Prospectus title

and subtitle!

Collin Nolte

April 29, 2022

Outline

1. bdots
    a. Methodology
    b. Updates
    c. Non-vwp data
2. Visual World Paradigm
    a. Eyetracking data
    b. Mathematical description
    c. Sampling and curves
3. Simulations
    a. Saccades
    b. Oculomotor delay

bdots

The idea behind bdots (bootstrapped difference in time series) was originally proposed by Oleson, Cavanaugh, McMurray and Brown (2017)

First packaged version for CRAN written by Michael Seedorff, with subsequent updates made by Brad Loeffler

bdots

Current implementation of `bdots` involves two steps:

1. **Curve Fitting:** Fitting parametric curve to observed data
2. **Bootstrap** Bootstrap curves to estimate group population curve

Additionally, there functions available to assist with refitting poorly fit curves, either with a manual refitting or batch uploading of new starting parameters

Updates

Fitting process has been simplified 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` are of class `bdObj`, inheriting from `data.table` class

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

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))
```

Output (old)

```
> summary(fit0)
             Length Class      Mode
data           7    data.table list
col            1    -none-     numeric
rho.0          1    -none-     numeric
N.time         1    -none-     numeric
N.sub1         1    -none-     numeric
N.sub2         1    -none-     numeric
coef.id1     150    -none-     numeric
coef.id2     150    -none-     numeric
sdev.id1     150    -none-     numeric
sdev.id2     150    -none-     numeric
sigma.id1     25    -none-     numeric
sigma.id2     25    -none-     numeric
coef.id3     150    -none-     numeric
coef.id4     150    -none-     numeric
sdev.id3     150    -none-     numeric
sdev.id4     150    -none-     numeric
sigma.id3     25    -none-     numeric
sigma.id4     25    -none-     numeric
id.nums.g1    25    factor     numeric
id.nums.g2    25    factor     numeric
groups         2    -none-     numeric
time.all     501    -none-     numeric
N.g1           1    -none-     numeric
N.g2           1    -none-     numeric
concave        2    -none-     logical
model          1    -none-     character
R2.g1.1       25    -none-     numeric
R2.g2.1       25    -none-     numeric
R2.g1.2       25    -none-     numeric
R2.g2.2       25    -none-     numeric
diffs          1    -none-     logical
```

Output (new)

```
> head(fit, n = 15)
    Subject Group LookType        fit      R2   AR1 fitCode
 1:       1    50   Cohort <gnls[18]> 0.96972  TRUE       0
 2:       1    65   Cohort <gnls[18]> 0.98049  TRUE       0
 3:       2    50   Cohort <gnls[18]> 0.98117  TRUE       0
 4:       2    65   Cohort <gnls[18]> 0.96975  TRUE       0
 5:       3    50   Cohort <gnls[18]> 0.97619  TRUE       0
 6:       3    65   Cohort <gnls[18]> 0.95349 FALSE       3
 7:       4    50   Cohort <gnls[18]> 0.97079  TRUE       0
 8:       4    65   Cohort <gnls[18]> 0.64374 FALSE       5
 9:       5    50   Cohort <gnls[18]> 0.97876  TRUE       0
10:       5    65   Cohort <gnls[18]> 0.97656  TRUE       0
11:       6    50   Cohort <gnls[18]> 0.93516  TRUE       1
12:       6    65   Cohort <gnls[18]> 0.92825  TRUE       1
13:       7    50   Cohort <gnls[18]> 0.84164  TRUE       1
14:       7    65   Cohort <gnls[18]> 0.93777  TRUE       1
15:       8    50   Cohort <gnls[18]> 0.98621  TRUE       0
```

Bootstrap with bdots

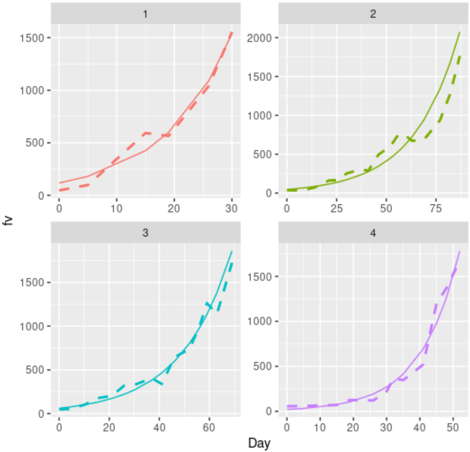
```r
## Old bdots
boot0 <- doubleGauss.boot(
  part1.list = fit0,
  paired = TRUE) # Must indicate if observations paired

