active sampling the two priors have indistinguishable effects on convergence. Because active learning represents a sequential Bayesian update procedure, the prior seen for later models is the posterior

from previous models, and jumpstarting that sequence with an informative prior seems to have little systematic effect under these testing conditions.

It is possible that any effect of an informative prior at the beginning is muted by the Halton sampling across the entire input domain. Halton sampling was included based on previous experience to prevent runaway degenerate sampling for the active learning, uninformative prior condition. Perhaps an informative prior can supplant this kind of heuristic, though there is practical value for behavioral testing scenarios in having participants start with a few trials that are very likely to be well above and well below their performance thresholds.

Experiment 2

The results of experiment 1 demonstrate that the general form of machine learning psychometric field estimation can successfully be applied to estimating visual contrast response functions for idealized phenotypes. In order to further illustrate the ability of this novel estimator to represent a variety of actual CSFs estimated from individuals in the course of human experiments, the types of manipulations from experiment 1 were replicated for different individuals in a second experiment.

Similarly to experiment 1, each individual CSF depicted in Figure 2 was used to create a generative ground truth model CSF. An example from a schizophrenic individual plotted according to the conventions of Figure 5 can be seen in **Figure 7**. The same observations relative to Figure 5 can be made in this case. The flexibility of the machine learning estimator is apparent in this example. By happenstance this example includes a lucky guess near (1.5, 0.01), which did not ultimately bias the final threshold estimate, although additional samples in that vicinity were required for the model to confirm that it was an outlier. This detail reveals the robustness of the MLCRF estimator as well as its efficiency.

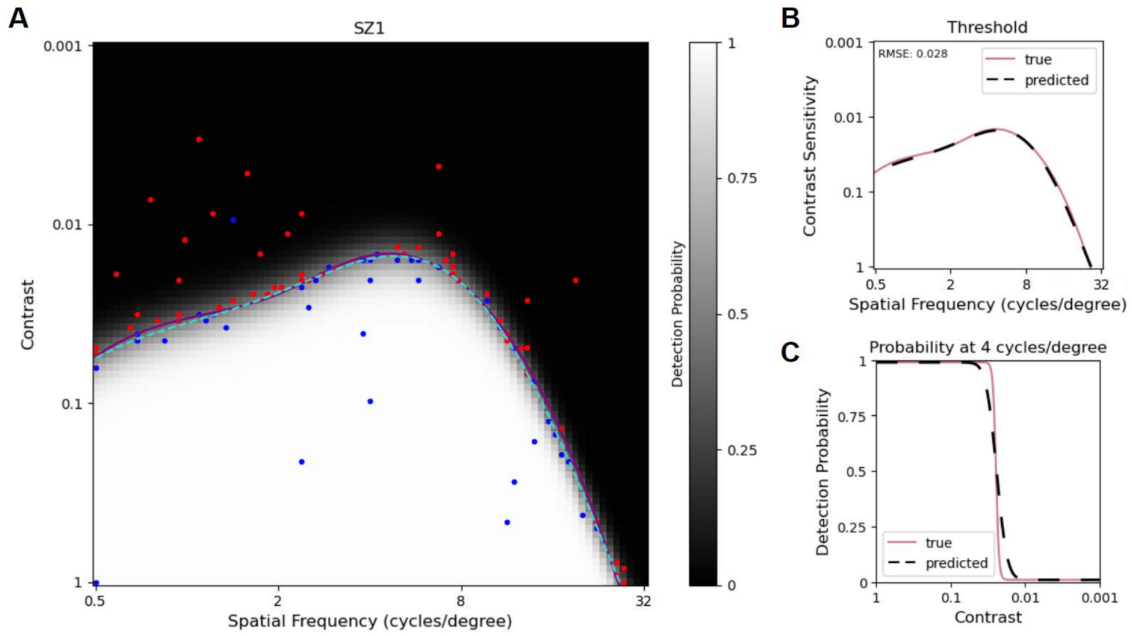

Figure 7: The MLCRF estimated with 100 samples of active learning using plotting conventions of Figure 5. **A**. Simulated behavioral responses, ground truth of a single schizophrenic individual CSF, and MLCSF estimate superimposed on the predictive posterior MLCRF mean. This example demonstrates recovery from a false positive response near (1.5, 0.01). **B**. MLCSF curve replotted along with the ground truth CSF curve. Ordinate units indicate stimulus contrast. **C**. MLCRF psychometric curve at 4 cycles per degree plotted along with the ground truth psychometric function curve.

