Prospectus title

and subtitle!

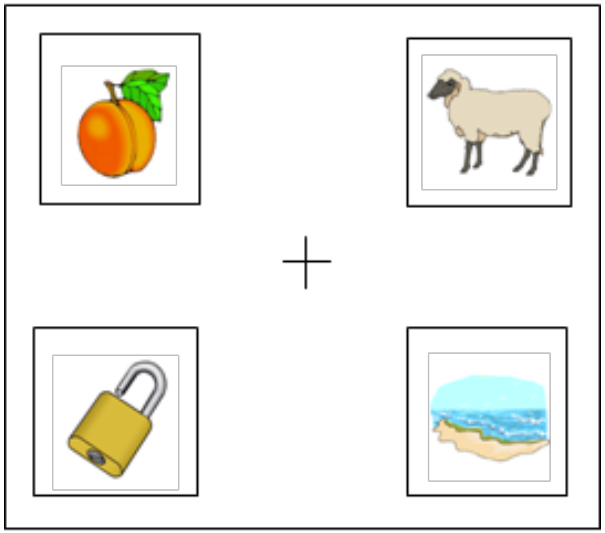
Collin Nolte

April 22, 2022

Outline

Problem statement

Overview



bdots2

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

package written by seedorff

blah blah blah

bdots

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

Fitting Process

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)

Bootstrapping 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,

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

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,

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

Updates

Fitting process has been reduced to a single function, `bdotsFit` which can accept arbitrary functions provided by the user, as well as an arbitrary number of experimental groups or conditions

Object returned by `bdotsFit` inherits `data.frame` class

Introduction of a number of useful generics including `plot`, `summary`, `coef`, etc.,

Allows for arbitrary user-defined function to be used

Formula definition introduced in bootstrapping step, removing need to prespecify differences or differences of differences between curves

Refitting step is interactive, can upload external data, saves progress

Fitting with bdots

```
## Old bdots
fit0 <- doubleGauss.fit(
  data = dat, # Requires columns "Subject", "Time", and "Group
  col = 4, # Specify outcome with numeric position
  concave = TRUE, # argument tied to curve function
  diffs = TRUE) # Requires column "Curve" with values 1,2

## New bdots
fit <- bdotsFit(data = dat,
  subject = "Subject",
  time = "Time",
  y = "Fixations",
  group = c("Group", "LookType"),
  curveType = doubleGauss(concave = TRUE))
```

Bootstrap with bdots

```
## Old bdots
boot0 <- doubleGauss.boot(
  part1.list = fit0,
  paired = TRUE) # Must indicate if observations paired

## New bdots
boot <- bdotsBoot(
  Fixations ~ Group(50, 65) + LookType(Cohort),
  bdObj = fit)

boot <- bdotsBoot(
  diffs(Fixations, Group(50, 65)) ~ LookType(Cohort, Unrelated),
  bdObj = fit)
```