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

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

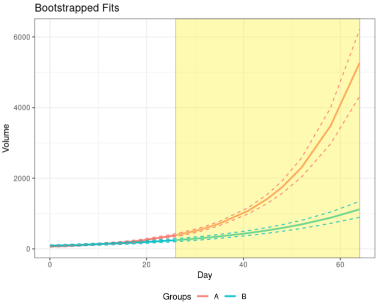
Non-vwp data

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

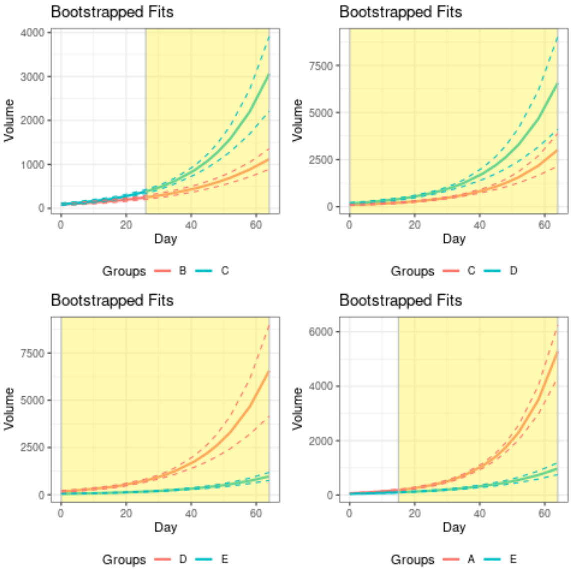
```r
## Using custom curve for fitting data
fit <- bdotsFit(data = dat,
                subject = "ID",
                time = "Day",
                y = "Volume",
                group = "Treatment",
                curveType = expCurve())
```

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





Other comparisons



Future work

Making package more robust 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)

Eyetracking data

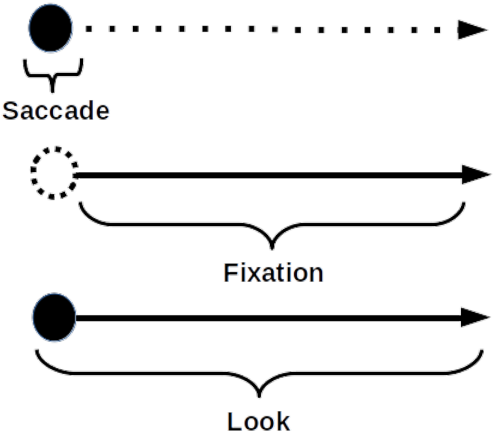
Two types of events make up eyetracking data: *saccades* and *fixations*

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*

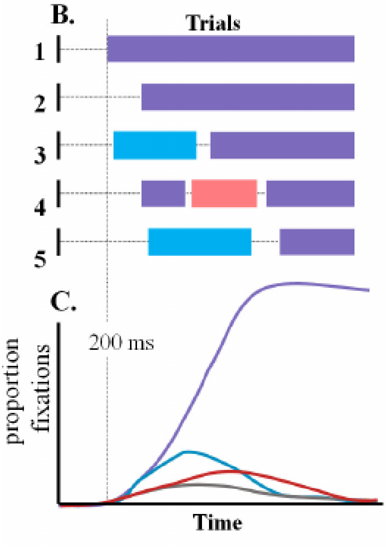
Saccade, Fixations, and Looks (oh my!)



Aggregation of Data

Fixations are captured in real time and aggregated across trials

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

What's in a name?