The individual ground truths compared against MLCSF estimates from 100 random samples with uninformative priors can be seen in **Figure 8**. Visually, the models show close alignment with the ground truths, also reflected in the low accompanying RMSE values. Notably, several examples deviate substantially from standard parametric forms typically used for CSF estimators, including the truncated parabola. This result further implies that the MLCSF estimator is flexible enough to accurately reflect a wide variety of individual CSF shapes.

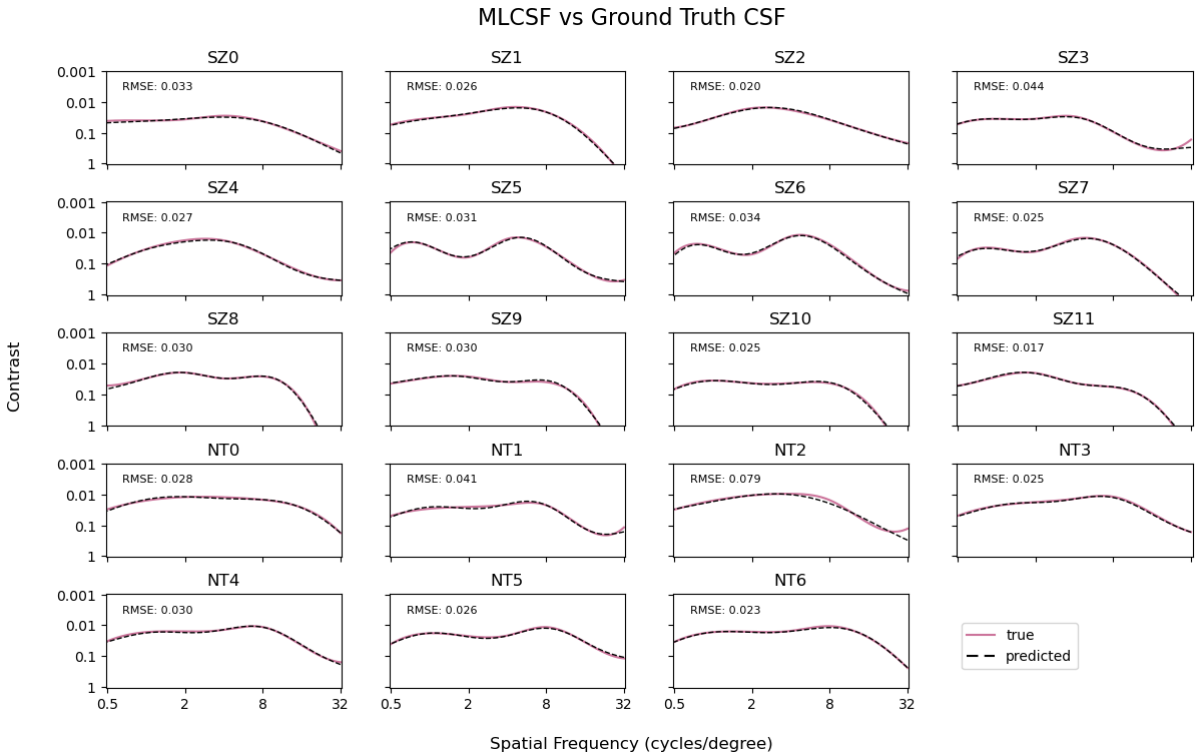

Figure 8: Comparison of ground truth CSF curves with MLCSF curves estimated using 100 active learning samples with uninformative prior. Twelve schizophrenic individuals and seven neurotypical individuals are represented.

As in experiment 1, the same two independent estimator configurations were evaluated for individuals in a 2×2 arrangement as a function of sample count and the results shown in **Figure 9**. This time, however, each CSF was modeled only once, and variation in performance was evaluated across the entire cohort. Once again, active learning showed an 80% to 90% efficiency gain in converging to accurate CSF estimates over random sampling. In this case random sampling with a prior performed better on average than random sampling without a prior, though with a higher variance. Given that this was not a consistent finding in experiment 1, however, concluding that an informative prior systematically benefits the random sampling condition is premature without further investigation.

