Prospectus title

and subtitle!

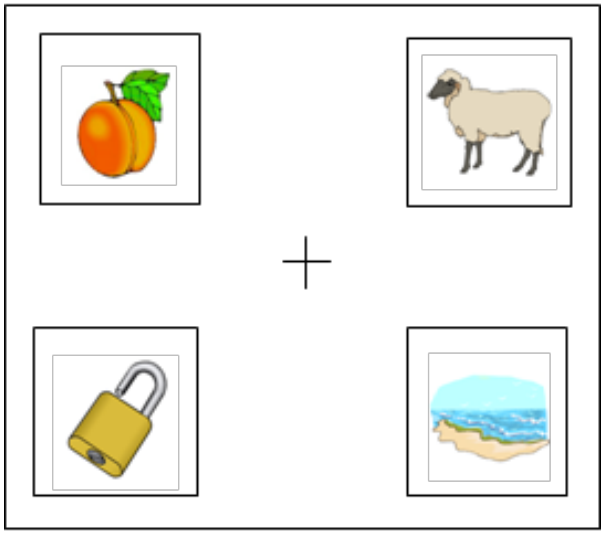
Collin Nolte

April 22, 2022

Outline

Problem statement

Overview



overview 2

The current method employed by `bdots` is to fit the observed y_{it} to an underlying curve f_θ, the fitting step given by

$$F : \{y\} \times f \rightarrow N\left(\hat{\theta}_i, \hat{\Sigma}_{\theta_i}\right)$$

Such that

$$\hat{\theta}_i = \text{argmin}_\theta \; ||y_{it} - f_\theta(t)||^2$$

These could be further specified by indicating a group for the observation $g = 1, \ldots, G$, where an individual may be in multiple groups (and ultimately, it will be group values being compared)

bdots

Bootstrapped difference in time series, paper originally by olseon, ecavanaugh, mcmurray, and brown

package written by seedorff

blah blah blah

bdots2

Current implementation of `bdots` involves three steps:

1. **Curve Fitting:** Fitting parametric curve to observed data
2. **Curve Refitting:** Manually estimating parameters in case of poor fit
3. **Bootstrap** Bootstrap curves to estimate group population curve

I also added a whole bunch of shit to make this package better

bdots add stuff

One function, `bdotsFit`, which can fit any curve to data (previously, one function for each curve)

Allows for arbitrary user-defined function to be used

Object returned by bdotsFit very flexible, allows formula definition of bootstrapping difference in bootstrap step

ggplot, correlation with fixed vector, just general all around better
Refitting step is interactive, can upload external data, saves progress

compare

maybe show old bdots vs new bdots, if i can find working code

bdots fit process

really, this covered in overview 2 slide

bdots bootstrap process

Here, we perform B bootstraps of the subject parameters to construct bootstrapped curves and confidence intervals. Assuming that each subject is in one *group*, we draw B samples of $\hat{\theta}_i$, where

$$\hat{\theta}_{ib} \sim N\left(\hat{\theta}_i, \Sigma_{\hat{\theta}_i}\right)$$

resulting in a $B \times p$ matrix, denoted M_i.

Doing this for each subject, we construct a $B \times p$ matrix of the average of bootstraps across iterations,

$$\overline{M} = \frac{1}{n}\sum_{i}^{n} M_i$$

(potentially confusing here – each subject has their own set of parameters, as distinct from group level pars?)

bootstrap cont

$\overline{M}$ is again a $B \times p$ matrix, each row representing the average parameter estimate of θ at each bootstrap b.

Each $1 \times p$ row of $\overline{M}$ returns a $1 \times T$ vector representing estimations of f_θ at each point t. Together, we have the $B \times T$ matrix $\overline{M}_f$. This gives an estimated fixation curve,

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \overline{M}_{\{b,\cdot\}_f}, \qquad \widehat{\text{se}}_f = \left[\frac{1}{B-1} \sum_{b=1}^{B} \left(\overline{M}_{\{i,\cdot\}_f} - \hat{f} \right)^2 \right]^{1/2}$$

bdots need

Have to tighten up functions

possibly jacknife

Look at replacing gnls with something else

still focus on not making vwp specific

Eyetracking data

In the context of vwp, subjects are lead through trials, with headsets monitoring eye position every 4ms.

