Commentary on the Analysis of the HumanFixationEvaluation Dataset in the gazeNet Paper:
gazeNet: End-to-end eye-movement event detection with deep neural networks (Zemblys, Niehorster, and Holmqvist, 2019)

Lee Friedman
Department of Computer Science
Texas State University
San Marcos, Texas, USA, 78666
lfriedman10@gmail.com

December 19, 2019

1 Introduction

In the gazeNet paper [1], several automatic classification schemes were evaluated and compared on the on the humanFixationEvaluation dataset made available by [2] [1] This report makes the case that this particular use of this dataset is highly problematic.

2 Statements in gazeNet paper about the humanFixationEvaluation dataset

Let us begin by citing the statements made in the gazeNet paper [1] on the humanFixationEvaluation dataset:

"...humanFixationEvaluation [3] containing data from [4]" (Page 16)

"The humanFixationEvaluation dataset [4, 2] was collected with a Tobii TX300 (300Hz) eye-tracker and derived from infant and adult participants. The total of almost 6 minutes of data were then hand labeled by 12 expert coders, who were asked to label only fixations. Because of the relatively small size of the dataset, we used the labels from all 12 coders in our evaluation." (Page 16)

"In the case of humanFixationEvaluation dataset, the samples that were left uncoded were artificially set to saccade samples and only originally missing samples were considered as missing when evaluating the algorithms." (Page 16)

"In particular this is evident for the humanFixationEvaluation dataset, where both the MNH and the NH2010 algorithms perform approximately twice as bad than IRF and gazeNet." (Page 17)

"Hooge et al. (2017) report that the average noise level in this dataset is 0.32 - 0.36 degrees RMS, which means it has a 10x larger noise than in the lund2013-image-test (see Table 4). Neither MNH nor NH2010 are able to cope with such noise, while both machine learning approaches seem to perform reasonably well (probably because they have seen data with such noise levels during training)." (Page 17-18)

"Moreover, the evaluation of gazeNet on two other datasets, GazeCom and humanFixationClasification, showed that gazeNet is not only able to generalize to unseen data with other qualities, but also agreed more with the hand-coding provided in these datasets than the other event detection algorithms." (Page 19)

---

[1]Made available by the authors at .

3 The humanFixationClasification dataset

In the humanFixationClasification [2], recordings were available from 10 adults and 60 infants. The sampling rate was 300 Hz. On average, 4,500 samples (approx. 15 sec) were recorded for each adult. On average, 1010 samples (approx. 3.33 sec) were recorded from each infant.

The dataset was rated by a number of different raters. All of the analyses presented herein were performed with only a single rater ('MN', Marcus Nyström). Dr. Nyström is a generally recognized expert in the classification of eye movements.

4 The Problem of "blink-saccades"

When a subject blinks, VOG-based eye trackers cannot measure eye position and often return a NaN value. Often, surrounding blink artifact, there are periods of data during which the eye-tracker can detect a signal, even though this periblink signal is not from an eye movement. The signals seen just before and after a blink can appear "saccade-like", and are often referred to as "blink-saccades". But these events are not saccades.

Recall that the gazeNet authors state: "In the case of humanFixationEvaluation dataset, the samples that were left uncoded were artificially set to saccade samples and only originally missing samples were considered as missing when evaluating the algorithms."