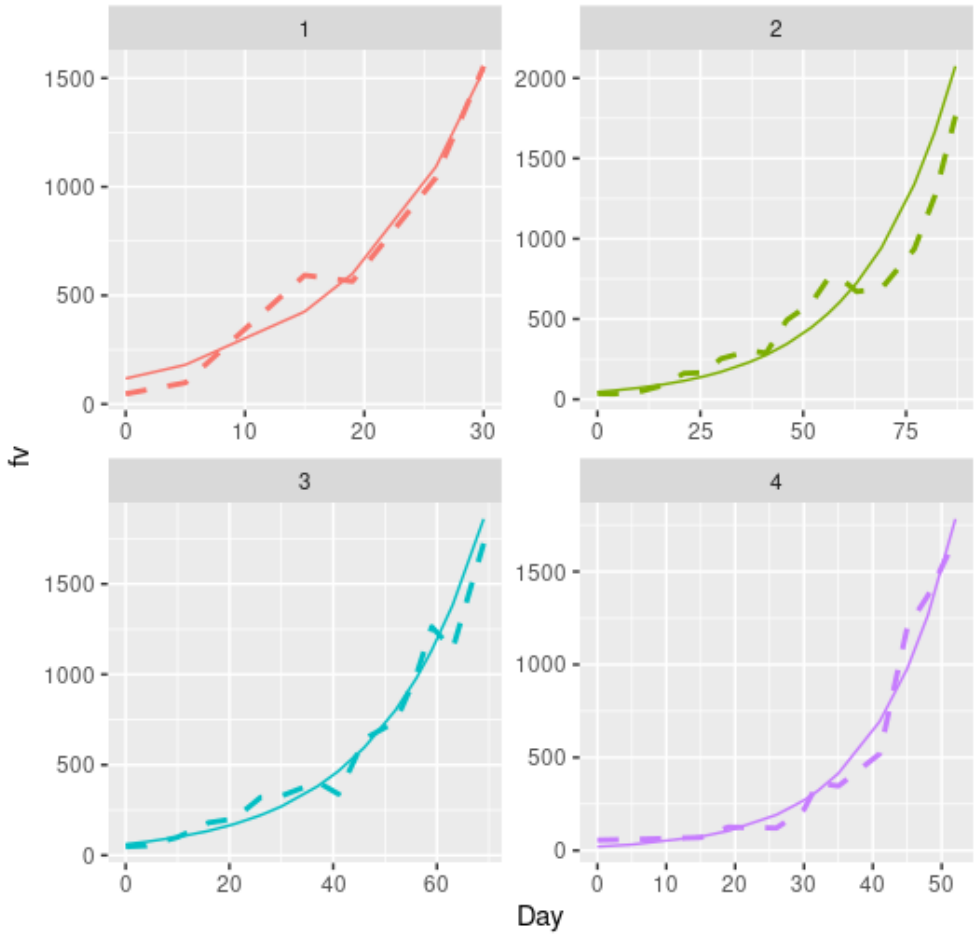
Non-vwp data

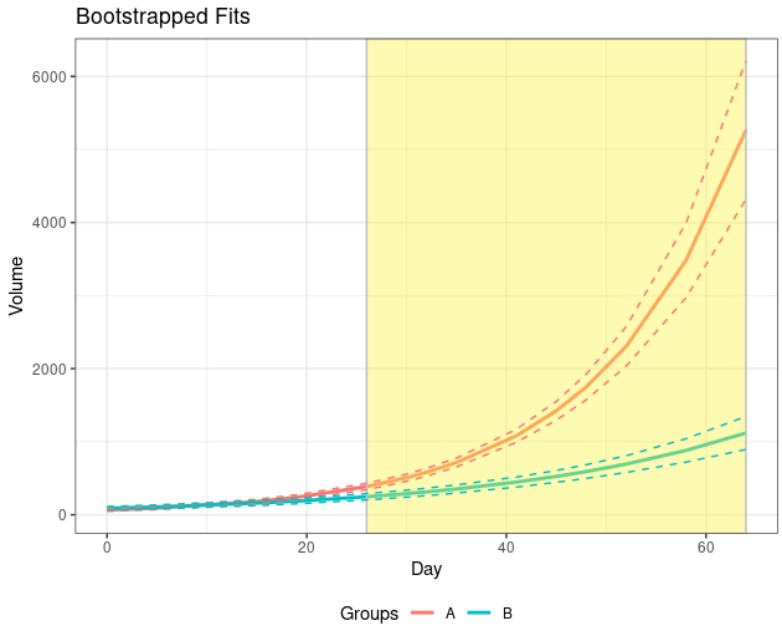
Data for 451LuBR cell line (metastatic melanoma) growth with repeated measures in mice with five treatment groups

```
## Using custom curve for fitting data
fit <- bdotsFit(data = dat,
                subject = "ID",
                time = "Day",
                y = "Volume",
                group = "Treatment",
                curveType = expCurve())
```

Plots

Representative curves for individual mice, as well as comparisons between two treatments with bootstrapped curves





Future work

Making package more robost to different types of data

Handling inconsistencies in time of observations

Investigate different optimization methods for improving fitted curves

Convenience functions

Visual World Paradigm

Broadly speaking, the visual world paradigm (vwp) is a paradigm in which subjects are placed in a "visual world" in which they are prompted to select an item in response to spoken language

Eyetracking software collects location of eye movement in real time as it responds to spoken language