Two types of events make up eyetracking data: *saccades* and *fixations*

A *saccade* represents the physical movement of an eye, lasting between 20-200ms. There is also about a 200ms oculomotor delay between planning an eye movement and it occuring

A *fixation* is characterized by a lack of movement, in which the eye is fixated on a particular location. The length of a fixation is more variable

Together, a saccade, followed by a subsequent fixation, is known as a *look*

Fixation curve

We define a *fixation curve*, $f_\theta(t)$ or $f(t|\theta)$ to represent a (usually parametric) function indicating the probability of fixating on a target at some time, t.

To disambiguate the term "target" when used in the VWP, we will let "Target" denote the object corresponding to a spoken word, while "target" will denote an object of interest

For example, with the four-parameter logistic, our target is typically the Target, while for the six-parameter double-gauss, our target is the Cohort

VWP and looks

Need for a grounding hypothesis linking eye tracking to cognitive activity

Magnuson paper on this, also assumptions from Bob's princess bride

For this dissertation, we will ignore the cognitive process and instead focus on recovering underlying fixation curves from observed data. The link between these two can be somebody elses' dissertation.

What we will consider here in review is high frequency sampling (HFS) and fixation-based sampling, augmented for target (FBS+T)

Sampling Paradigms

Two main assumptions focusing on right now: HFS and FBS+T (bob's paper)

HFS: High frequency sampling assumption, "if researcher is sampling at 4ms intervals, the fixation curve is assumped to derive from a probabilistic sample every 4ms"

FBS+T: Fixation-based sampling + target, series of discrete fixations with reasonable refractory period, treats fixations as primarily a readout ouf the unfolding decision, ignores the role of the fixation as an information gather behavior (pg 26 of princess bride), allows fixations to target to be slightly longer (once fixated, subject more likely to stay)

Importance of sampling paradigms

Why do we care? Well, it's relevant to the reconstruction of the curve. Specifically, either of these assumptions can be used to simulate eyetracking data. With recovery of the underlying fixation curve being our goal, we should be able to recover the underlying curve according to a particular hypothesis

To be covered later, but ideally we have a simulation the empirically matches data collected from eyetracking. While we don't quite have that yet, we have two we can explore

High Frequency Sampling

For subjects $i = 1, \ldots, n$, trials $j = 1, \ldots, J$, and time points $t = 1, \ldots, T$, the current method of estimating this curve is

$$y_{it} = \frac{1}{J} \sum_{j=1}^{J} z_{ijt}$$

where $z_{ijt} = \{0, 1\}$, conditional on the measured fixation at timepoint t in trial j.

Everything above this is independent of HFS and should go on slide comparing saccades to proportion fixations

Under the HFS assumption (employed by bdots), the vector y_i serves as a direct observation of $f_\theta(t)$ for subject i ($f_{\theta_i}(t)$?)

HFS wrong

its wrong. bob even said so

but look, plots are fine

FBS+T

1. fixations follow gamma distribution
2. extra time for target
3. doesn't take into account prior information (no learning)

there are also some plots showing bias

Mathematical Structure of Problem

The primary variable of interest is the *underlying activation curve*, $x \equiv x(t)$, a cognitive process neither measured or observed

Some function of this process is captured by a *fixation curve*, $H(x) = y(?)$, which, depending on the context (target or competitor), is represented by a parametric function $f(t|\theta)$ or $f_\theta(t)$

For subjects $i = 1, \ldots, n$, trials $j = 1, \ldots, J$, and time points $t = 1, \ldots, T$, the current method of estimating this curve is

$$y_{it} = \frac{1}{J} \sum_{j=1}^{J} z_{ijt}$$

where $z_{ijt} = \{0, 1\}$, conditional on the measured fixation at timepoint t in trial j.