The gazeNet paper, according to their own description of their approach, would consider these events, which are either preceded or followed by NaNs, as "saccades". **see Figure 1**.
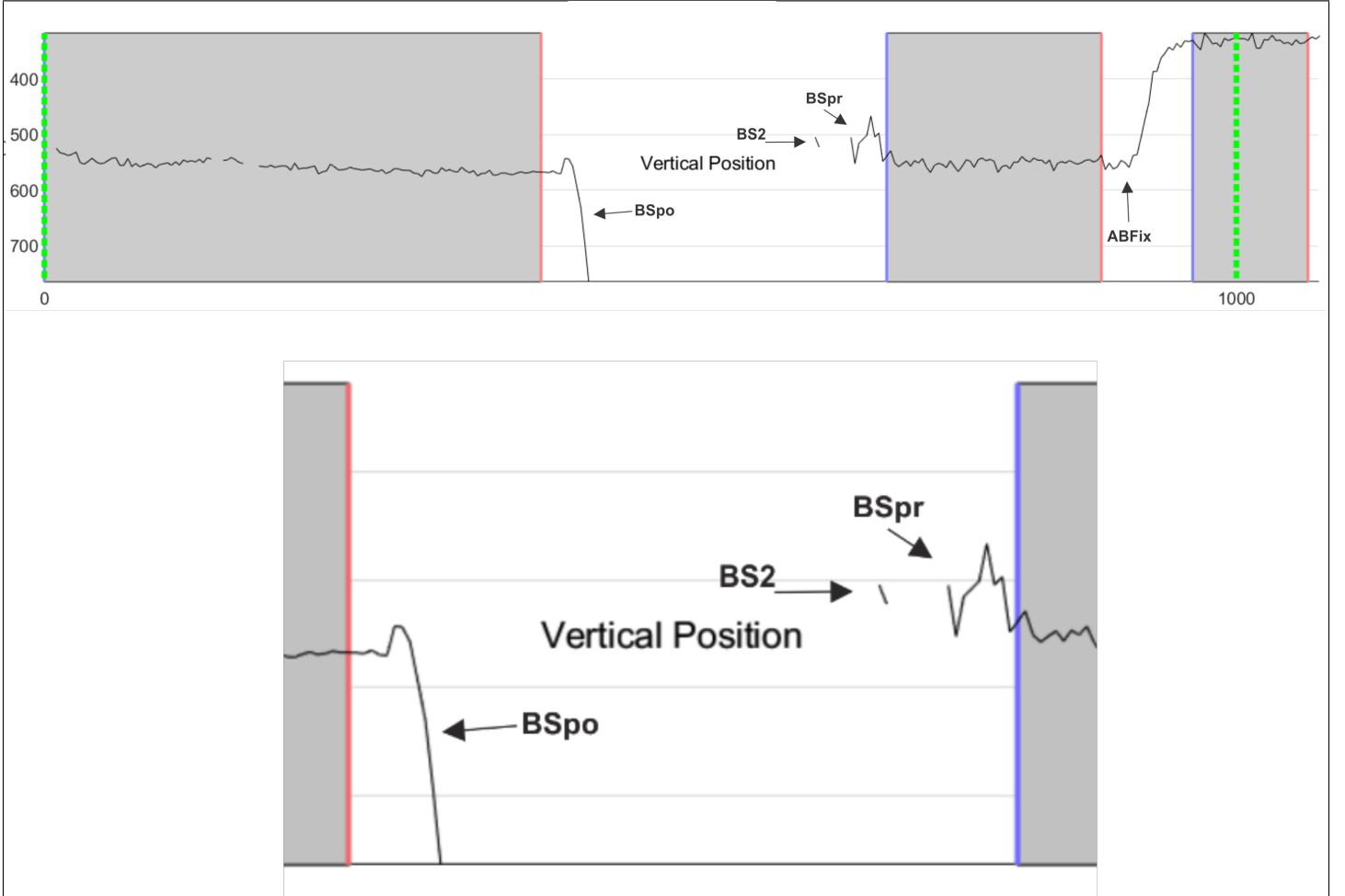

Figure 1: Illustration of the problem of blink-saccades for the gazeNet approach to the humanFixationEvaluation dataset. **TOP:** This is the vertical position signal from the 11th infant study (infant-4592207091392917904). The gray sections are the portions of the data classified as fixation by rater MN. The arrows point to different types of blink-saccades. **BSpo** is a blink-saccade that is followed by NaNs. **BSpre** is a blink-saccade that is preceded by NaNs. **BS2** is a blink-saccade that is both preceded and followed by NaNs. **ABFix** is an event that is both after and before fixation periods. **Bottom:** A magnified version of the top of the figure.

.

Using fixations marked only by rater MN, there are 929 defined "potential saccades" in the humanFixationEvaluation dataset as processed by gazeNet. Of these, 130 were both preceded and followed by NaNs, 106 were followed by NaNs and and 96 were preceded by NaNs. Therefore, there were 332 or 36% of potential saccades that were actually blink-saccades and not saccades.

5 The Problem of Non-Blink-Saccades that are Ill Formed

Of the 597 events that were not blink-saccades, there were a number of potential saccades that had very non-saccade-like trajectories. To detect such events automatically we followed the steps in Algorithm 1[2].

Algorithm 1: Steps in to Identify Poorly Formed Events treated as if they were Saccades by the gazeNet Authors

---

1. Determine the instantaneous velocity $(x_t - x_{t-1})$ of the horizontal and vertical position signals (Vel_x, Vel_y).

2. Compute the radial velocity (Vel_{rad}) as $\sqrt{Vel_x + Vel_y}$.

3. Integrate the radial velocity to estimate a eye-position radial saccade from the data. (**Figure 2B**)

4. Use [5] to estimate the form of a theoretical saccade by equation, using the time and amplitude of the radial saccade. .

5. Zscore transform both the radial saccade and the theoretical saccade.

6. Subtract the radial saccade position values from the theoretical saccade position values and save them as residuals.

7. Sum the absolute value of the residuals and divide by the number of samples to get a Residual-Per-Sample metric for each potential saccade.

