A Semiparametric Model for Bayesian Reader Identification

Ahmed Abdelwahab[1] and **Reinhold Kliegl**[2] and **Niels Landwehr**[1]
[1] Department of Computer Science, Universität Potsdam
August-Bebel-Straße 89, 14482 Potsdam, Germany
{ahmed.abdelwahab, niels.landwehr}@uni-potsdam.de
[2] Department of Psychology, Universität Potsdam
Karl-Liebknecht-Straße 24/25, 14476 Potsdam OT/Golm
kliegl@uni-potsdam.de

Abstract

We study the problem of identifying individuals based on their characteristic gaze patterns during reading of arbitrary text. The motivation for this problem is an unobtrusive biometric setting in which a user is observed during access to a document, but no specific challenge protocol requiring the user's time and attention is carried out. Existing models of individual differences in gaze control during reading are either based on simple aggregate features of eye movements, or rely on parametric density models to describe, for instance, saccade amplitudes or word fixation durations. We develop flexible semiparametric models of eye movements during reading in which densities are inferred under a Gaussian process prior centered at a parametric distribution family that is expected to approximate the true distribution well. An empirical study on reading data from 251 individuals shows significant improvements over the state of the art.

1 Introduction

Eye-movement patterns during skilled reading consist of brief fixations of individual words in a text that are interleaved with quick eye movements called *saccades* that change the point of fixation to another word. Eye movements are driven both by low-level visual cues and high-level linguistic and cognitive processes related to text understanding; as a reflection of the interplay between vision, cognition, and motor control during reading they are frequently studied in cognitive psychology (Kliegl et al., 2006; Rayner, 1998). Computational models (Engbert et al., 2005; Reichle et al., 1998) as well

as models based on machine learning (Matties and Søgaard, 2013; Hara et al., 2012) have been developed to study how gaze patterns arise based on text content and structure, facilitating the understanding of human reading processes.

A central observation in these and earlier psychological studies (Huey, 1908; Dixon, 1951) is that eye movement patterns strongly differ between individuals. Holland et al. (2012) and Landwehr et al. (2014) have developed models of individual differences in eye movement patterns during reading, and studied these models in a biometric problem setting where an individual has to be identified based on observing her eye movement patterns while reading arbitrary text. Using eye movements during reading as a biometric feature has the advantage that it suffices to observe a user during a routine access to a device or document, without requiring the user to react to a specific challenge protocol. If the observed eye movement sequence is unlikely to be generated by an authorized individual, access can be terminated or an additional verification requested. This is in contrast to approaches where biometric identification is based on eye movements in response to an artificial visual stimulus, for example a moving (Kasprowski and Ober, 2004; Komogortsev et al., 2010; Rigas et al., 2012b; Zhang and Juhola, 2012) or fixed (Bednarik et al., 2005) dot on a computer screen, or a specific image stimulus (Rigas et al., 2012a).

The model studied by Holland & Komogortsev (2012) uses aggregate features (such as average fixation duration) of the observed eye movements. Landwehr et al. (2014) showed that readers can be identified more accurately with a model that captures aspects of individual-specific distributions over

eye movements, such as the distribution over fixation durations or saccade amplitudes for word refixations, regressions, or next-word movements. Some of these distributions need to be estimated from very few observations; a key challenge is thus to design models that are flexible enough to capture characteristic differences between readers yet robust to sparse data. Landwehr et al. (2014) used a fully parametric approach where all densities are assumed to be in the gamma family; gamma distributions were shown to approximate the true distribution of interest well for most cases (see Figure 1). This model is robust to sparse data, but might not be flexible enough to capture all differences between readers.

The model we study in this paper follows ideas developed by Landwehr et al. (2014), but employs more flexible semiparametric density models. Specifically, we place a Gaussian process prior over densities that concentrates probability mass on densities that are close to the gamma family. Given data, a posterior distribution over densities is derived. If data is sparse, the posterior will still be sharply peaked around distributions in the gamma family, reducing the effective capacity of the model and minimizing overfitting. However, given enough evidence in the data, the model will also deviate from the gamma-centered prior—depending on the kernel function chosen for the GP prior, any density function can in principle be represented. Integrating over the space of densities weighted by the posterior yields a marginal likelihood for novel observations from which predictions are inferred. We empirically study this model in the same setting as studied by Landwehr et al. (2014), but using an order of magnitude more individuals. Identification error is reduced by more than a factor of three compared to the state of the art.

The rest of the paper is organized as follows. After defining the problem setting in Section 2, Section 3 presents the semiparametric probabilistic model. Section 4 discusses inference, Section 5 presents an empirical study on reader identification.

2   Problem Setting

Assume R different readers, indexed by $r \in \{1, \ldots, R\}$, and let $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ denote a set of texts. Each $r \in \mathcal{R}$ generates a set of eye movement patterns $\mathcal{S}^{(r)} = \{\mathbf{S}_1^{(r)}, \ldots, \mathbf{S}_n^{(r)}\}$ on $\mathcal{X}$, by

$$\mathbf{S}_i^{(r)} \sim p(\mathbf{S}|\mathbf{X}_i, r, \mathbf{\Gamma})$$

where $p(\mathbf{S}|\mathbf{X}_i, r, \mathbf{\Gamma})$ is a reader-specific distribution over eye movement patterns given a text $\mathbf{X}_i$. Here, r is a variable indicating the reader generating the sequence, and $\mathbf{\Gamma}$ is a true but unknown model that defines all reader-specific distributions. We assume that $\mathbf{\Gamma}$ can be broken down into reader-specific models, $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k)$, such that the distribution

$$p(\mathbf{S}|\mathbf{X}_i, r, \mathbf{\Gamma}) = p(\mathbf{S}|\mathbf{X}_i, \boldsymbol{\gamma}_r) \tag{1}$$

is defined by the partial model $\boldsymbol{\gamma}_r$. We aggregate the observations of all readers on the training data into a variable $\mathcal{S}^{(1:R)} = (\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(R)})$.