Briefly, we need to address the fact that there are a few distinct but similar "curves" in question here

1. "Activation" curve, latent mental process
2. "Empirical fixation" curve, single trial or aggregate
3. Idealized "looking" curve, data generating mechanism

Still unsettled

VWP and looks

Underlying assumptions relating cognitive activation of an object and the resulting fixations, referred to as a grounding hypothesis (Magnuson 2019)

We 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 (*Fixation Curves in the Visual World Paradigm*, 2022?), and here we will focus on two:

- High frequency sampling (HFS)
- Fixation-based sampling, augmented for target (FBS+T)

Sampling Paradigms

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

FBS+T: Fixation-based sampling + target, series of discrete fixations with reasonable refractory period

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"

bdots to account for autocorrelation

Limit our attention to FBS+T in consideration of potential biases introduced by the mechanics of eye movement

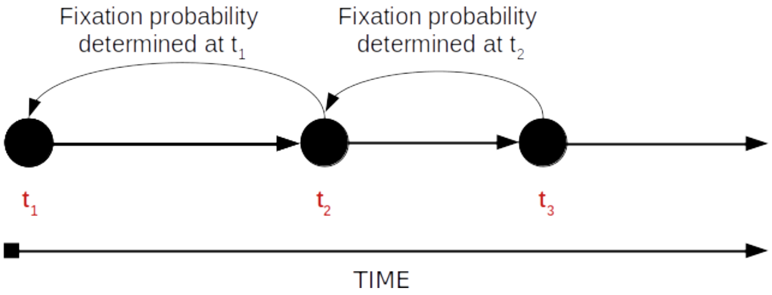
FBS+T

Fixation based sampling introduces a more realistic situation

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

FBS+T


Fixation probability determined at t_1
Fixation probability determined at t_2
t_1
t_2
t_3
TIME

Simulations

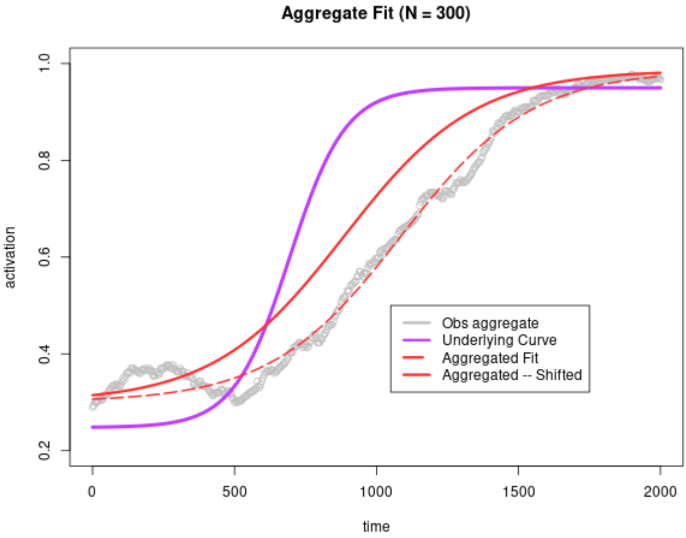
Simulations were run with a single subject with $N = 300$ trials

Following FBS+T to generate eyetracking data

Plots include

- data generating curve
- observed data from simulation
- estimated bdots curve, with shift

FBS+T plot



Aggregate Fit (N = 300)
activation
time
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
Obs aggregate
Underlying Curve
Aggregated Fit
Aggregated -- Shifted

Fixation vs Saccade

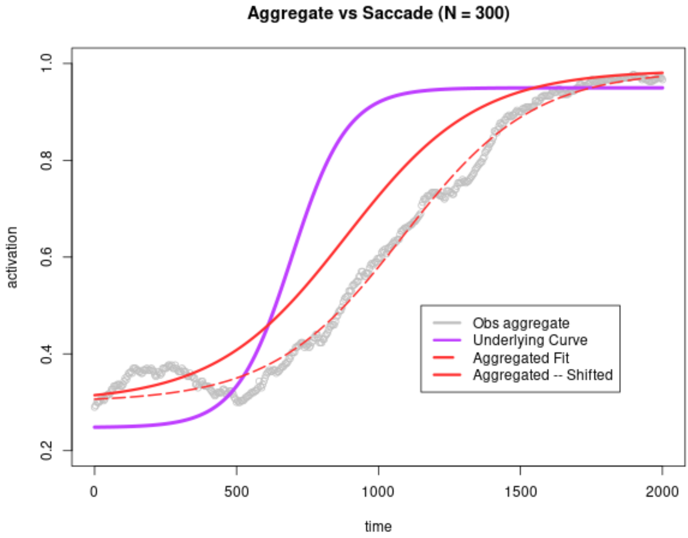
Here, we reflect again on the construction of our empirical fixation curve,