"An increasingly popular approach to visual world data is to fit some nonlinear function of time to visualizations of the data...as descriptors of how the trajectories change over time" (Oleson, 2017)

This leads to the idea of using eyetracking data to construct a *fixation curve*, a (typically) parametric function fit to the data collected

Eyetracking data

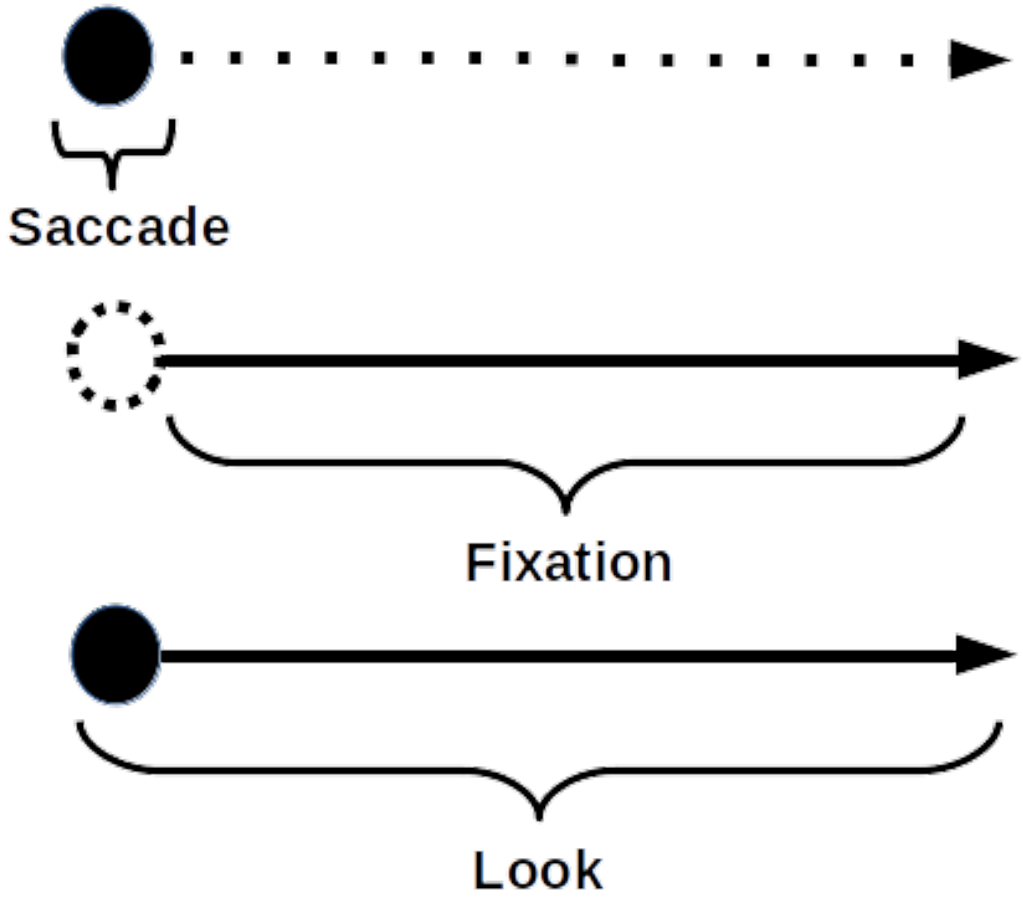
Two types of events make up eyetracking data: *saccades* and *fixations*

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*

Saccade, Fixations, and Looks (oh my!)