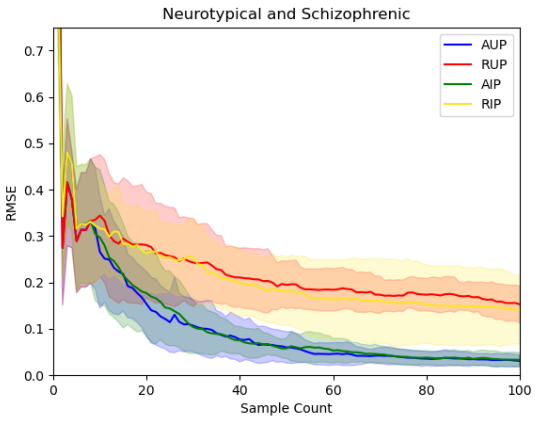

Figure 9: Mean +/– standard deviation RMSE values of all neurotypical and schizophrenic individuals for the same 4 experimental conditions as in Figure 6, with the same plotting conventions.

One notable distinction between experiments 1 and 2 is an apparently slower average convergence and more variation in the latter case. Variability in experiment 1 is akin to test-retest reliability of the MLCSF estimator, whereas variability in experiment 2 also includes the variation from different CSF shapes. The latter is therefore likely to be a more appropriate representation of the variation expected in estimator performance for an arbitrary CSF shape. It is also worth observing that variation from active sampling is systematically lower than from random sampling in all conditions, as expected.

Discussion

Psychometric estimator design for research or clinical applications faces common considerations. Nonparametric methods such as adaptive staircases are individually efficient at estimating single thresholds by themselves. When strung together they do produce the most systematically accurate conventional estimators of threshold curves such as the CSF (Watson & Ahumada, 2005), but with diminishing returns for efficiency at high dimensions and resolutions. Figure 2 shows examples of the kinds of curves that result. A major disadvantage of this approach is the need to interpolate threshold values between discretized, pre-selected abscissa values on the curve. For phenotypes with steep dropoffs or other local nuances, this uniform sampling approach decreases estimation accuracy. Furthermore, this discretization limits the ability to exploit additional information to improve estimation efficiency without relying on parametric assumptions about the functional form of the CSF.

Alternatively, parametric methods can incorporate outside information, such as constraints on CSF shape, and provide continuous threshold estimates. CSF estimators designed along these principles can achieve high efficiency at the expense of some accuracy (Hou et al., 2015; Lesmes et al., 2010; Watson & Ahumada, 2005; Zhao et al., 2021). Neither conventional parametric nor nonparametric models of CSF learn item-level prediction, however, because they are formulated as explanatory models.

This paper describes the development of a probabilistic machine learning classifier designed to deliver fully predictive, probabilistic models of item-level contrast detection responses from human participants for patterned visual stimuli as a function of spatial frequency and visual contrast. From these machine learning contrast response functions (MLCRFs), machine learning contrast sensitivity functions (MLCSFs) can be estimated as a single performance threshold contour, here taken to be the

0.5 detection probability contour. Similar methods have been successful at estimating detection thresholds for audibility (Cox & de Vries, 2021; Schlittenlacher et al., 2018; Song et al., 2015) and visual fields (Chesley & Barbour, 2020).

As expected for a nonparametric estimator, accuracy for the MLCSF was generally high for a variety of shapes. Efficiency was poor for the tested estimator configuration with random sampling, however, which was not unexpected given previous findings (Song et al., 2018). Active learning led to more informative stimulus delivery, which yielded faster convergence to accurate CSF estimates. Including an informative prior did not systematically improve estimation efficiency, which was an unexpected result. Perhaps the other elements of the estimator configuration, such as the initial deterministic sampling or the active learning of hyperparameters, muted the impact of an informative prior. In any case, this characteristic of the general MLCRF estimator provides for this level of ready configuration flexibility should it be useful in other contexts.

This simulation study successfully demonstrated the utility of the MLCRF estimator design for CSF estimation. It is both flexible and efficient, able to estimate arbitrary CSF shapes to reasonable accuracies in a few dozen samples. This performance compares favorably with perhaps the most advanced current CSF estimator capable of incorporating knowledge across individuals and testing sessions, as well as testing variables, with a hierarchical design (Zhao et al., 2021). Early evidence raises the possibility that the MLCRF estimator represents a more favorable tradeoff between accuracy and efficiency than conventional methods, essentially combining the increased accuracy of nonparametric methods with the increased efficiency of parametric methods.