Fixation vs Sacade

Current methods employ

$$y_{it} = \frac{1}{J}\sum_{j=1}^{J} z_{ijt}$$

where z_{ijt} denotes the *fixation* of subject i in trial j at time t. Insomuch as y_{it} represents the observed $f_\theta(t)$, we have an issue in that $f_\theta(t)$ represents the probability of *fixating* on a target at time t, rather than the proportion of fixations already on the target at t

In other words, we looking at the wrong thing

fix vs saccade sim

just use fbs+t paradigm, really, since other one is wrong (and not biased)

identity theroem and time windows

The identity theorem for analytic functions states: given functions f and g analytic on (open and connected) domain D, if $f = g$ on some $S \subseteq D$, where S has an accumulation point, then $f = g$ on D (wikipedia)

In other words, letting D be time, if we were able to identify values $\hat{\theta}$ such that $f_{\hat{\theta}} = f_{\theta}$ some interval S, then we should find that $f_{\hat{\theta}} = f_{\theta}$ on D

In other words again, if we could identify some interval where our approximation of f was best, we could extrapolate this for the rest of the domain

id theorem + window

In most contexts, we have $f_\theta(t) \sim Bern(p(t))$ where $var(p(t)) = p(t)(1 - p(t))$, the least amount of variability will occur as $p(t)$ approaches 0 or 1

Taking the logistic curve as an example, this will be towards the beginnings and ends of a trial, with maximal variability occuring at the crossover points

window sims

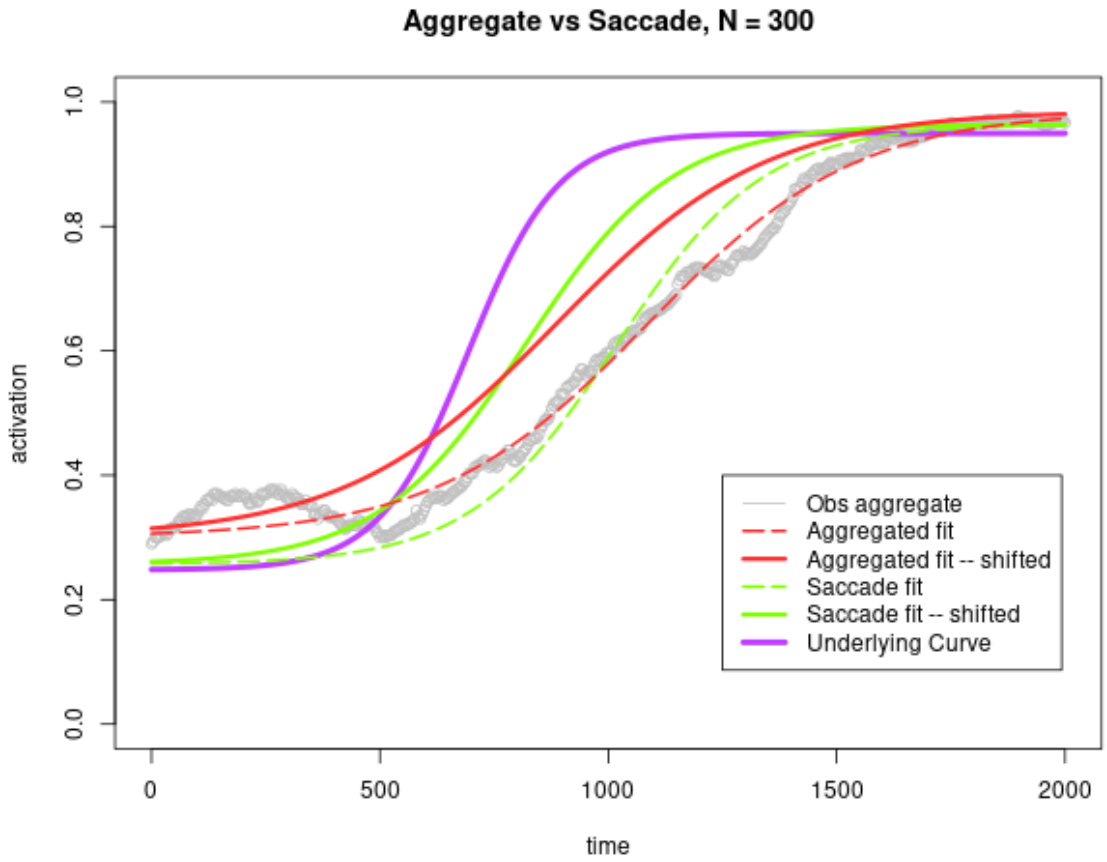
Here, we run simulations assuming fbst and reconstruct fits using bdots