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

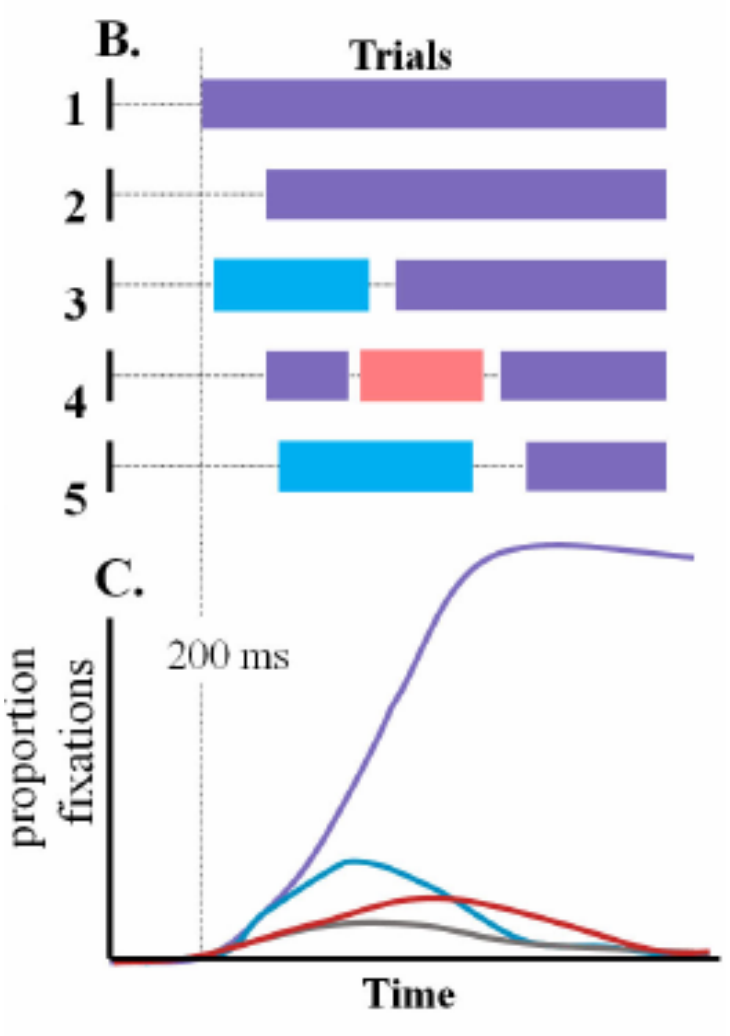
For example, with the four-parameter logistic, our target is typically the Target, while for the six-parameter double-gauss, our target is the Cohort

Aggregation of Data

Subjects are measured across a series of trials in which they are asked to identify the Target

Fixations are captured in real time and aggregated across trials, with each time point representing the proportions of trials in which a subject was fixed on any item

The resulting proportion of fixations then serves as the empirical observation of the *fixation curve*



source: Princess Bride paper

Mathematical Expression of Aggregate Data

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.

Here, the vector y_i serves as a direct observation of $f_\theta(t)$ for subject i

VWP and looks

Of critical importance here are the underlying assumptions relating cognitive activation of an object and the resulting fixations, referred to as a grounding hypothesis (Magnuson)

We can begin by assuming that there is some underlying function dictating fixations, though how this is mediated with observed data is still up for debate

Bob explored a number of these assumptions in his Princess Bride paper, and here we will focus on two: high frequency sampling and fixation-based sampling, augmented for target

Sampling Paradigms

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, allows fixations to target to be slightly longer (once fixated, subject more likely to stay)

Why do we care?

The differences in sampling assumptions raises questions as to biases introduced in our estimate of the fixation curve.

- Can each observed time point be considered an independent draw from a fixation curve?
- How do the lengths of each fixation impact potential bias?
- Does the duration of these fixations change over time and in response to previously identified items?

With recovery of the underlying fixation curve being our goal, we should be able to recover the underlying curve from observed data according to a particular hypothesis

High Frequency Sampling

On the positive side, simulations run with the HFS assumption were able to correctly recover the underlying fixation curve

As Bob has noted, however, the HFS assumption is "patently untrue"

While the bdots package was originally created as a means of modeling the data to account for autocorrelation, it is unable to take into consideration the dynamics of fixations

We will then limit our attention to FBS+T in consideration of potential biases introduced by the mechanics of eye movement

High Frequency Sampling

On the positive sic nption were able to correctly recove

As Bob has noted, ently untrue"

While the bdots p ns of modeling the data to accour ake into consideration the

We will then limit ion of potential biases introduced

FBS+T

Fixation based sampling introduces a more realistic situation in which looks to a particular target are initiated at some point, but then remain fixated for a random period of time