Additionally, the nature of the MLCSF estimator allows for explicit tradeoffs between accuracy and efficiency. While in the current study the model hyperparameters were learned as data were collected, these could be fixed to imbue the model with particular assumptions. To give one example, if very strong prior beliefs about the curvature of the CSF curve existed, the length constant in the squared exponential kernel could be fixed at a particular value. More appropriately for a Bayesian estimator, a narrow prior belief centered around that value could be applied to that hyperparameter. In such cases, the average convergence to the final estimate would be expected to be faster at the expense of average accuracy. This type of direct manipulation of the constraints on function shape is not available for conventional parametric CSF estimators.

More compelling reasons also exist to consider a complex machine learning model for CSF estimation over conventional procedures. Extreme efficiency gains are possible by formulating specific hypotheses to be ruled in or out in real time. Bayesian active model selection offers this hyper-efficient means of identifying potential disease states at arbitrary levels of confidence (Gardner et al.,

2015) or fluctuations in visual processing over time (Larsen et al., 2020, 2021). In this framework, each new data point is used to accumulate evidence for or against discrete hypotheses defined either by phenotypic examples or an individual's previous test results. This procedure retains the flexibility of nonparametric estimation while requiring quite a small number of samples to reach a conclusion. Its potential utility might lie in beginning each eye's CSF test as a screening procedure to first try to rule out pathology efficiently. If this cannot be done with high confidence using a small number of samples, the acquisition function could then switch to optimally determining the actual shape of the CSF, as in the current study. In this framework at each phase of testing the question of greatest interest is addressed optimally, which will lead to considerable time savings for a given diagnostic test sensitivity and specificity. This efficiency advantage is likely to be highest with highly asymmetric pretest probabilities, such as for screening tests in populations with no known visual disorders.

The MLCRF estimator is designed for ready extendibility, which enables combinations of different types of data observed through various likelihoods. Thus, researchers can readily incorporate diverse, but related, sources of information that collectively integrate into more confident and comprehensive models. For example, by linking the two ears together in real time, hearing thresholds from both ears can be obtained simultaneously in the same average amount of time normally needed to obtain thresholds from one ear (Barbour et al., 2018; Heisey et al., 2018). The same improvements would be expected for contrast tasks simultaneously performed by an individual's two eyes.

The MLCRF learns a psychometric spread at every spatial frequency rather than assuming it. Unfortunately, examples of such multidimensional models do not exist to compare against in terms of efficiency. The utility of this capability is unclear at present because spread estimates are not generally employed, but there could be subtle variations in spreads for different phenotypes that might make sensitive testing procedures more likely to resolve certain pathologies. Thus, this fully predictive model could open up avenues in research for the exploration of deeper behavioral response relationships not yet fully considered in psychophysics and vision science.

Finally, while the MLCRF as coded can make use of small or large amounts of data to improve estimation properties, the framework it is based upon could easily accommodate different machine learning algorithms as data available to constrain models increase in number. It is not unreasonable to consider that in the future, once large numbers of individuals have completed MLCRF tests, the mass of data could be used to train even more flexible machine learning models that can exploit many small correlations in widely divergent data streams to accommodate more refined individualized phenotypes. The models themselves could grow in complexity, incorporating more details about visual and brain health than simply a person's ability to resolve visual images at different spatial

frequencies and contrasts. This scalability in predictive model form is not generally available to other estimator frameworks.

Conclusions

A nonparametric Bayesian estimator learning from simulated contrast detection behavioral responses can provide accurate estimates for a variety of contrast sensitivity function shapes in a few dozen samples under ideal conditions. This estimator makes use of recent developments in probabilistic machine learning classification that provide it with potential advantages over classical estimators at the cost of increased model complexity. Notably, it can accommodate a wide set of constraints on functional form, it is a fully predictive model, and it can be extended to incorporate related data from a variety of sources. Future work will extend the estimator algorithm in ways described above and evaluate its performance in real-time human data acquisition.

Acknowledgments

Supported by R21EY033553 and R01EY023582. DLB has equity ownership in Bonauria and a patent on technology used in this study.

References

Barbour, D. L., DiLorenzo, J., Sukesan, K. A., Song, X. D., Chen, J. Y., Degen, E. A., Heisey, K. L., & Garnett, R. (2018). Conjoint psychometric field estimation for bilateral audiometry. *Behav Res Meth*. https://doi.org/10.3758/s13428-018-1062-3

Basha, S. M., & Rajput, D. S. (2019). Survey on evaluating the performance of machine learning algorithms: Past contributions and future roadmap. In *Deep Learning and Parallel Computing Environment for Bioengineering Systems* (pp. 153–164). Elsevier.