8. Sort potential saccades (not blink-saccades) by their Residual-Per-Sample metric.

9. Illustrate some well formed and poorly formed good saccades found in this way (**Figures 2 and 3**).

10. Illustrate 162 of the most poorly formed of these potential saccade events. (**Figures 4-30**).

---

[2]The code for estimating theoretical saccade trajectories used below is at: https://www.mathworks.com/matlabcentral/fileexchange/62880-saccade-model-a-parametric-model-for-saccadic-eye-movement
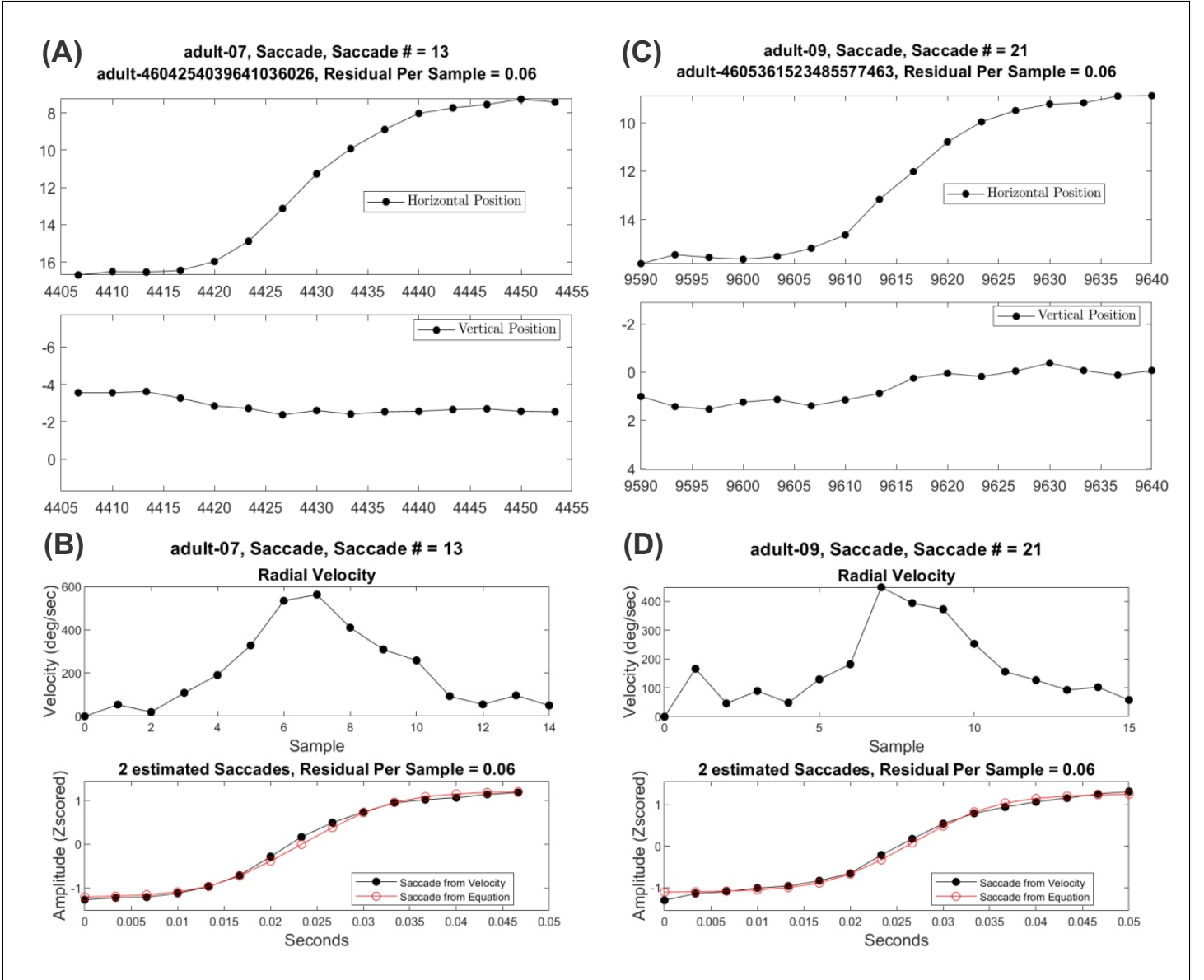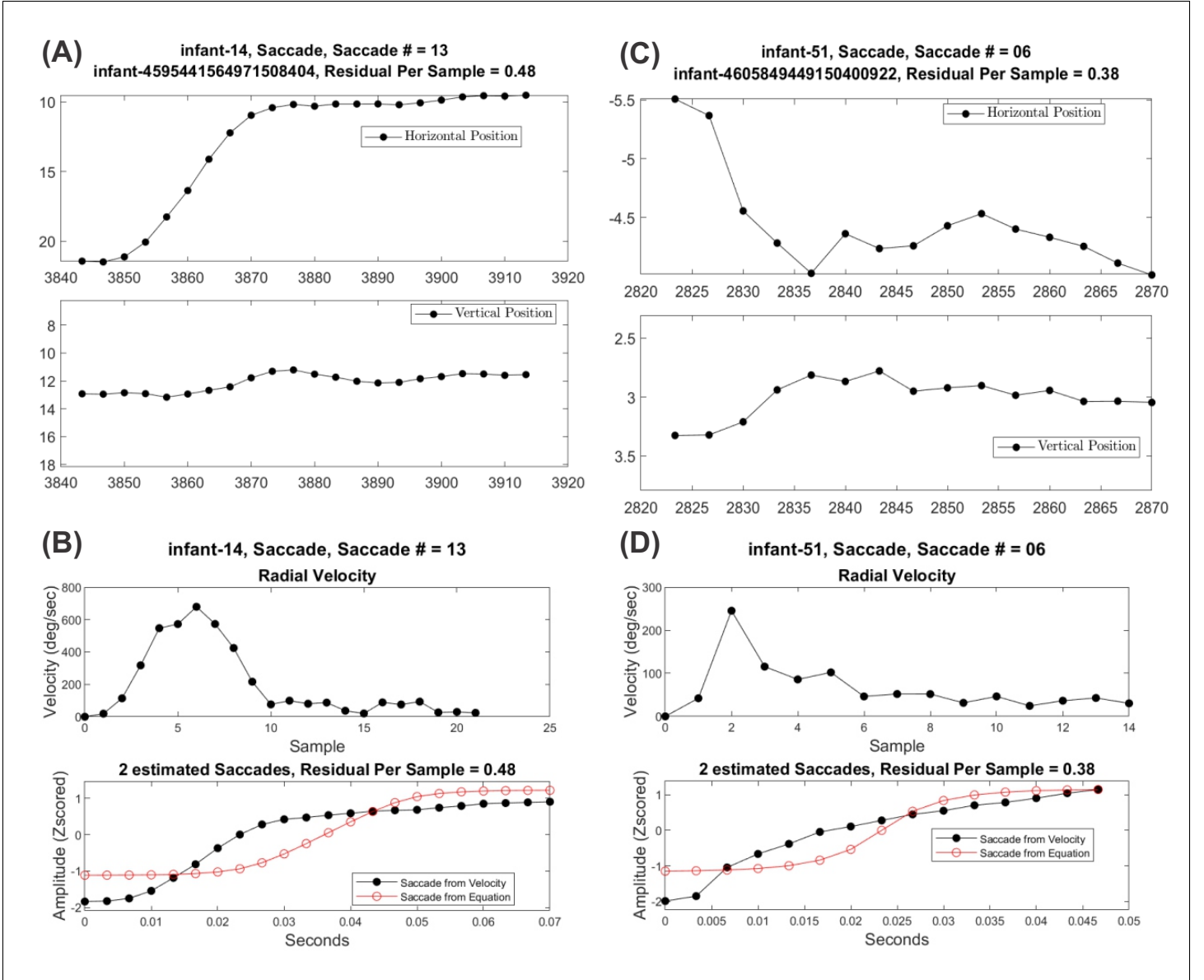
Figure 2: Detection of a Well-Formed Saccades. **(A)** The horizontal position and vertical position of an event considered as a saccade in the gazeNet processing of the humanFixationEvaluation dataset. The saccade is from the 7th adult recording. The low residual per sample (0.06) indicates that this is a well shaped saccade. **(B)** Illustrates steps in the process of determining if the event trajectory is consistent with a saccade (see Algorithm 1). Analysis pertains to saccade shown in **A**. **Upper Panel:** Radial velocity of the saccade. **Lower Panel:** Saccade trajectory constructed from integrating the velocity signal, and a trajectory from the saccade equation provided by [5]. **(C,D)** Saccade from the 9th adult recording. See **(A,B)** for details.