Empirically, fixations on the Target tend to last longer than others, adding an additional mechanic to the generation of data

Finally, there is the adjustment for oculomotor delay; when a saccade occurs at 1200ms, it is likely that it was planned around 1000ms

We will briefly review how this simulation is conducted

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 from 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
}
```

fbst plot

plot with only aggregate data for fbst

Fixation vs Saccade

Here, we reflect again on the construction of our empirical fixation curve,

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

Critically, we realize that the only sample from the fixation curve that we observe *is at the saccade*.

In other words, if the subject fixates on the target at t for 500ms, we have introduced "observed" data at $t + 500$ from a sample of the fixation curve taken at time t

fix vs saccade plots

First plots showing aggregate vs saccade

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

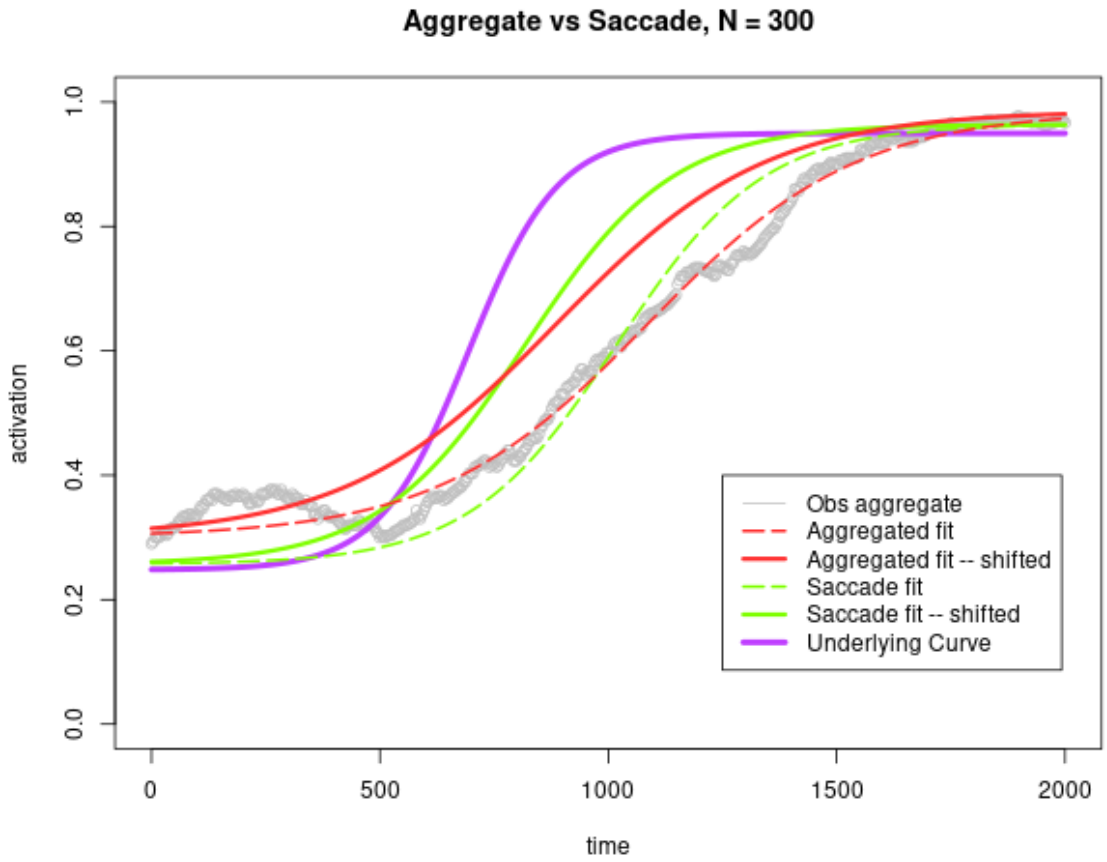
Here, we run simulations assuming fbst and reconstruct fits using bdots

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

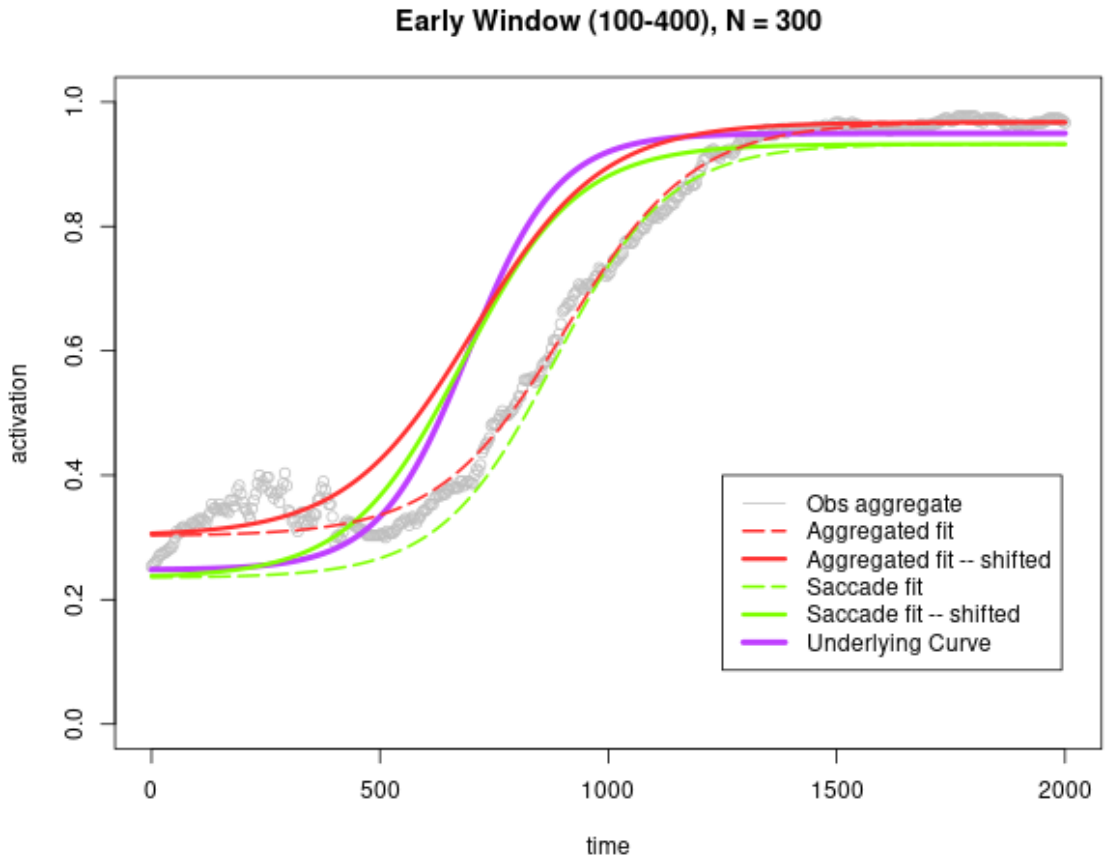
both



MISE

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

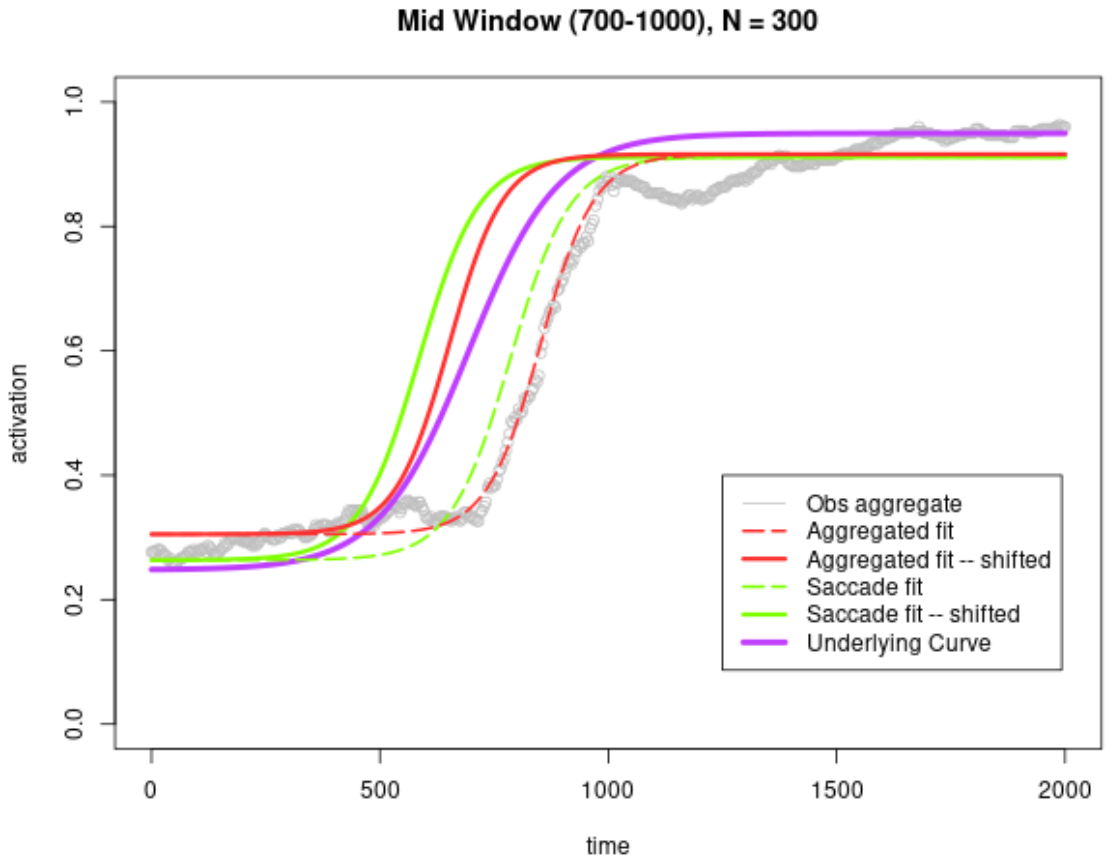