Plot them as-is, along with 200ms shift to account for occulomotor delay

To examine various time windows, have included simulations that will sample at a fixed rate (every 25ms) within a specified time interval, rather than relying on length of fixations or identified target

Each simulation is based off a singular underlying fixation function with identical parameters across trials

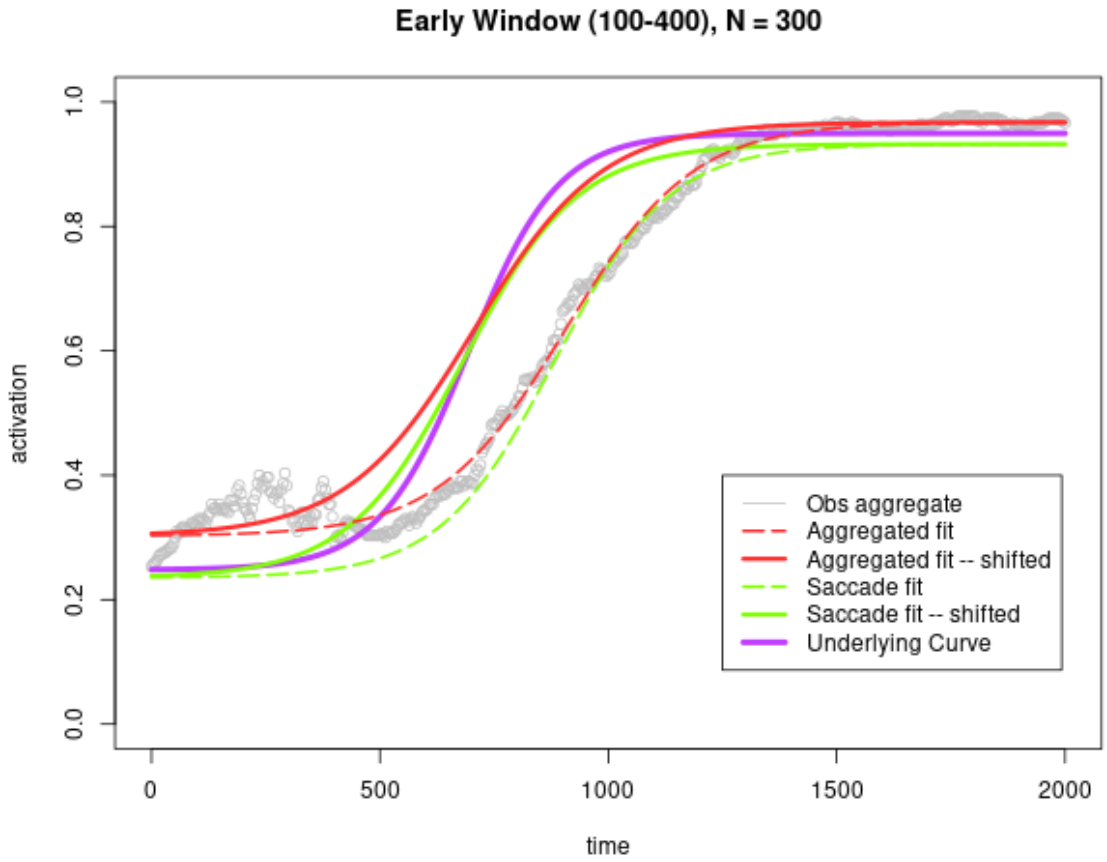
both



MISE

Aggregate	Saccade	Aggregate – Shifted	Saccade – Shifted	Underlying
57.86	58.10	20.65	10.84	0.00