Calderone, D. J., Martinez, A., Zemon, V., Hoptman, M. J., Hu, G., Watkins, J. E., Javitt, D. C., & Butler, P. D. (2013). Comparison of psychophysical, electrophysiological, and fMRI assessment of visual contrast responses in patients with schizophrenia. *NeuroImage*, *67*, 153–162. https://doi.org/10.1016/j.neuroimage.2012.11.019

Chesley, B., & Barbour, D. L. (2020). Visual field estimation by probabilistic classification. *IEEE Journal of Biomedical and Health Informatics*, *24*(12), 3499–3506. https://doi.org/10.1109/JBHI.2020.2999567

Chung, S. T. L., & Legge, G. E. (2016). Comparing the shape of contrast sensitivity functions for normal and low vision. *Investigative Ophthalmology & Visual Science*, *57*(1), 198–207. https://doi.org/10.1167/iovs.15-18084

Cox, M., & de Vries, B. (2021). Bayesian pure-tone audiometry through active learning under informed priors. *Frontiers in Digital Health*, *3*, 723348. https://doi.org/10.3389/fdgth.2021.723348

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology: CB*, *24*(4), R146-147. https://doi.org/10.1016/j.cub.2014.01.004

DiMattina, C. (2015). Fast adaptive estimation of multidimensional psychometric functions. *J Vis*, *15*(9), 5–5.

Gardner, J. M., Malkomes, G., Garnett, R., Weinberger, K. Q., Barbour, D., & Cunningham, J. P. (2015). Bayesian active model selection with an application to automated audiometry. *Adv Neural Inf Process Syst*, 2377–2385.

Gardner, J. M., Pleiss, G., Weinberger, K. Q., Bindel, D., & Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in Neural Information Processing Systems*, *31*.

Ginsburg, A. P. (2003). Contrast sensitivity and functional vision. *International Ophthalmology Clinics*, *43*(2), 5–15. https://doi.org/10.1097/00004397-200343020-00004

GPyTorch 1.9.1. (n.d.). Retrieved February 26, 2023, from https://docs.gpytorch.ai/en/stable/

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons, Inc.

Gu, H., Kim, W., Hou, F., Lesmes, L. A., Pitt, M. A., Lu, Z.-L., & Myung, J. I. (2016). A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function. *Journal of Vision*, *16*(6), 15. https://doi.org/10.1167/16.6.15

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Commun ACM*, *7*(12EDITED), 701–702.

Heisey, K. L., Buchbinder, J. M., & Barbour, D. L. (2018). Concurrent Bilateral Audiometric Inference. *Acta Acustica United with Acustica*, *104*(5), 762–765. https://doi.org/info:doi/10.3813/AAA.919218

Hensman, J., Matthews, A., & Ghahramani, Z. (2015). Scalable variational Gaussian process classification. *Artificial Intelligence and Statistics*, 351–360.

Hou, F., Lesmes, L., Bex, P., Dorr, M., & Lu, Z.-L. (2015). Using 10AFC to further improve the efficiency of the quick CSF method. *Journal of Vision*, *15*(9), 2. https://doi.org/10.1167/15.9.2

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *ArXiv Preprint ArXiv:1112.5745*.

Kalloniatis, M., & Luu, C. (1995). Visual Acuity. In H. Kolb, E. Fernandez, & R. Nelson (Eds.), *Webvision: The Organization of the Retina and Visual System*. University of Utah Health Sciences Center. http://www.ncbi.nlm.nih.gov/books/NBK11509/

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A Practical Introduction*. Elsevier.

King-Smith, P. E. (1984). Efficient threshold estimates from yes-no procedures using few (about 10) trials. *American Journal of Optometry and Physiological Optics*, *61*, 119.

King-Smith, P. E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, *37*(12), 1595–1604. https://doi.org/10.1016/s0042-6989(96)00310-0

Kokol, P., Kokol, M., & Zagoranski, S. (2021). Machine learning on small size samples: A synthetic knowledge synthesis. *ArXiv Preprint ArXiv:2103.01002*.

Kolb, H., Fernandez, E., & Nelson, R. (Eds.). (1995). *Webvision: The Organization of the Retina and Visual System*. University of Utah Health Sciences Center. http://www.ncbi.nlm.nih.gov/books/NBK11530/

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Res*, *39*(16), 2729–2737.