We follow a Bayesian approach, defining a prior $p(\mathbf{\Gamma})$ over the joint model that factorizes into priors over reader-specific models, $p(\mathbf{\Gamma}) = \prod_{r=1}^{R} p(\boldsymbol{\gamma}_r)$. At test time, we observe novel eye movement patterns $\bar{\mathcal{S}} = \{\bar{\mathbf{S}}_1, \ldots, \bar{\mathbf{S}}_m\}$ on a novel set of texts $\bar{\mathcal{X}} = \{\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_m\}$ generated by an unknown reader $r \in \mathcal{R}$. We assume a uniform prior over readers, that is, each $r \in \mathcal{R}$ is equally likely to be observed at test time. The goal is to infer the most likely reader to have generated the novel eye movement patterns. In a Bayesian setting, this means inferring the most likely reader given the training observations $(\mathcal{X}, \mathcal{S}^{(1:R)})$ and test observation $(\bar{\mathcal{X}}, \bar{\mathcal{S}})$:

$$r_* = \arg\max_{r \in \mathcal{R}} p(r|\bar{\mathcal{X}}, \bar{\mathcal{S}}, \mathcal{X}, \mathcal{S}^{(1:R)}). \tag{2}$$

We can rewrite Equation 2 to

$$r_* = \arg\max_{r \in \mathcal{R}} p(\bar{\mathcal{S}}|r, \bar{\mathcal{X}}, \mathcal{X}, \mathcal{S}^{(1:R)}) \tag{3}$$

$$= \arg\max_{r \in \mathcal{R}} \int p(\bar{\mathcal{S}}|r, \bar{\mathcal{X}}, \mathbf{\Gamma}) p(\mathbf{\Gamma}|\mathcal{X}, \mathcal{S}^{(1:R)}) \mathrm{d}\mathbf{\Gamma}$$

$$= \arg\max_{r \in \mathcal{R}} \int p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}_r) p(\boldsymbol{\gamma}_r|\mathcal{X}, \mathcal{S}^{(r)}) \mathrm{d}\boldsymbol{\gamma}_r \tag{4}$$

where

$$p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}_r) = \prod_{i=1}^{m} p(\bar{\mathbf{S}}_i|\bar{\mathbf{X}}_i, \boldsymbol{\gamma}_r) \tag{5}$$

$$p(\boldsymbol{\gamma}_r|\mathcal{X}, \mathcal{S}^{(r)}) \propto p(\boldsymbol{\gamma}_r) \prod_{i=1}^{n} p(\mathbf{S}_i^{(r)}|\mathbf{X}_i, \boldsymbol{\gamma}_r). \tag{6}$$

586

In Equation 3 we exploit that readers are uniformly chosen at test time, and in Equation 4 we exploit the factorization $p(\mathbf{\Gamma}) = \prod_{r=1}^{R} p(\boldsymbol{\gamma}_r)$ of the prior, which together with Equation 1 entails a factorization $p(\mathbf{\Gamma}|\mathcal{X}, \mathcal{S}^{(1:R)}) = \prod_{r=1}^{R} p(\boldsymbol{\gamma}_r|\mathcal{X}, \mathcal{S}^{(r)})$ of the posterior. Note that Equation 4 states that at test time we predict the reader r for which the marginal likelihood (that is, after integrating out the reader-specific model $\boldsymbol{\gamma}_r$) of the test observations is highest. The next section discusses the reader-specific models $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma}_r)$ and prior distributions $p(\boldsymbol{\gamma}_r)$.

3 Probabilistic Model

The probabilistic model we employ follows the general structure proposed by Landwehr et al. (2014), but employs semiparametric density models and allows for fully Bayesian inference. To reduce notational clutter, let $\boldsymbol{\gamma} \in \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_R\}$ denote a particular reader-specific model, and let $\mathbf{X} \in \mathcal{X}$ denote a text. An eye movement pattern is a sequence $\mathbf{S} = ((s_1, d_1), \ldots, (s_T, d_T))$ of gaze fixations, consisting of a fixation position s_t (position in text that was fixated) and duration $d_t \in \mathbb{R}$ (length of fixation in milliseconds). In our experiments, individual sentences are presented in a single line on screen, thus we only model a horizontal gaze position $s_t \in \mathbb{R}$. We model $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma})$ as a dynamic process that successively generates fixation positions s_t and durations d_t in $\mathbf{S}$, reflecting how a reader generates a sequence of saccades in response to a text stimulus $\mathbf{X}$:

$$p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma}) = p(s_1, d_1|\mathbf{X}, \boldsymbol{\gamma}) \prod_{t=2}^{T} p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma}),$$

where $p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$ models the generation of the next fixation position and duration given the old fixation position s_{t-1}. In the psychological literature, four different *saccade types* are distinguished: a reader can refixate the current word (*refixation*), fixate the next word in the text (*next word movement*), move the fixation to a word after the next word, that is, skip one or more words (*forward skip*), or regress to fixate a word occurring earlier in the text (*regression*), see, e.g., Heister et al. (2012). We observe empirically that for each saccade type, there is a characteristic distribution over saccade amplitudes and fixation durations, and that both approximately follow gamma distributions—see Fig-
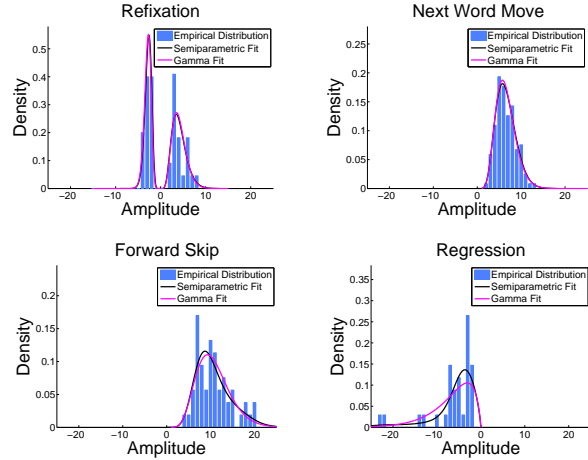

Figure 1: Empirical distributions of saccade amplitudes in training data for first individual, with fitted Gamma distributions and semiparametric distribution fits.

ure 1. We therefore model $p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$ using a mixture over distributions for the four different saccade types. At each time t, the model first draws a saccade type $u_t \in \{1, 2, 3, 4\}$, and then draws a saccade amplitude a_t and fixation duration d_t from type-specific distributions $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$ and $p(d|u_t, \boldsymbol{\gamma})$. More formally,

$$u_t \sim p(u|\boldsymbol{\pi}) \tag{7}$$
$$a_t \sim p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) \tag{8}$$
$$d_t \sim p(d|u_t, \boldsymbol{\delta}), \tag{9}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ is decomposed into components $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}$. Afterwards, the model updates the fixation position according to $s_t = s_{t-1} + a_t$, concluding the definition of $p(s_t, d_t|s_{t-1}, \mathbf{X}, \boldsymbol{\gamma})$. Figure 2 shows a slice in the dynamical model.