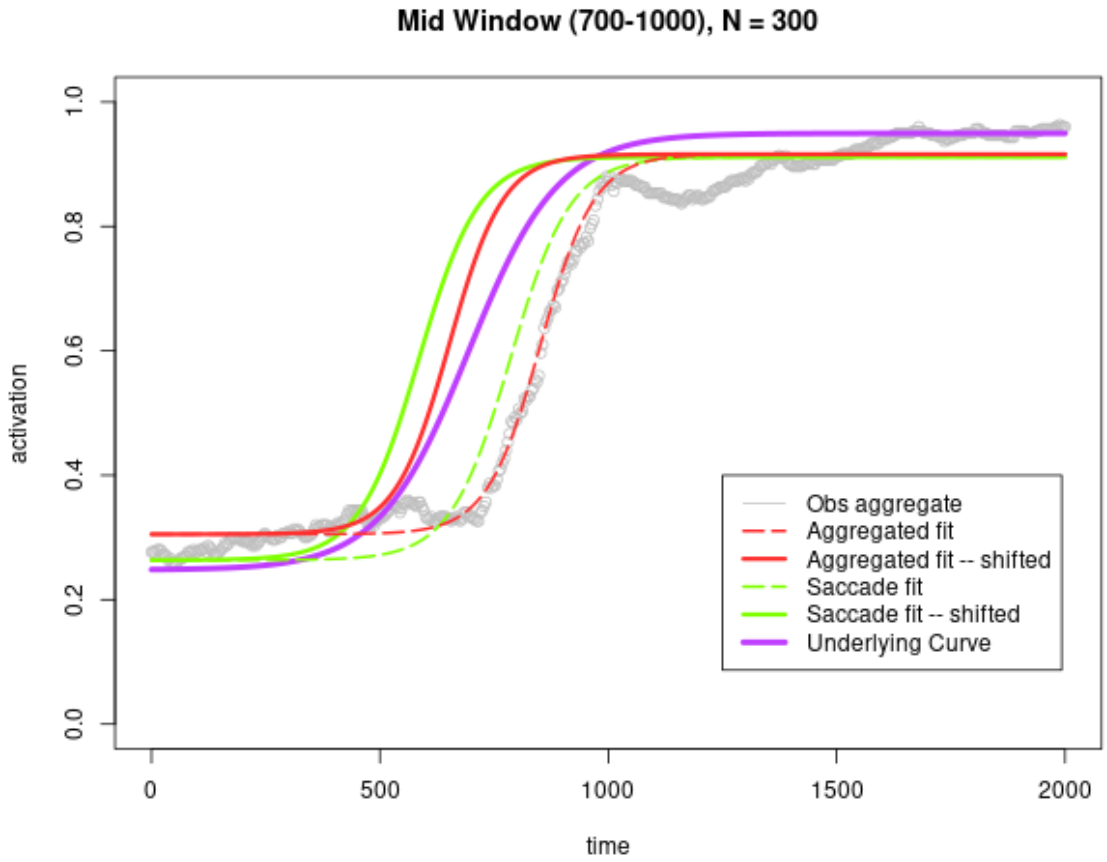
early window



MISE

Aggregate	Saccade	Aggregate – Shifted	Saccade – Shifted	Underlying
22.51	29.94	4.18	1.14	0.00