Larsen, T. J., Malkomes, G., & Barbour, D. L. (2020). Accelerating Psychometric Screening Tests With Bayesian Active Differential Selection. *ArXiv Preprint ArXiv:2002.01547*.

Larsen, T. J., Malkomes, G., & Barbour, D. L. (2021). Accelerating Psychometric Screening Tests with Prior Information. In A. Shaban-Nejad, M. Michalowski, & D. L. Buckeridge (Eds.), *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (pp. 305–311). Springer International Publishing. https://doi.org/10.1007/978-3-030-53352-6_29

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Percept Psychophys*, *63*(8), 1279–1292.

Lesmes, L. A., Jeon, S.-T., Lu, Z.-L., & Dosher, B. A. (2006). Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. *Vision Research*, *46*(19), 3160–3176. https://doi.org/10.1016/j.visres.2006.04.022

Lesmes, L. A., Lu, Z.-L., Baek, J., & Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *Journal of Vision*, *10*(3), 17.1-21. https://doi.org/10.1167/10.3.17

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477.

PyTorch 1.13. (n.d.). Retrieved February 26, 2023, from https://pytorch.org/docs/stable/index.html

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Rohaly, A. M., & Owsley, C. (1993). Modeling the contrast-sensitivity functions of older adults. *JOSA A*, *10*(7), 1591–1599. https://doi.org/10.1364/JOSAA.10.001591

Schlittenlacher, J., Turner, R. E., & Moore, B. C. J. (2018). Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, *144*(1), 421. https://doi.org/10.1121/1.5047436

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Song, X. D., Garnett, R., & Barbour, D. L. (2017). Psychometric function estimation by probabilistic classification. *The Journal of the Acoustical Society of America*, *141*(4), 2513. https://doi.org/10.1121/1.4979594

Song, X. D., Sukesan, K. A., & Barbour, D. L. (2018). Bayesian active probabilistic classification for psychometric field estimation. *Attention, Perception & Psychophysics*, *80*(3), 798–812. https://doi.org/10.3758/s13414-017-1460-0

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., & Barbour, D. L. (2015). Fast, continuous audiogram estimation using machine learning. *Ear and Hearing*, *36*(6), e326-335. https://doi.org/10.1097/AUD.0000000000000186

Tahir, H. J., Parry, N. R. A., Pallikaris, A., & Murray, I. J. (2009). Higher-order aberrations produce orientation-specific notches in the defocused contrast sensitivity function. *Journal of Vision*, *9*(7), 11. https://doi.org/10.1167/9.7.11

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. *Artificial Intelligence and Statistics*, 567–574.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Res*, *35*(17), 2503–2522.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Percept Psychophys*, *61*(1), 87–106.

Wang, X., Wang, H., Huang, J., Zhou, Y., & Tzvetanov, T. (2016). Bayesian Inference of Two-Dimensional Contrast Sensitivity Function from Data Obtained with Classical One-Dimensional Algorithms Is Efficient. *Frontiers in Neuroscience*, *10*, 616. https://doi.org/10.3389/fnins.2016.00616

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, *5*(9), 717–740. https://doi.org/10.1167/5.9.6

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Woods, R. L., Bradley, A., & Atchison, D. A. (1996). Consequences of monocular diplopia for the contrast sensitivity function. *Vision Research*, *36*(22), 3587–3596. https://doi.org/10.1016/0042-6989(96)00091-0

Yaghoubi, K. C., Jayakumar, S., Ahmed, A. O., Butler, P. D., Silverstein, S., Thompson, J. L., & Seitz, A. R. (2022). Characterization of training profiles between individuals with schizophrenia and healthy individuals on Contrast Detection and Contour Integration tasks. *Journal of Vision*, *22*(14), 3728–3728.

Yarmohammadi, A., Zangwill, L. M., Diniz-Filho, A., Suh, M. H., Yousefi, S., Saunders, L. J., Belghith, A., Manalastas, P. I., Medeiros, F. A., & Weinreb, R. N. (2016). Relationship between optical coherence tomography angiography vessel density and severity of visual field loss in glaucoma. *Ophthalmology*, *123*(12), 2498–2508.

Zhao, Y., Lesmes, L. A., Hou, F., & Lu, Z.-L. (2021). Hierarchical Bayesian modeling of contrast sensitivity functions in a within-subject design. *Journal of Vision*, *21*(12), 9. https://doi.org/10.1167/jov.21.12.9

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. (2017). Machine Learning on Big Data: Opportunities and Challenges. *Neurocomputing*, *237*, 350–361.