The distribution $p(u|\boldsymbol{\pi})$ over saccade types (Equation 7) is multinomial with parameter vector $\boldsymbol{\pi} \in \mathbb{R}^4$. The distributions over amplitudes and durations (Equations 8 and 9) are modeled semiparametrically as discussed in the following subsections.

3.1 Model of Saccade Amplitudes

We first discuss the amplitude model $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$ (Equation 8). We first define a distribution $p(a|u_t, \boldsymbol{\alpha})$ over amplitudes for saccade type u_t, and subsequently discuss conditioning on the text $\mathbf{X}$ and old fixation position s_{t-1},
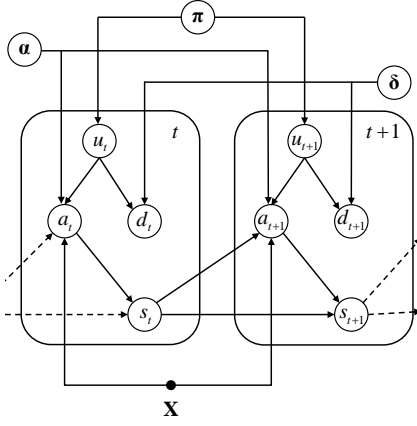
Figure 2: Plate notation of of a slice in the dynamic model.

leading to $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$. We define

$$p(a|u_t = 1, \boldsymbol{\alpha}) = \begin{cases} \mu\alpha_1(a) & : a > 0 \\ (1-\mu)\bar{\alpha}_1(-a) & : a \le 0 \end{cases} \quad (10)$$

where μ is a mixture weight and α_1, $\bar{\alpha}_1$ are densities defining the distribution over positive and negative amplitudes for the saccade type *refixation*, and

$$p(a|u_t = 2, \boldsymbol{\alpha}) = \alpha_2(a) \quad (11)$$
$$p(a|u_t = 3, \boldsymbol{\alpha}) = \alpha_3(a) \quad (12)$$
$$p(a|u_t = 4, \boldsymbol{\alpha}) = \alpha_4(-a) \quad (13)$$

where $\alpha_2(a)$, $\alpha_3(a)$, and $\alpha_4(a)$ are densities defining the distribution over amplitudes for the remaining saccade types. Finally, the distribution

$$p(s_1|\mathbf{X}, \boldsymbol{\alpha}) = \alpha_0(s_1) \quad (14)$$

over the initial fixation position is given by another density function α_0. The variables $\mu, \alpha_0, \alpha_1, \bar{\alpha}_1,$ $\alpha_2, \alpha_3,$ and α_4 are aggregated into model component $\boldsymbol{\alpha}$. For resolving the most likely reader at test time (Equation 4), densities in $\boldsymbol{\alpha}$ will be integrated out under a prior based on Gaussian processes (Section 3.3) using MCMC inference (Section 4).

Given the old fixation position s_{t-1}, the text $\mathbf{X}$, and the chosen saccade type u_t, the amplitude is constrained to fall within a specific interval. For instance, for a refixation the amplitude has to be chosen such that the novel fixation position lies within the beginning and the end of the currently fixated word; a regression implies an amplitude that is negative and makes the novel fixation position lie before the beginning of the currently fixated word.

These constraints imposed by the text structure define the conditional distribution $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$. More formally, $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$ is the distribution $p(a|u_t, \boldsymbol{\alpha})$ conditioned on $a \in [l, r]$, that is,

$$p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = p(a|a \in [l, r], u_t, \boldsymbol{\alpha}),$$

where l and r are the minimum and maximum amplitude consistent with the constraints. Recall that for a distribution over a continuous variable x given by density $\alpha(x)$, the distribution over x conditioned on $x \in [l, r]$ is given by the truncated density

$$\alpha(x|x \in [l, r]) = \begin{cases} \frac{\alpha(x)}{\int_l^r \alpha(\bar{x})\mathrm{d}\bar{x}} & : x \in [l, r] \\ 0 & : x \notin [l, r]. \end{cases} \quad (15)$$

We derive $p(a|u_t, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha})$ by truncating the distributions given by Equations 10 to 13 to the minimum and maximum amplitude consistent with the current fixation position s_{t-1} and text $\mathbf{X}$. Let w_l° (w_r°) denote the position of the left-most (right-most) character of the currently fixated word, and let w_l^+, w_r^+ denote these positions for the next word in $\mathbf{X}$. Let furthermore $l^\circ = w_l^\circ - s_{t-1}$, $r^\circ = w_r^\circ - s_{t-1}$, $l^+ = w_l^+ - s_{t-1}$, and $r^+ = w_r^+ - s_{t-1}$. Then

$$p(a|u_t = 1, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) =$$
$$\begin{cases} \mu\alpha_1(a|a \in [0, r^\circ]) : a > 0 \\ (1-\mu)\bar{\alpha}_1(-a|a \in [l^\circ, 0]) : a \le 0 \end{cases} \quad (16)$$
$$p(a|u_t = 2, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \alpha_2(a|a \in [l^+, r^+]) \quad (17)$$
$$p(a|u_t = 3, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \alpha_3(a|a \in (r^+, \infty)) \quad (18)$$
$$p(a|u_t = 4, s_{t-1}, \mathbf{X}, \boldsymbol{\alpha}) = \alpha_4(-a|a \in (-\infty, l^\circ)) \quad (19)$$

defines the appropriately truncated distributions.