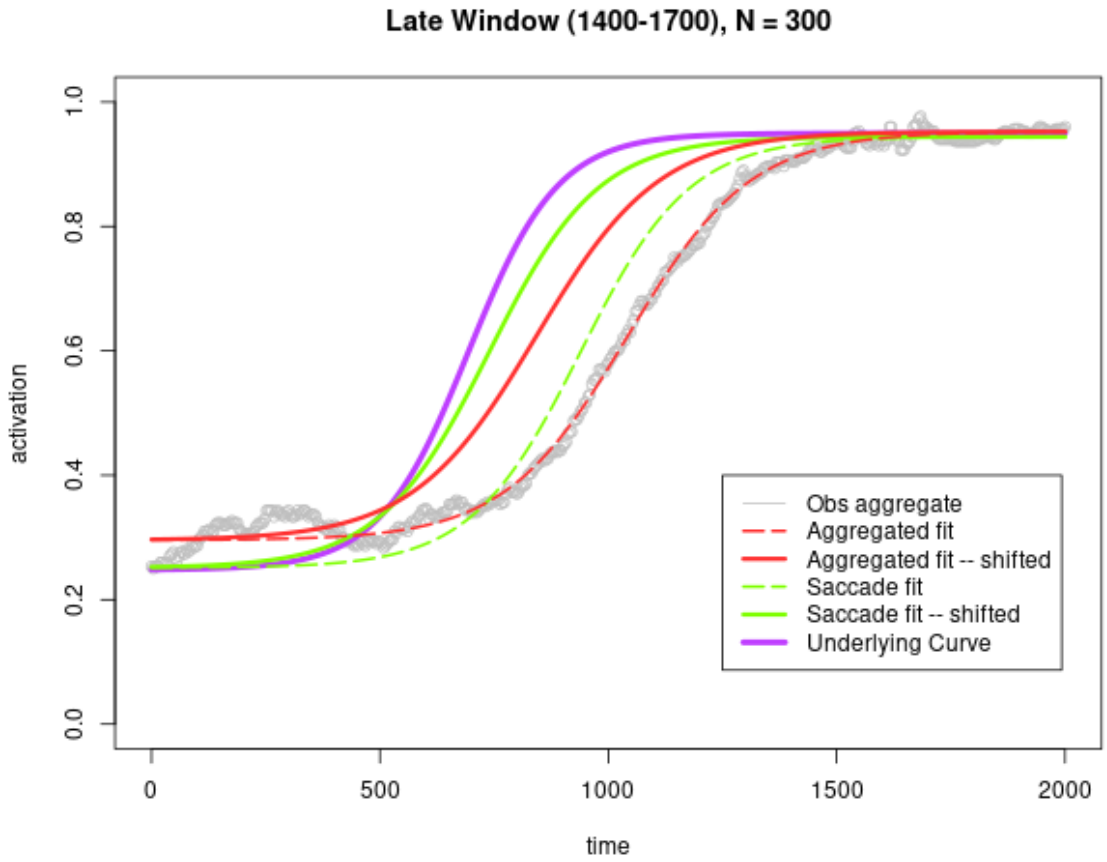
mid window



MISE

Aggregate	Saccade	Aggregate – Shifted	Saccade – Shifted	Underlying
21.16	10.74	4.75	10.76	0.00