Figure 3: Detection of Poorly-Formed Saccades. **(A)** The horizontal position and vertical position of an event considered as a saccade in the gazeNet processing of the humanFixationEvaluation dataset. The saccade is from the 14th infant recording. The high residual per sample (0.48) indicates that this is a poorly shaped saccade. See caption for Figure **2** for details. **(C,D)** Data from the data from the 51st infant recording.

6   Unusual Potential Saccade Trajectories

What follows are 27 figures **Figures 4-30** illustrating unusual potential saccade trajectories. Each page contains 6 potential saccades. The 162 potential saccades are presented in descending order of the Residual-Per-Sample metric. Saccades with lower Residual-Per-Sample values than 0.24 started to look more acceptable and so the presentation stopped there.

The **Discussion** section follows the presentation of these 162 potential saccades.

Figure 4:

Figure 5:

Figure 6:

Figure 7:
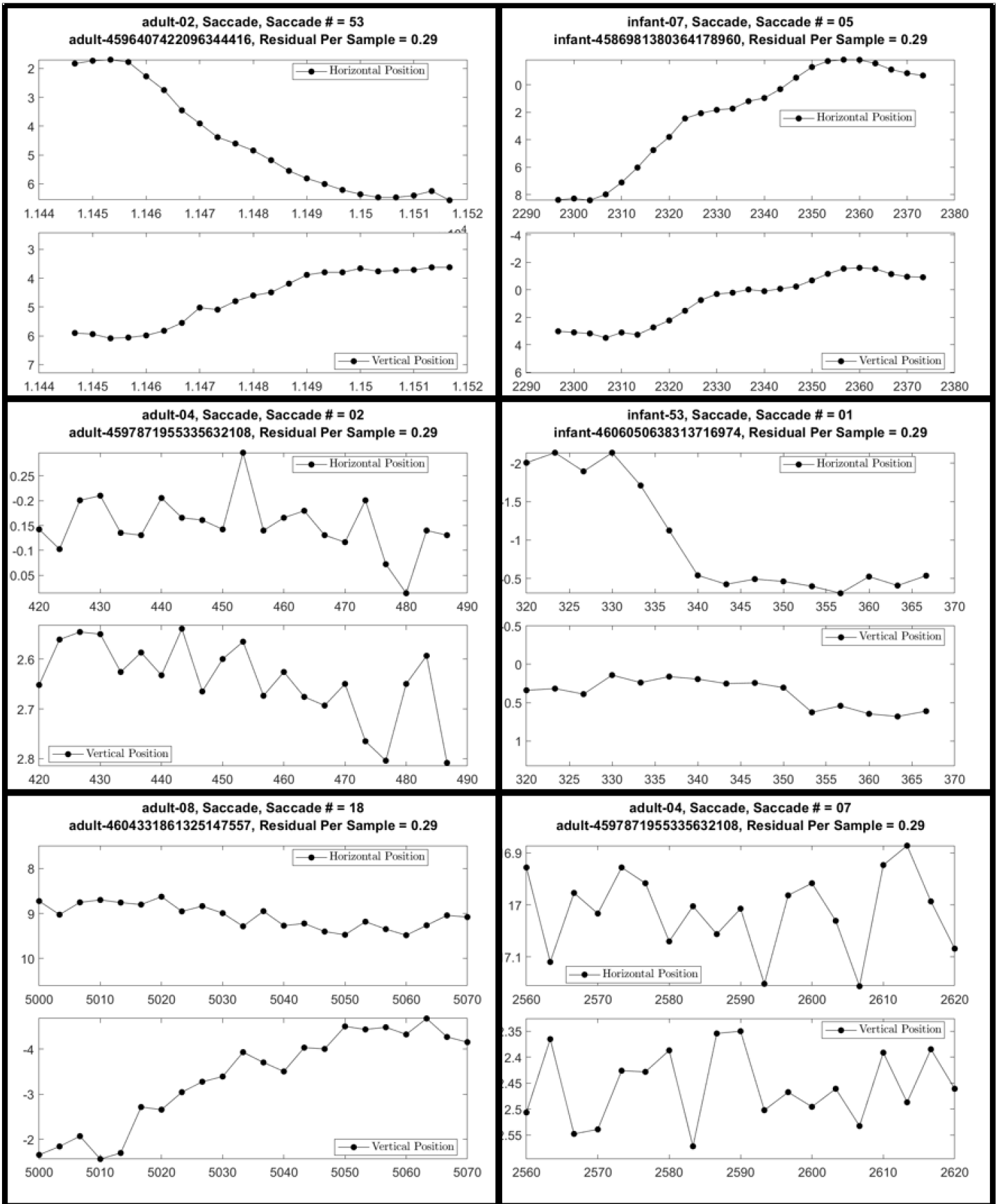
Figure 8:

Figure 9:

Figure 10:
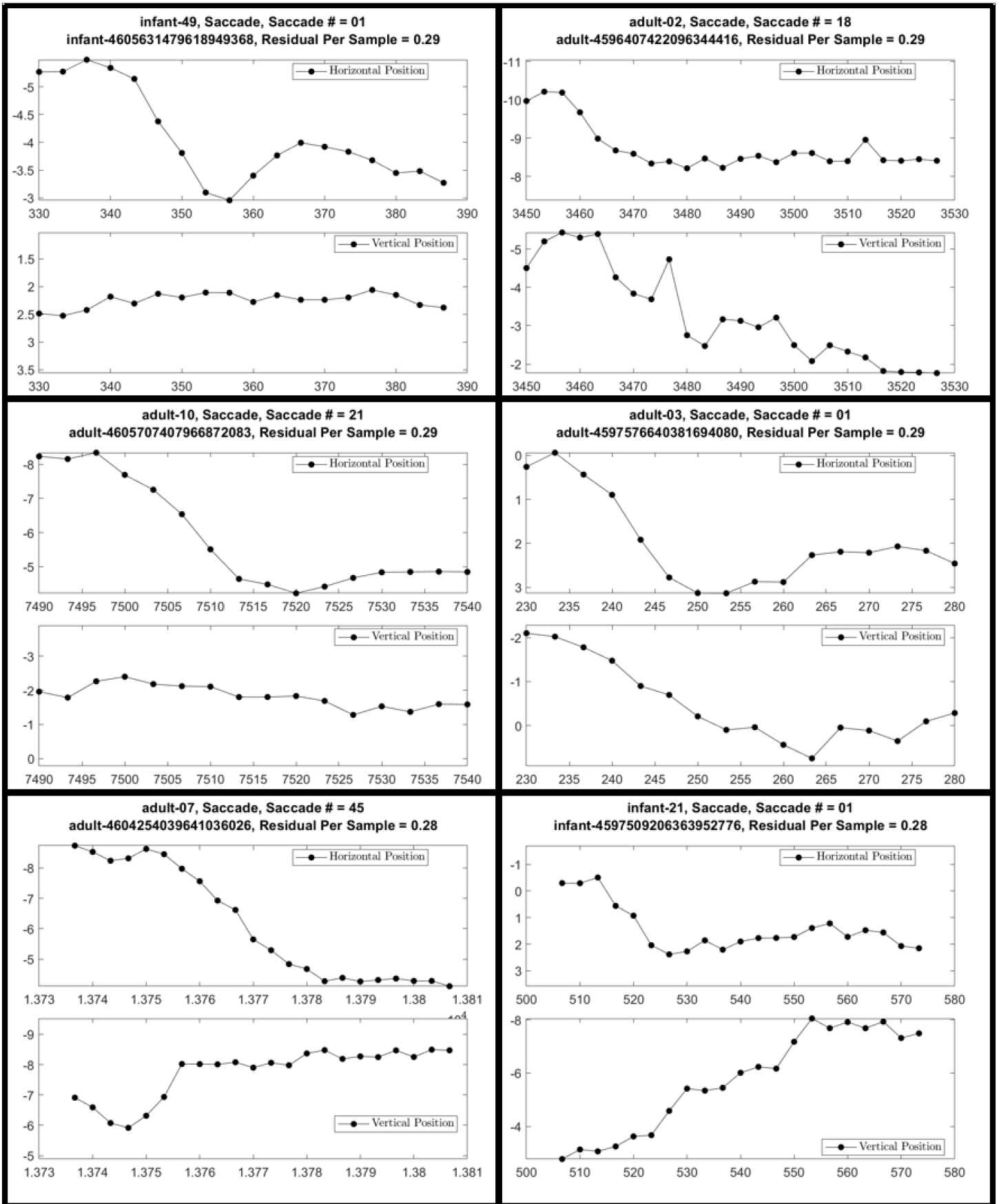