3.2 Model of Fixation Durations

The model for fixation durations (Equation 9) is similarly specified by saccade type-specific densities,

$$p(d|u_t = u, \boldsymbol{\delta}) = \delta_u(d) \quad \text{for } u \in \{1, 2, 3, 4\} \quad (20)$$

and a density for the initial fixation durations

$$p(d_1|\mathbf{X}, \boldsymbol{\delta}) = \delta_0(d_1) \quad (21)$$

where $\delta_0, ..., \delta_4$ are aggregated into model component $\boldsymbol{\delta}$. Unlike saccade amplitude, the fixation duration is not constrained by the text structure and accordingly densities are not truncated. This concludes the definition of the model $p(\mathbf{S}|\mathbf{X}, \boldsymbol{\gamma})$.

3.3 Prior Distributions

The prior distribution over the entire model $\boldsymbol{\gamma}$ factorizes over the model components as

$$p(\boldsymbol{\gamma}|\lambda, \rho, \kappa) = \tag{22}$$

$$p(\boldsymbol{\pi}|\lambda)p(\mu|\rho)p(\bar{\alpha}_1|\kappa) \prod_{i=0}^{4} p(\alpha_i|\kappa) \prod_{i=0}^{4} p(\delta_i|\kappa)$$

where $p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\lambda)$ is a symmetric Dirichlet prior and $p(\mu) = \text{Beta}(\mu|\rho)$ is a Beta prior. The key challenge is to develop appropriate priors for the densities defining saccade amplitude $(p(\bar{\alpha}_1|\kappa), p(\alpha_i|\kappa))$ and fixation duration $(p(\delta_i|\kappa))$ distributions. Empirically, we observe that amplitude and duration distributions tend to be close to gamma distributions—see the example in Figure 1.

Our goal is to exploit the prior knowledge that distributions tend to be closely approximated by gamma distributions, but allow the model to deviate from the gamma assumption in case there is enough evidence in the data. To this end, we define a prior over densities that concentrates probability mass around the gamma family. For all densities $f \in \{\bar{\alpha}_1, \alpha_0, ..., \alpha_4, \delta_0, ..., \delta_4\}$, we employ identical prior distributions $p(f|\kappa)$. Intuitively, the prior is given by first drawing a density function from the gamma family and then drawing the final density from a Gaussian process (with covariance function κ) centered at this function. More formally, let

$$\mathcal{G}(x|\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x))}{\int \exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x'))\mathrm{d}x'} \tag{23}$$

denote the gamma distribution in exponential family form, with sufficient statistics $\mathbf{u}(x) = (\log(x), x)^\mathsf{T}$ and parameters $\boldsymbol{\eta} = (\eta_1, \eta_2)$. Let $p(\boldsymbol{\eta})$ denote a prior over the gamma parameters, and define

$$p(f|\kappa) = \int p(\boldsymbol{\eta})p(f|\boldsymbol{\eta}, \kappa)\mathrm{d}\boldsymbol{\eta} \tag{24}$$

where $p(f|\boldsymbol{\eta}, \kappa)$ is given by drawing

$$g \sim \mathcal{GP}(0, \kappa) \tag{25}$$

from a Gaussian process prior $\mathcal{GP}(0, \kappa)$ with mean zero and covariance function κ, and letting

$$f(x) = \frac{\exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x) + g(x))}{\int \exp(\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(x') + g(x'))\mathrm{d}x'}. \tag{26}$$

Note that decreasing the variance of the Gaussian process means regularizing $g(x)$ towards zero, and therefore Equation 26 towards Equation 23. This concludes the specification of the prior $p(\boldsymbol{\gamma}|\lambda, \rho, \kappa)$.

The density model defined by Equations 24 to 26 draws on ideas from the large body of literature on GP-based density estimation, for example by Adams et al. (2009), Leonard (1978), or Tokdar et al. (2010), and semiparametric density estimation, e.g. as discussed by Yang (2009), Lenk (2003) or Hjort & Glad (1995). However, note that existing density estimation approaches are not applicable off-the-shelf as in our domain distributions are truncated differently at each observation due to constraints that arise from the way eye movements interact with the text structure (Equations 16 to 19).

4 Inference

To solve Equation 4, we need to integrate for each $r \in \mathcal{R}$ over the reader-specific model $\boldsymbol{\gamma}_r$. To reduce notational clutter, let $\boldsymbol{\gamma} \in \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_R\}$ denote a reader-specific model, and let $\mathcal{S} \in \{\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(R)}\}$ denote the eye movement observations of that reader on the training texts $\mathcal{X}$. We approximate

$$\int p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma})p(\boldsymbol{\gamma}|\mathcal{X}, \mathcal{S})\mathrm{d}\boldsymbol{\gamma} \approx \frac{1}{K} \sum_{k=1}^{K} p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}^{(k)})$$

by a sample $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(K)}$ of models drawn by

$$\boldsymbol{\gamma}^{(k)} \sim p(\boldsymbol{\gamma}|\mathcal{X}, \mathcal{S}, \lambda, \rho, \kappa),$$

where $p(\boldsymbol{\gamma}|\mathcal{X}, \mathcal{S}, \lambda, \rho, \kappa)$ is the posterior as given by Equation 6 but with the dependence on the prior hyperparameters λ, ρ, κ made explicit. Note that with $\mathcal{X}$ and $\mathcal{S}$, all saccade types u_t are observed. Together with the factorizing prior (Equation 22), this means that the posterior factorizes according to

$$p(\boldsymbol{\gamma}|\mathcal{X}, \mathcal{S}, \lambda, \rho, \kappa) = p(\boldsymbol{\pi}|\mathcal{X}, \mathcal{S}, \lambda)p(\mu|\mathcal{X}, \mathcal{S}, \rho)$$

$$\cdot p(\bar{\alpha}_1|\mathcal{X}, \mathcal{S}, \kappa) \prod_{i=0}^{4} p(\alpha_i|\mathcal{X}, \mathcal{S}, \kappa) \prod_{i=0}^{4} p(\delta_i|\mathcal{X}, \mathcal{S}, \kappa)$$