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

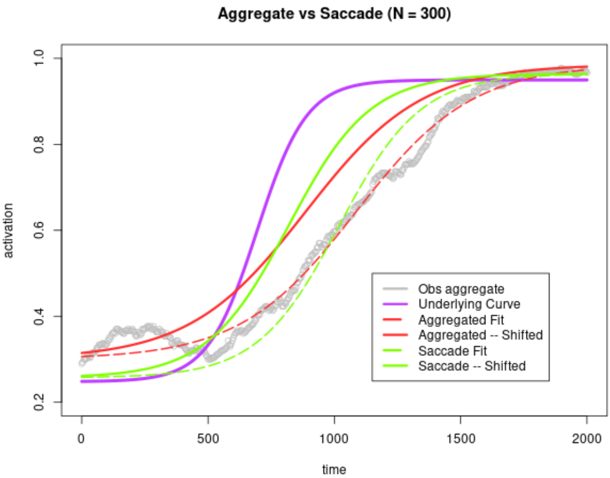
Critically, we realize that the only sample from the fixation curve that we observe *is at the saccade*.

In other words, if the subject fixates on the target at t for 500ms, we have introduced “observed” data through $(t+1, t+500)$ from a sample of the fixation curve taken at time t

Aggregate vs Saccade

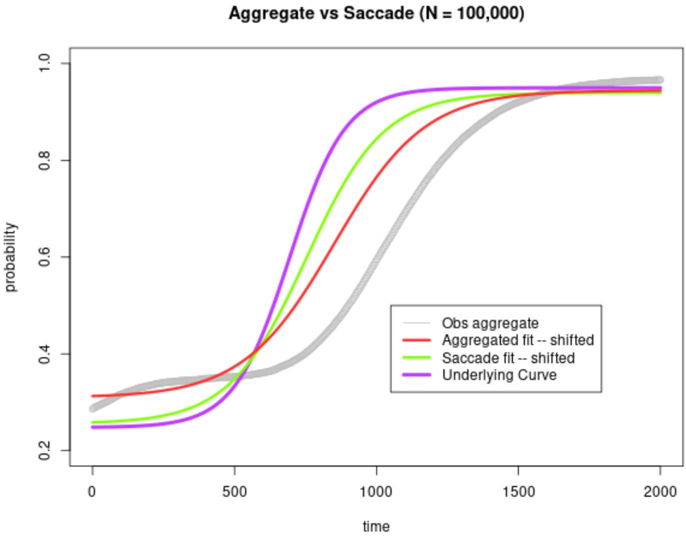


Aggregate vs Saccade



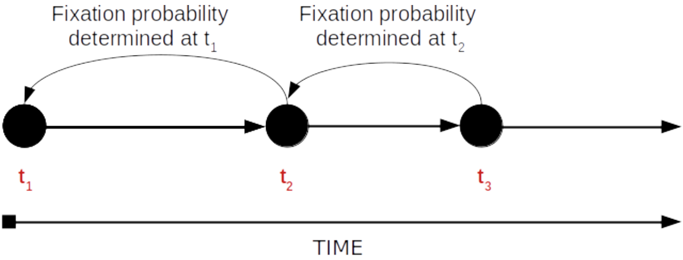
	Aggregate	Saccade	Aggregate – Shifted	Saccade – Shifted	Underlying
MISE	21.96	19.34	5.04	1.77	0.00