early window



MISE

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

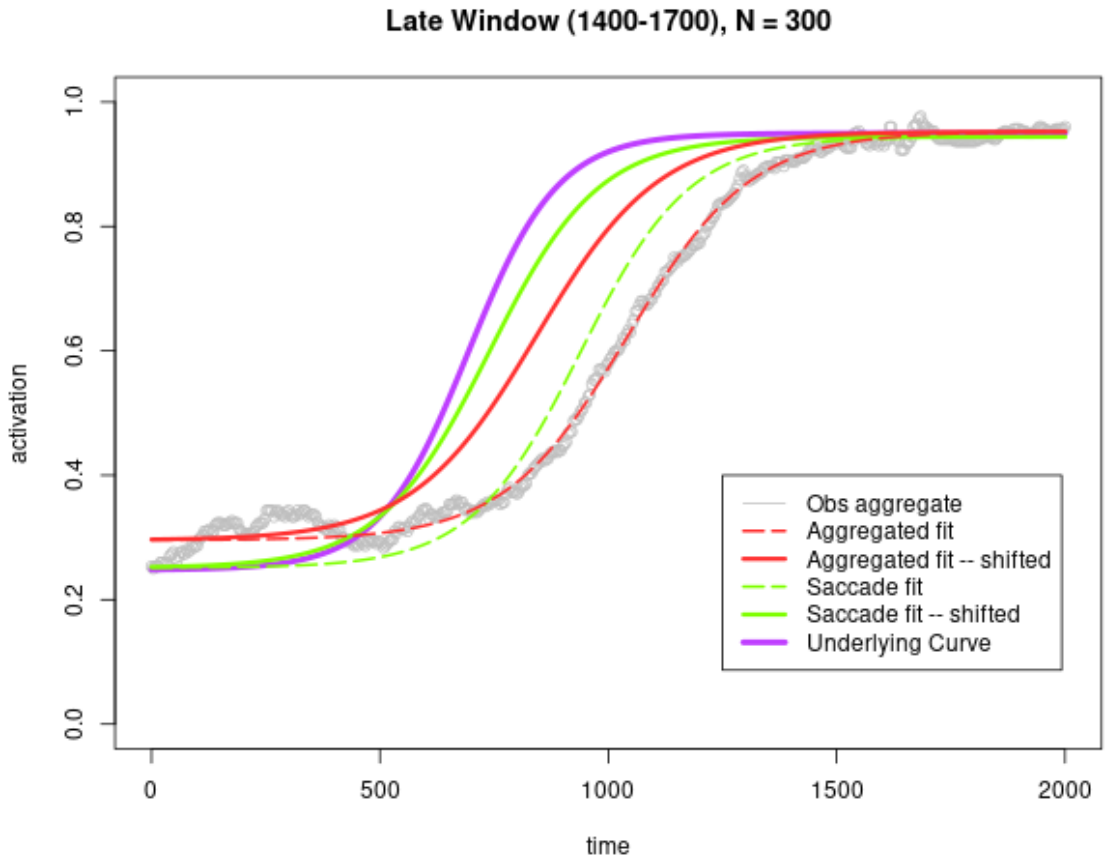
mid window



MISE

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

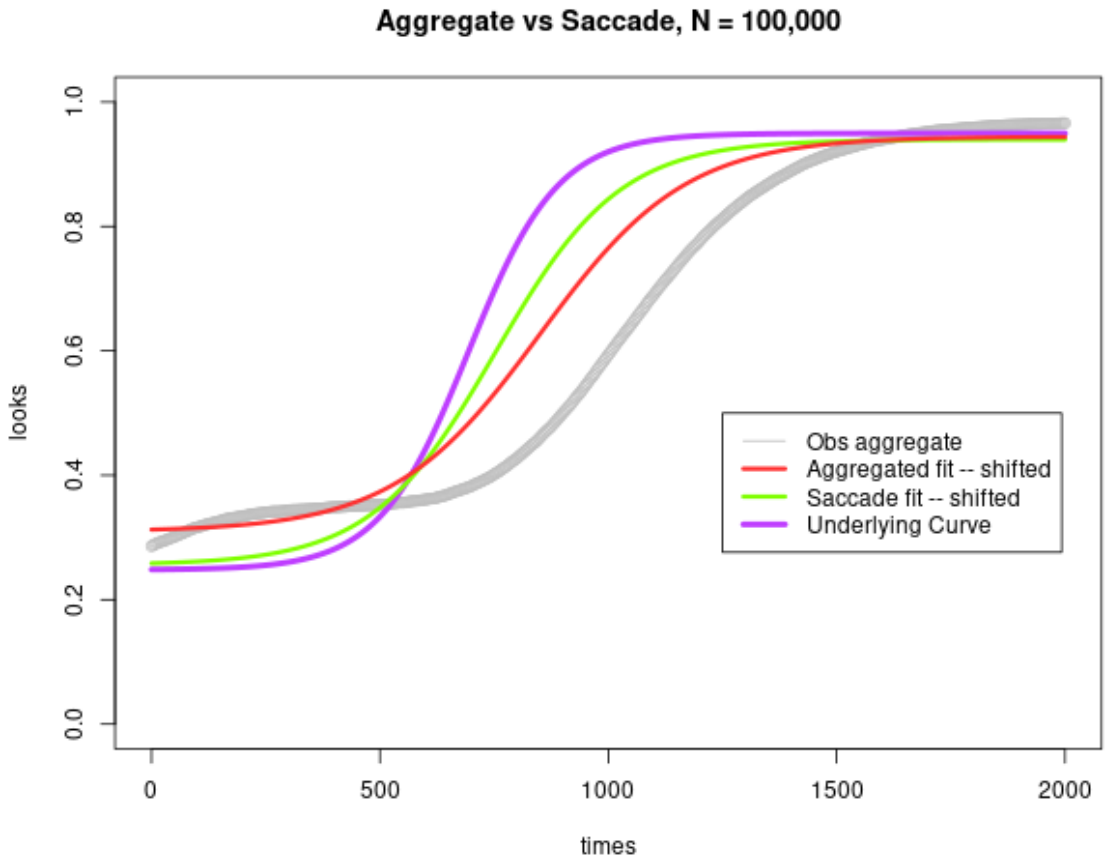
late window



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

Limitations

Not really recovering cognitive curve

Assumes that all trials are drawn from the same functional curve (that is, independent of trial conditions)