as is easily seen from the graphical model in Figure 2. Obtaining samples $\boldsymbol{\pi}^{(k)} \sim p(\boldsymbol{\pi}|\mathcal{X}, \mathcal{S})$ and $\mu^{(k)} \sim p(\mu|\mathcal{X}, \mathcal{S})$ is straightforward because their prior distributions are conjugate to the likelihood terms. Let now $f \in \{\bar{\alpha}_1, \alpha_0, ..., \alpha_4, \delta_0, ..., \delta_4\}$

denote a particular density in the model. The posterior $p(f|\mathcal{X}, \mathcal{S}, \kappa)$ is proportional to the prior $p(f|\kappa)$ (Equation 24) multiplied by the likelihood of all observations that are generated by this density, that is, that are generated according to Equation 14, 16, 17, 18, 19, 20, or 21. Let $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|})^{\mathsf{T}} \in \mathbb{R}^{|\mathbf{y}|}$ denote the vector of all observations generated from density f, and let $\mathbf{l} = (l_1, \ldots, l_{|\mathbf{l}|})^{\mathsf{T}} \in \mathbb{R}^{|\mathbf{l}|}$, $\mathbf{r} = (r_1, \ldots, r_{|\mathbf{r}|})^{\mathsf{T}} \in \mathbb{R}^{|\mathbf{r}|}$ denote the corresponding left and right boundaries of the truncation intervals (again see Equations 14 to 21), where for densities that are not truncated we take $l_i = 0$ and $r_i = \infty$ throughout. Then the likelihood of the observations generated from f is

$$p(\mathbf{y}|f, \mathbf{l}, \mathbf{r}) = \prod_{i=1}^{|\mathbf{y}|} f(y_i|y_i \in [l_i, r_i]) \qquad (27)$$

and the posterior over f is given by

$$p(f|\mathcal{X}, \mathcal{S}, \kappa) \propto p(f|\kappa)p(\mathbf{y}|f, \mathbf{l}, \mathbf{r}). \qquad (28)$$

Note that $\mathbf{y}$, $\mathbf{l}$ and $\mathbf{r}$ are observable from $\mathcal{X}, \mathcal{S}$.

We obtain samples from the posterior given by Equation 28 from a Metropolis-Hastings sampler that explores the space of densities $f : \mathbb{R} \to \mathbb{R}$, generating density samples $f^{(1)}, \ldots, f^{(K)}$. A density f is given by a combination of gamma parameters $\boldsymbol{\eta} \in \mathbb{R}^2$ and function $g : \mathbb{R} \to \mathbb{R}$; specifically, f is obtained by multiplying the gamma distribution with parameters $\boldsymbol{\eta}$ by $\exp(g)$ and normalizing appropriately (Equation 26). During sampling, we explicitly represent a density sample $f^{(k)}$ by its gamma parameters $\boldsymbol{\eta}^{(k)}$ and function $g^{(k)}$. The proposal distribution of the Metropolis-Hastings sampler is

$$q(\boldsymbol{\eta}^{(k+1)}, g^{(k+1)}|\boldsymbol{\eta}^{(k)}, g^{(k)}) =$$
$$p(g^{(k+1)}|\kappa)\mathcal{N}(\boldsymbol{\eta}^{(k+1)}|\boldsymbol{\eta}^{(k)}, \sigma^2 \mathbf{I})$$

where $p(g^{(k+1)}|\kappa)$ is the probability of $g^{(k+1)}$ according to the GP prior $\mathcal{GP}(0, \kappa)$ (Equation 25), and $\mathcal{N}(\boldsymbol{\eta}^{(k+1)}|\boldsymbol{\eta}^{(k)}, \sigma^2 \mathbf{I})$ is a symmetric proposal that randomly perturbs the old state $\boldsymbol{\eta}^{(k)}$ according to a Gaussian. In every iteration k a proposal $\boldsymbol{\eta}^\star, g^\star \sim q(\boldsymbol{\eta}, g|\boldsymbol{\eta}^{(k)}, g^{(k)})$ is drawn based on the old state $(\boldsymbol{\eta}^{(k)}, g^{(k)})$. The acceptance probability is $A(\boldsymbol{\eta}^\star, g^\star|\boldsymbol{\eta}^{(k)}, g^{(k)}) = \min(1, Q)$ with

$$Q =$$
$$\frac{q(\boldsymbol{\eta}^{(k)}, g^{(k)}|\boldsymbol{\eta}^\star, g^\star)p(\boldsymbol{\eta}^\star)p(g^\star|\kappa)p(\mathbf{y}|f^\star, \mathbf{l}, \mathbf{r})}{q(\boldsymbol{\eta}^\star, g^\star|\boldsymbol{\eta}^{(k)}, g^{(k)})p(\boldsymbol{\eta}^{(k)})p(g^{(k)}|\kappa), p(\mathbf{y}|f^{(k)}, \mathbf{l}, \mathbf{r})}.$$

Here, $p(\boldsymbol{\eta}^\star)$ is the prior probability of gamma parameters $\boldsymbol{\eta}^\star$ (Section 3.3) and $p(\mathbf{y}|f^\star, \mathbf{l}, \mathbf{r})$ is given by Equation 27 where $f^\star$ is obtained from $\boldsymbol{\eta}^\star$, $g^\star$ according to Equation 26.

To compute the likelihood terms $p(\mathbf{y}|f^{(k)}, \mathbf{l}, \mathbf{r})$ (Equation 27) and also to compute the likelihood of test data under a model (Equation 5), the density $f : \mathbb{R} \to \mathbb{R}$ needs to be evaluated. According to Equation 26, f is represented by parameter vector $\boldsymbol{\eta}$ together with the nonparametric function $g : \mathbb{R} \to \mathbb{R}$. As usual when working with distributions over functions in a Gaussian process framework, the function g only needs to be represented at those points for which we need to evaluate it. Clearly, this includes all observations of saccade amplitudes and fixation durations observed in the training and test set. However, we also need to evaluate the normalizer in Equation 26, and (for $f \in \{\alpha_1, \bar{\alpha}_1, \alpha_2, \alpha_3, \alpha_4\}$) the additional normalizer required when truncating the distribution (see Equation 15). As these integrals are one-dimensional, they can be solved relatively accurately using numerical integration; we use 2-point Newton-Cotes quadrature. Newton-Cotes integration requires the evaluation (and thus representation) of g at an additional set of equally spaced supporting points.