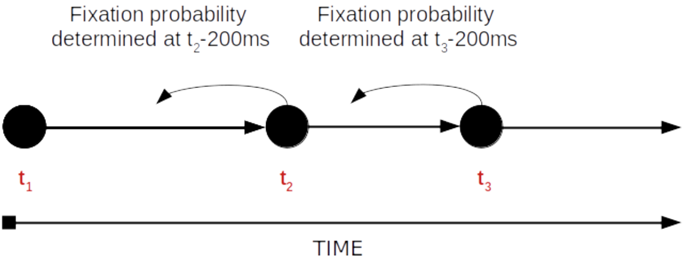
Asymptotic



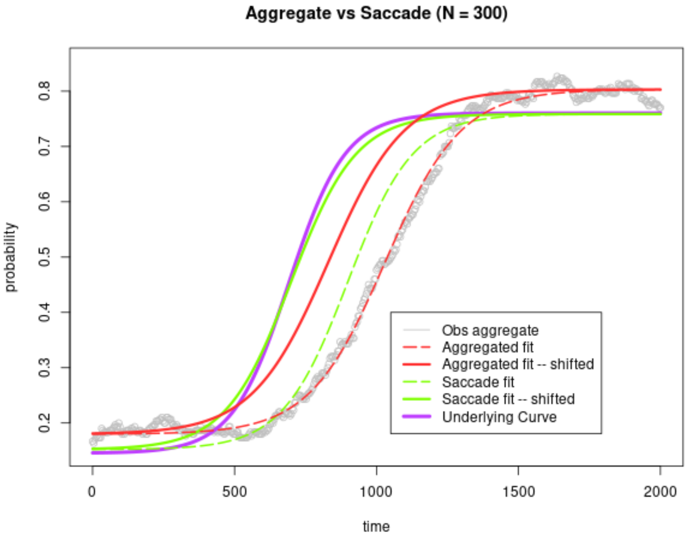
Aggregate vs Saccade (N = 100,000)
probability
1.0
0.8
0.6
0.4
0.2
time
0
500
1000
1500
2000
Obs aggregate
Aggregated fit -- shifted
Saccade fit -- shifted
Underlying Curve

Adjusting Occulomotor Delay



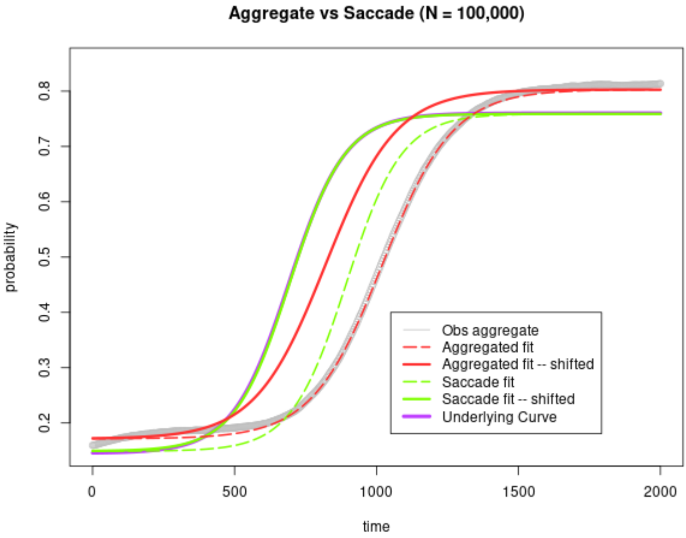


200ms delay



Aggregate vs Saccade (N = 300)
probability
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
500
1000
1500
2000
time
Obs aggregate
Aggregated fit
Aggregated fit -- shifted
Saccade fit
Saccade fit -- shifted
Underlying Curve

200ms delay



Aggregate vs Saccade (N = 100,000)
probability
0.2
0.3
0.4
0.5
0.6
0.7
0.8
time
0
500
1000
1500
2000
Obs aggregate
Aggregated fit
Aggregated fit -- shifted
Saccade fit
Saccade fit -- shifted
Underlying Curve

Future Directions?

Revist fitting/refitting process to make more robust to bad fits

Investigate bootstrap confidence intervals for correct coverage

Consideration of nonparametric fits for data

Window/variability theory

"Information gathering behavior"

References

Magnuson, James S. **Fixations in the visual world paradigm: where, when, why?** 2019-09 *Journal of Cultural Cognitive Science*, Vol. 3, No. 2 Springer Science and Business Media LLC p. 113-139

McMurray, Bob **Fixation Curves in the Visual World Paradigm** 2020(?)

Oleson, Jacob J; Cavanaugh, Joseph E, McMurray, Bob; Brown, Grant **Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm** 2017 *Statistical Methods in Medical Research*, Vol. 26, No. 6 p 2708-2725

Fitting Process

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

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

Such that

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

Bootstrapping Process

Here, we perform B bootstraps of the subject parameters to construct bootstrapped curves and confidence intervals. Taking parameter estimates from the fitting step, for each subject 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,

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

Bootstrapping, 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,

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