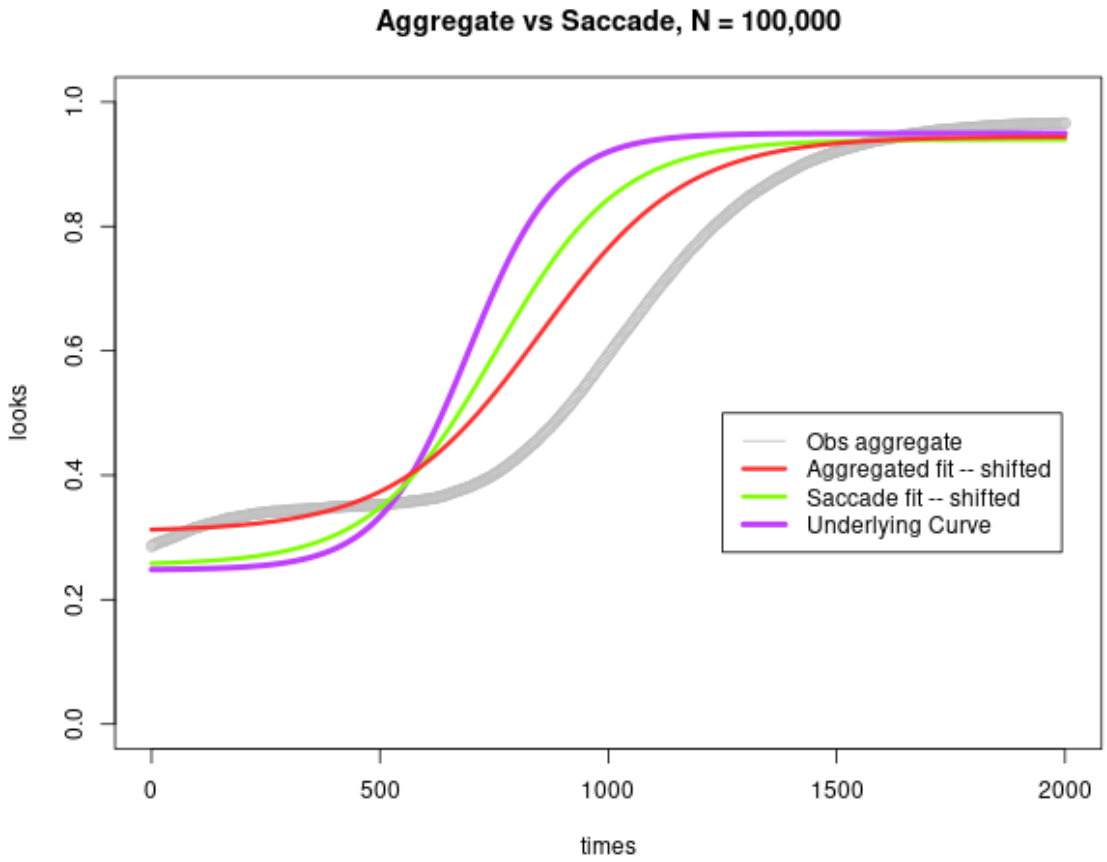
late window



MISE

Aggregate	Saccade	Aggregate – Shifted	Saccade – Shifted	Underlying
61.89	42.18	12.77	2.18	0.00

asymptotics



MISE

Aggregate – Shifted	Saccade – Shifted	Underlying
15.57	4.31	0.00

mise

It's worth pointing out that each time window will appear better than the trial with no time window, as these ultimately ended up including more samples per trial

MISE

	Standard	Early	Mid	Late	N
Aggregate	57.86	22.51	21.16	61.89	NA
Saccade	58.10	29.94	10.74	42.18	NA
Aggregate – Shifted	20.65	4.18	4.75	12.77	15.57
Saccade – Shifted	10.84	1.14	10.76	2.18	4.31
Underlying	0.00	0.00	0.00	0.00	0.00

other issues

Density of saccades over time

identity theorem

still does 0-2000ms (how to deal with identifying target/RT)

"Ignores the role of the fixation as an information gathering behavior"

Can show sims demonstrating time/density/etc with saccades and wololololo

current simulation (fbs+t)

Begin by creating a subject

1. Draw $\theta \sim N(\theta^*, \Sigma_{\theta^*})$ s.t $max(f_\theta) \in (0.6, 1)$
2. Draw parameters for distribution Γ for duration of fixation to non-target
3. Draw parameters for distribution Γ_T for duration of fixation to target
4. Return $(\theta, \Gamma, \Gamma_T)$

Current sim (fbs+t)

Single trial. With vector of times, `time`, run the following simulation:

```
pars <- makeSubject()
currTime <- min(time) - runif(1)E(Γ)
lastTime <- currTime - Γ - runif(1)E(Γ)

while(currTime < max(time) {
   p <- f(lasttime|θ)
   target <- runif(1) < p
   duration <- ifelse(target, Γ, Γ_T)
   lasttime <- currTime
   currTime <- currTime + duration
}
```

Limitations

Not really recovering cognitive curve