When the set of test observations $\bar{\mathcal{S}}, \bar{\mathcal{X}}$ is large, the need to evaluate $p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}^{(k)})$ for all $\boldsymbol{\gamma}_k$ and all test observations leads to computational challenges. In our experiments, we use a heuristic to reduce computational load. While generating samples, densities are only represented at the training observations and the supporting points needed for Newton-Cotes integration. We then estimate the mean of the posterior by $\hat{\boldsymbol{\gamma}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\gamma}^{(k)}$, and approximate $\frac{1}{K} \sum_{k=1}^{K} p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \boldsymbol{\gamma}^{(k)}) \approx p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \hat{\boldsymbol{\gamma}})$. To evaluate $p(\bar{\mathcal{S}}|\bar{\mathcal{X}}, \hat{\boldsymbol{\gamma}})$, we infer the approximate value of the density $\hat{\boldsymbol{\gamma}}$ at a test observation by linearly interpolating based on the available density values at the training observations and supporting points.

5 Empirical Study

We conduct a large-scale study of biometric identification performance using the same setup as discussed by Landwehr et al. (2014) but a much larger set of individuals (251 rather than 20).
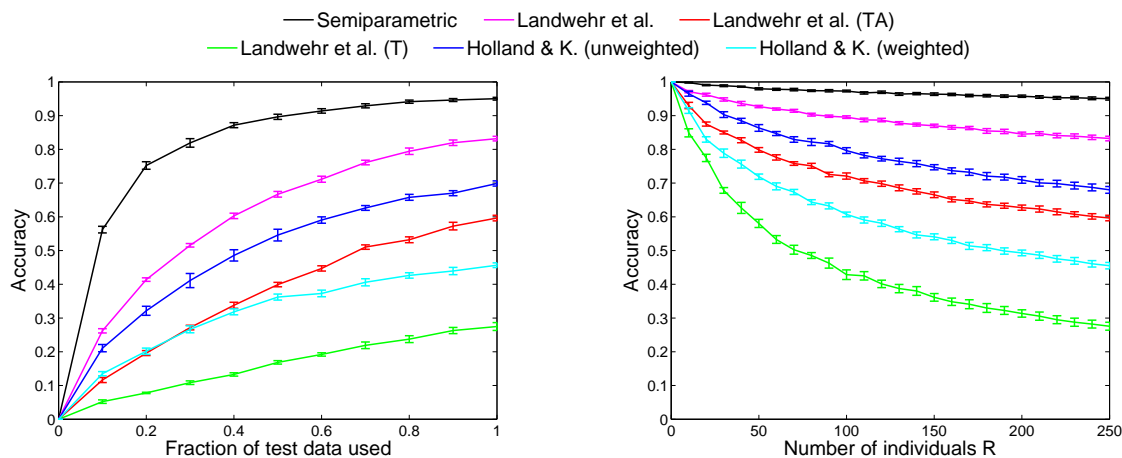
Eye movement records for 251 individuals are

590

Figure 3: Multiclass accuracy over number of test observations (left) and number of individuals R (right) with standard errors.

Method	Accuracy
Semiparametric	0.9502 ± 0.0130
Semiparametric (TD)	0.8853 ± 0.0142
Semiparametric (TA)	0.7717 ± 0.0361
Landwehr et al.	0.8319 ± 0.0218
Landwehr et al. (TA)	0.5964 ± 0.0262
Landwehr et al. (T)	0.2749 ± 0.0369
Holland & K. (unweighted)	0.6988 ± 0.0241
Holland & K. (weighted)	0.4566 ± 0.0220

Table 1: Multiclass identification accuracy $\pm$ standard error.

obtained from an EyeLink II system with a 500-Hz sampling rate (SR Research, Ongoode, Ontario, Canada) while reading sentences from the *Potsdam Sentence Corpus* (Kliegl et al., 2006). There are 144 sentences in the corpus, which we split into equally sized sets of training and test sentences. Individuals read between 100 and 144 sentences, the training (testing) observations for one individual are the observations on those sentences in the training (testing) set of sentences that the individual has read. Results are averaged over 10 random train-test splits. Each sentence is shown as a single line on the screen.

We study the semiparametric model discussed in Section 3 with MCMC inference as presented in Section 4 (denoted *Semiparametric*[1]). We employ a squared exponential covariance function $\kappa(x, x') = \alpha \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$, where the multiplicative constant α is tuned on the training data by cross-

validation and the bandwidth σ is set to the average distance between points in the training data. The Beta and Dirichlet parameters λ and ρ are set to one (Laplace smoothing), the prior $p(\boldsymbol{\eta})$ for the Gamma parameters is uninformative. We use backoff-smoothing as discussed by Landwehr et al. (2014). We initialize the sampler with the maximum-likelihood Gamma fit and perform 10000 sampling iterations, 5000 of which are burn-in iterations. As a baseline, we study the model by Landwehr et al. (2014) (*Landwehr et al.*) and simplified versions proposed by them that only use saccade type and amplitude (*Landwehr et al. (TA)*) or saccade type (*Landwehr et al. (T)*). We also study the weighted and unweighted version of the feature-based model of Holland & Komogortsev (2012) with a feature set adapted to the Potsdam Sentence Corpus data as described in Landwehr et al. (2014).

We note that there are two recent extensions of the feature-based model (by Rigas et al. (2016) and Abdulin & Komogortsev (2015)) that are unfortunately not applicable in our empirical setting but might yield improved results in other scenarios. Rigas et al. (2016) study a model that is focused on representing reader-specific differences in saccadic vigor and acceleration, which are both derived from the dynamics of saccadic velocity. In the preprocessed data set that we use, saccadic velocities are not available, therefore we do not make use of velocities in our model and cannot easily compare against their model. Abdulin & Komogortsev (2015) study a model that is based on features that relate eye move-
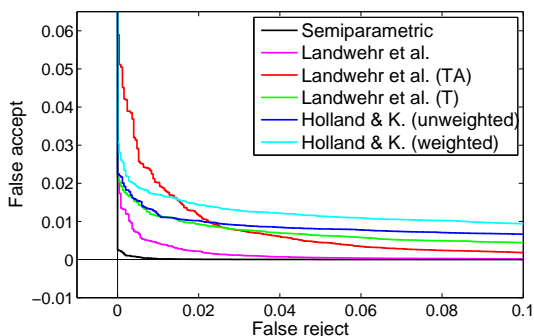
---

[1]An implementation is available at github.com/abdelwahab/SemiparametricIdentification

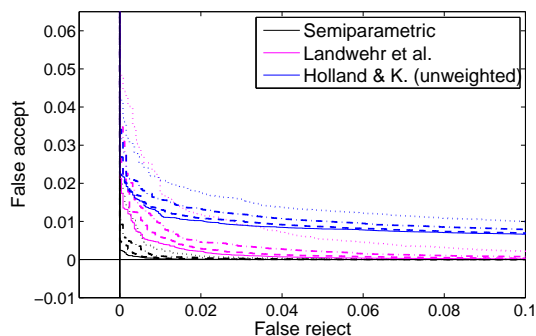**Figure 4:** False-accept over false-reject rate when varying τ.

Figure 5: False-accept over false-reject rate when using 40% (dotted), 60% (dashed-dotted), 80% (dashed), and 100% (solid) of test observations, for selected subset of methods.

Method	Area under curve
Semiparametric	0.0000119
Semiparametric (TD)	0.0000821
Semiparametric (TA)	0.0001833
Landwehr et al.	0.0001743
Landwehr et al. (TA)	0.0010371
Landwehr et al. (T)	0.0017040
Holland & K. (unweighted)	0.0027853
Holland & K. (weighted)	0.0039978

Table 2: Area under the curve in binary classification setting.

ments to the 2D text structure, that is, to the way words are arranged into lines in a text. As in our empirical study each sentence is presented as a single line on screen, this 2D structure does not exist. Moreover, Abdulin & Komogortsev (2015) only report accuracy improvements for their method in a setting where individuals have to be identified in the future based on data collected in the past (*aging test*), which is not the focus of our study.

We first study multiclass identification accuracy. All test observations of one particular individual constitute one test example; the task is to infer the individual that has generated these test observations. Multiclass identification accuracy is the fraction of cases in which the correct individual is identified. Table 1 shows multiclass identification accuracy for all methods, including variants of *Semiparametric* discussed below. We observe that *Semiparametric* outperforms *Landwehr et al.*, reducing the error by more than a factor of three. Consistent with results reported in Landwehr et al. (2014), *Holland & K. (unweighted)* is less accurate than *Landwehr et al.*, but more accurate than the simplified variants. We next study how the amount of data available at test time—that is, the amount of time we can observe a reader before having to make a decision—influences accuracy. Figure 3 (left) shows identification accuracy as a function of the fraction of test data available, obtained by randomly removing a fraction of sentences from the test set. We observe that identification accuracy steadily improves with more test observations for all methods. Figure 3 (right) shows identification accuracy when varying the number R of individuals that need to be distinguished. We randomly draw a subset of R individuals from the set

of 251 individuals, and perform identification based on only these individuals. Results are averaged over 10 such random draws. As expected, accuracy improves if fewer individuals need to be distinguished.

We next study a binary setting in which for each individual and each set of test observations a decision has to be made whether or not the test observations have been generated by that individual. This setting more closely matches typical use cases for the deployment of a biometric system. Let $\bar{\mathcal{X}}$ denote the text being read at test time, and let $\bar{\mathcal{S}}$ denote the observed eye movement sequences. Our model infers for each reader $r \in \mathcal{R}$ the marginal likelihood $p(\bar{\mathcal{S}}|r, \bar{\mathcal{X}}, \mathcal{X}, \mathcal{S}^{(1:R)})$ of the eye movement observations under the reader-specific model (Equation 3). The binary decision is made by dividing this marginal likelihood by the average marginal likelihood assigned to the observations by all reader-specific models, and comparing the result to a threshold τ. Figure 4 shows the fraction of false accepts as a function of false rejects as the threshold τ is varied, averaged over all individuals. The *Landwehr et al.* model and variants also assign a
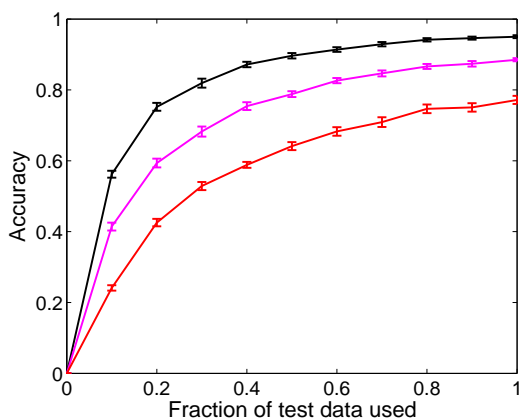
Figure 6: Multiclass accuracy over number of test observations with standard errors for *Semiparametric* variants.

Figure 7: False-accept over false-reject rate when varying τ for the *Semiparametric* variants.

reader-specific likelihood to novel test observations; we compute the same statistics again by normalizing the likelihood and comparing to a threshold τ. Finally, *Holland & K. (unweighted)* and *Holland & K. (weighted)* compute a similarity measure for each combination of individual and set of test observations, which we normalize and threshold analogously. We observe that *Semiparametric* accomplishes a false-reject rate of below 1% at virtually no false accepts; *Landwehr et al.* and variants tend to perform better than *Holland & K. (unweighted)* and *Holland & K. (weighted)*. Table 2 shows the error under the curve for the experiment shown in Figure 4, as well as for variants of *Semiparametric* discussed below.

We finally study the contribution of the individual model components for saccade type, saccade amplitude, and fixation duration (see Figure 2) by removing the corresponding model components, as in Landwehr et al. (2014). By *Semiparametric (TD)* we denote a variant of *Semiparametric* in which the variable a_t and the corresponding distribution is removed, that is, only the distribution over the saccade type and duration is modeled. *Semiparametric (TA)* denotes a variant in which the variable d_t and the corresponding distribution is removed. Figure 6 shows identification accuracy as a function of the fraction of test data available for model variants *Semiparametric (TD)* and *Semiparametric (TA)* in comparison to *Semiparametric*; results for these variants are also included in Table 1. Figure 7 shows the fraction of false accepts as a function of
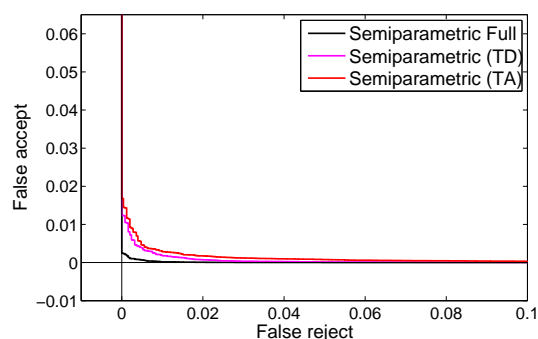
false rejects in the binary classification setting discussed above for these two model variants; Table 2 includes area under the curve results for the experiment shown in Figure 7. We observe that accuracy is substantially reduced when removing any model component. Note that if both the amplitude and duration components of the model are removed, it becomes identical to the model *Landwehr et al. (T)*.

Training the joint model for all 251 individuals takes 46 hours on a single eight-core CPU (Intel Xeon E5520, 2.27GHz); predicting the most likely individual to have generated a set of 72 test sentences takes less than 2 seconds.

6 Conclusions

We have studied the problem of identifying readers unobtrusively during reading of arbitrary text. For fitting reader-specific distributions, we employ a Bayesian semiparametric approach that infers densities under a Gaussian process prior centered at the gamma family of distributions, striking a balance between robustness to sparse data and modeling flexibility. In an empirical study with 251 individuals, the model was shown to reduce identification error by more than a factor of three compared to earlier approaches to reader identification proposed by Landwehr et al. (2014) and Holland & Komogortsev (2012).

Acknowledgements

References

Evgeniy Abdulin and Oleg Komogortsev. 2015. Person verification via eye movement-driven text reading model. In *Proceedings of the Sixth International Conference on Biometrics: Theory, Applications and Systems*.

Ryan P. Adams, Iain Murray, and David J.C. MaxKay. 2009. Gaussian process density sampler. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*.

Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. 2005. Eye-movements as a biometric. In *Proceedings of the 14th Scandinavian Conference on Image Analysis*.

W. Robert Dixon. 1951. Studies in the psychology of reading. In W. S. Morse, P. A. Ballantine, and W. R. Dixon, editors, *Univ. of Michigan Monographs in Education No. 4*. Univ. of Michigan Press.

Ralf Engbert, Antje Nuthmann, Eike M. Richter, and Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Tadayoshi Hara, Daichi Mochihashi, Yoshino Kano, and Akiko Aizawa. 2012. Predicting word fixations in text with a CRF model for capturing general reading strategies among readers. In *Proceedings of the First Workshop on Eye-Tracking and Natural Language Processing*.

Julian Heister, Kay-Michael Würzner, and Reinhold Kliegl. 2012. Analysing large datasets of eye movements during reading. In James S. Adelman, editor, *Visual word recognition. Vol. 2: Meaning and context, individuals and development*, pages 102–130.

Nils L. Hjort and Ingrid K. Glad. 1995. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23(3):882–904.

Corey Holland and Oleg V. Komogortsev. 2012. Biometric identification via eye movement scanpaths in reading. In *Proceedings of the 2011 International Joint Conference on Biometrics*.

Edmund B. Huey. 1908. *The psychology and pedagogy of reading*. Cambridge, Mass.: MIT Press.

Pawel Kasprowski and Jozef Ober. 2004. Eye movements in biometrics. In *Proceedings of the 2004 International Biometric Authentication Workshop*.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135(1):12–35.

Oleg V. Komogortsev, Sampath Jayarathna, Cecilia R. Aragon, and Mechehoul Mahmoud. 2010. Biometric identification via an oculomotor plant mathematical model. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*.

Niels Landwehr, Sebastian Arzt, Tobias Scheffer, and Reinhold Kliegl. 2014. A model of individual differences in gaze control during reading. In *Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing*.

Peter J. Lenk. 2003. Bayesian semiparametric density estimation and model verification using a logistic-Gaussian process. *Journal of Computational and Graphical Statistics*, 12(3):548–565.

Tom Leonard. 1978. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society*, 40(2):113–146.

Franz Matties and Anders Søgaard. 2013. With blinkers on: robust prediction of eye movements across readers. In *Proceedings of the 2013 Conference on Empirical Natural Language Processing*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Erik D. Reichle, Alexander Pollatsek, Donald L. Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157.

Ioannis Rigas, George Economou, and Spiros Fotopoulos. 2012a. Biometric identification based on the eye movements and graph matching techniques. *Pattern Recognition Letters*, 33(6).

Ioannis Rigas, George Economou, and Spiros Fotopoulos. 2012b. Human eye movements as a trait for biometrical identification. In *Proceedings of the IEEE 5th International Conference on Biometrics: Theory, Applications and Systems*.

Ioannis Rigas, Oleg Komogortsev, and Reza Shadmehr. 2016. Biometric recognition via eye movements: Saccadic vigor and acceleration cues. *ACM Transaction on Applied Perception*, 13(2):1–21.

Surya T. Tokdar, Yu M. Zhuy, and Jayanta K. Ghoshz. 2010. Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian Analysis*, 5(2):319–344.

Ying Yang. 2009. Penalized semiparametric density estimation. *Statistics and Computing*, 19(1):355–366.

Youming Zhang and Martti Juhola. 2012. On biometric verification of a user by means of eye movement data mining. In *Proceedings of the 2nd International Conference on Advances in Information Mining and Management*.