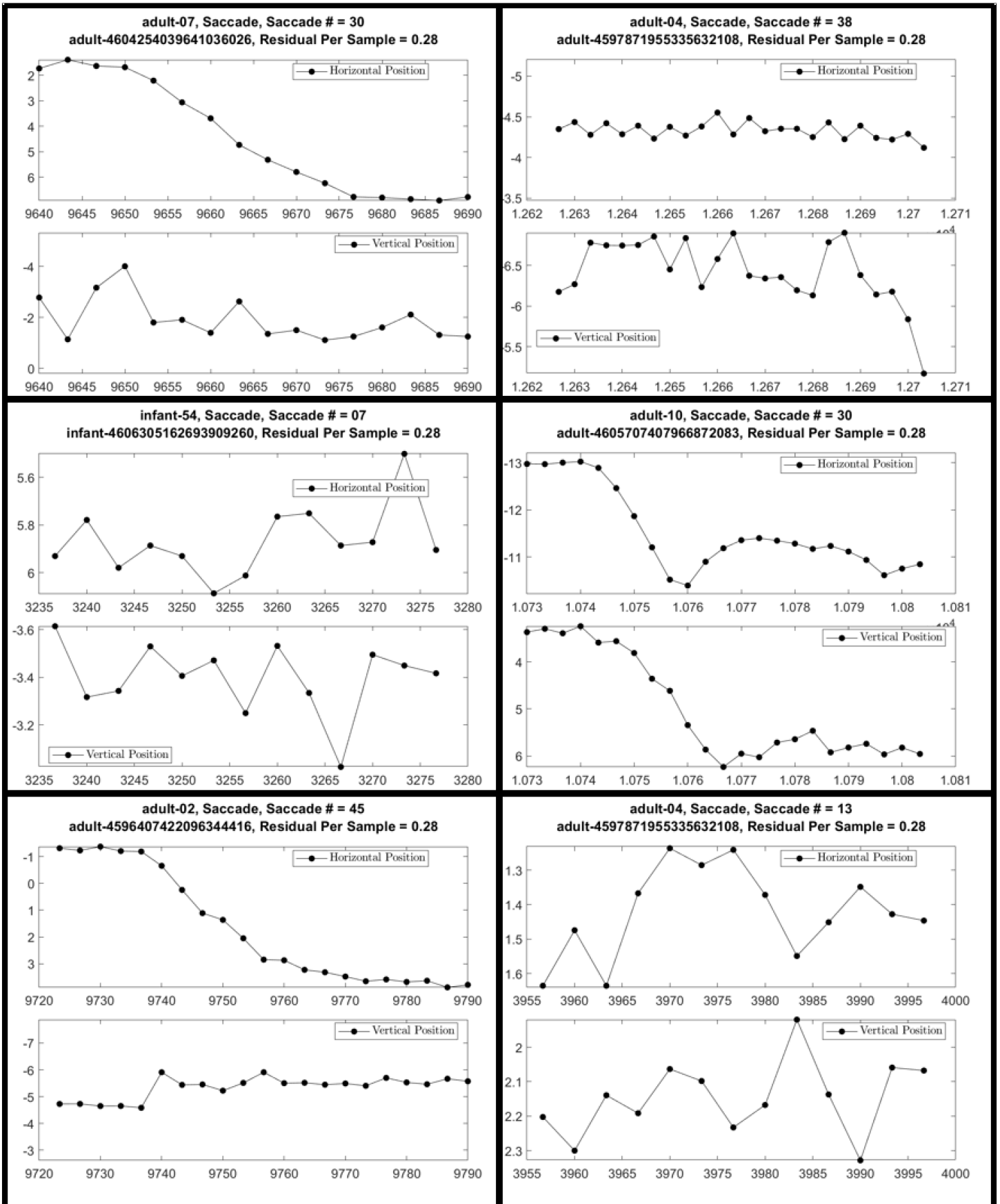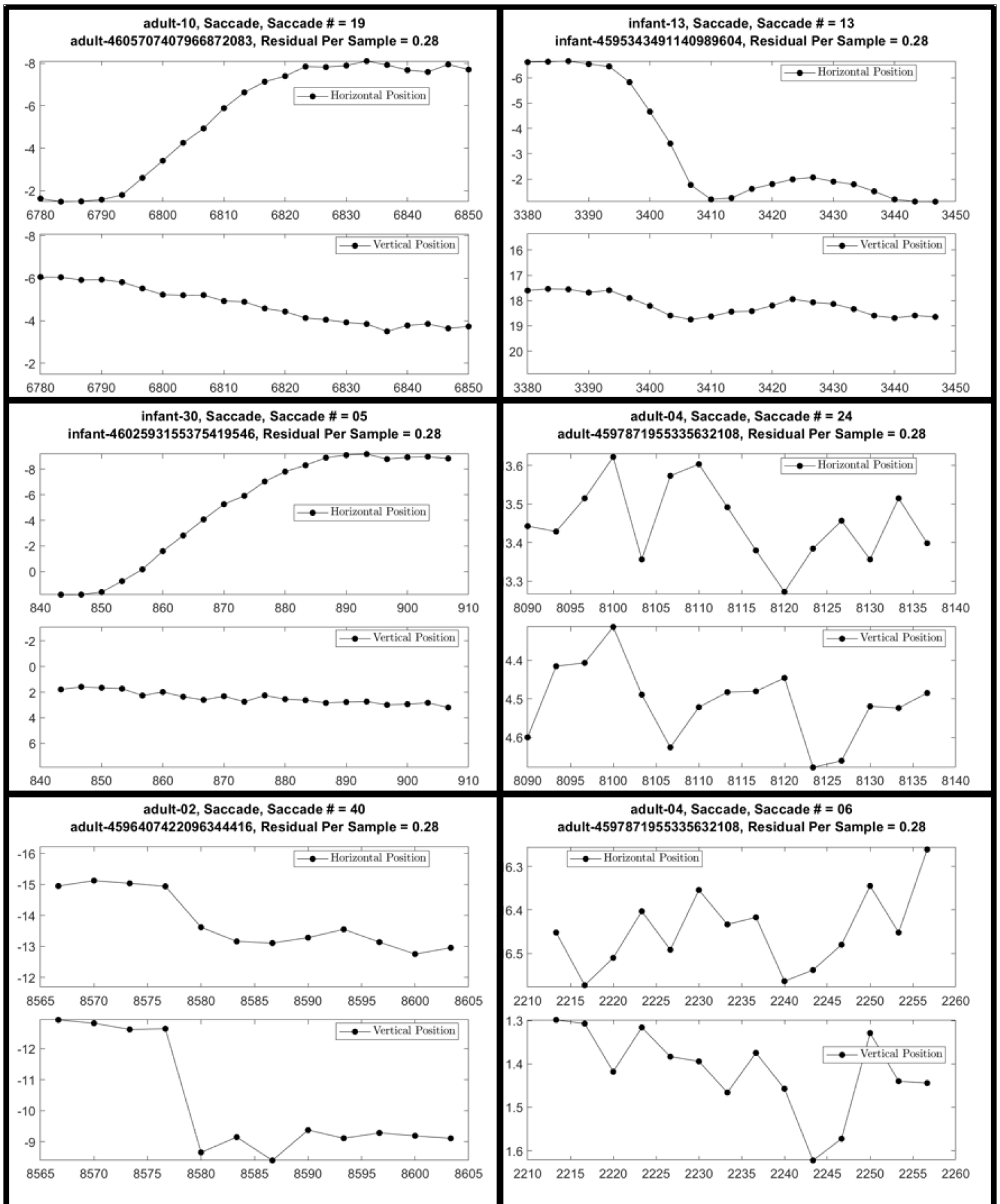
Figure 11:

Figure 12:

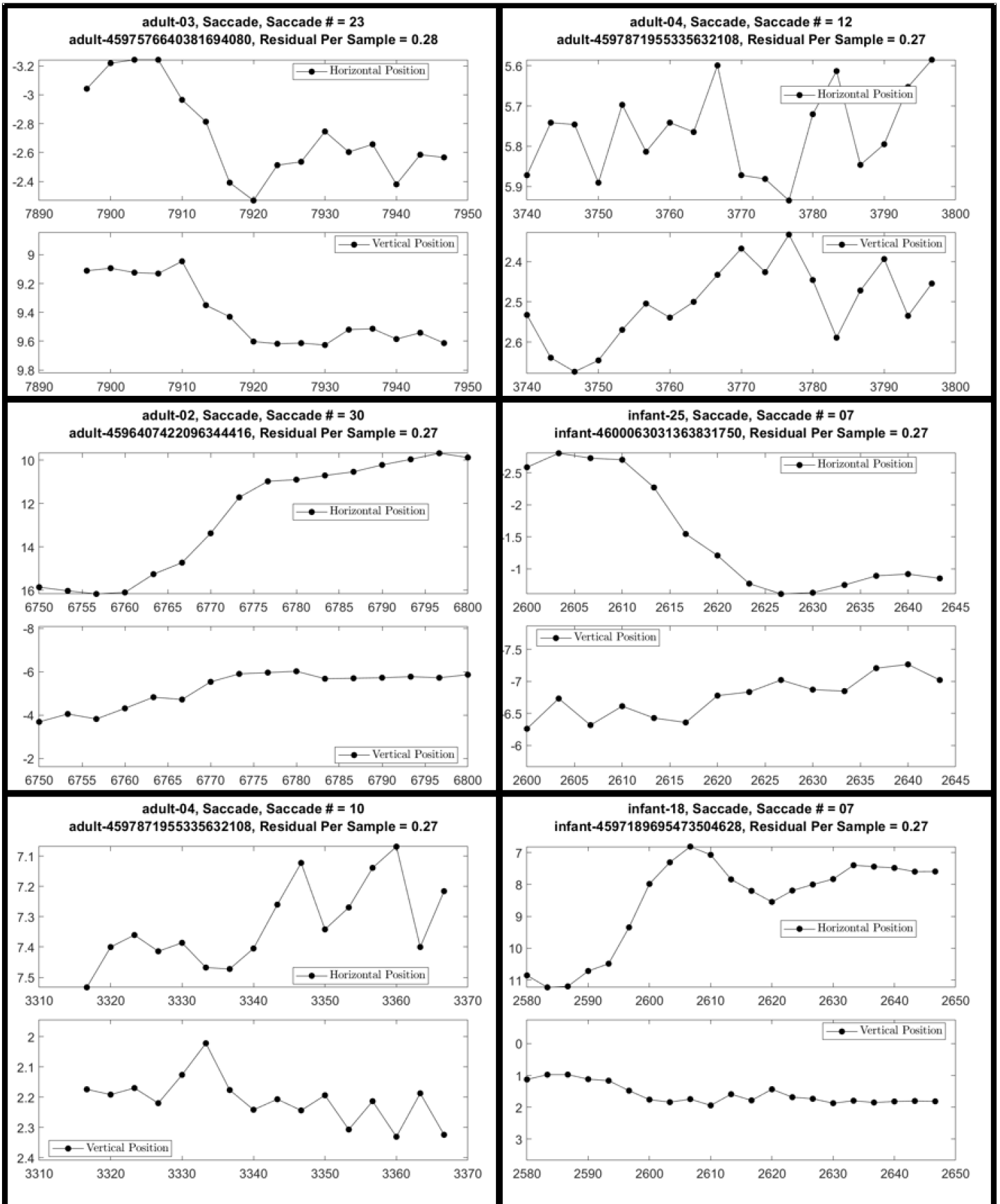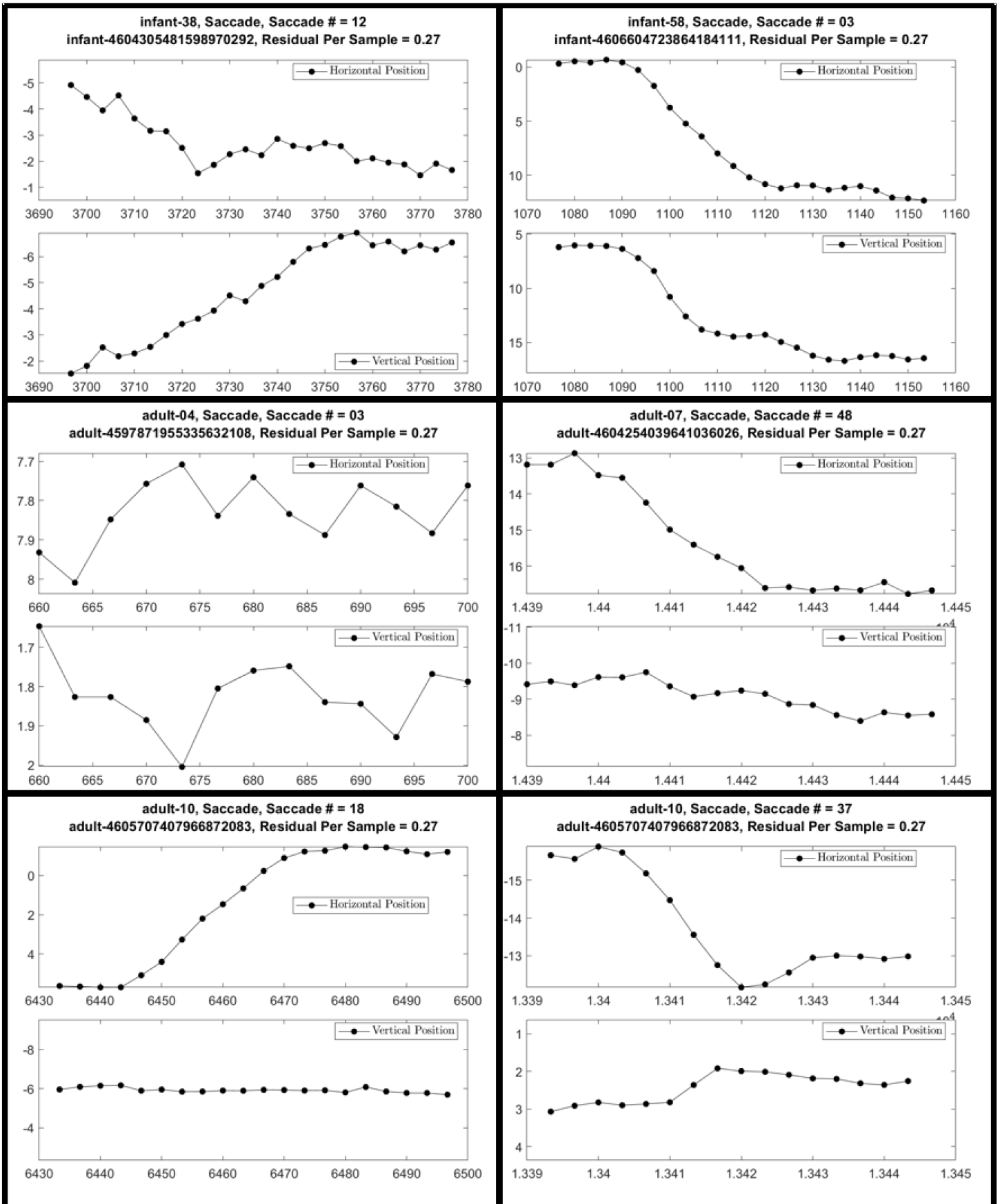Figure 13:

Figure 14:

Figure 15:
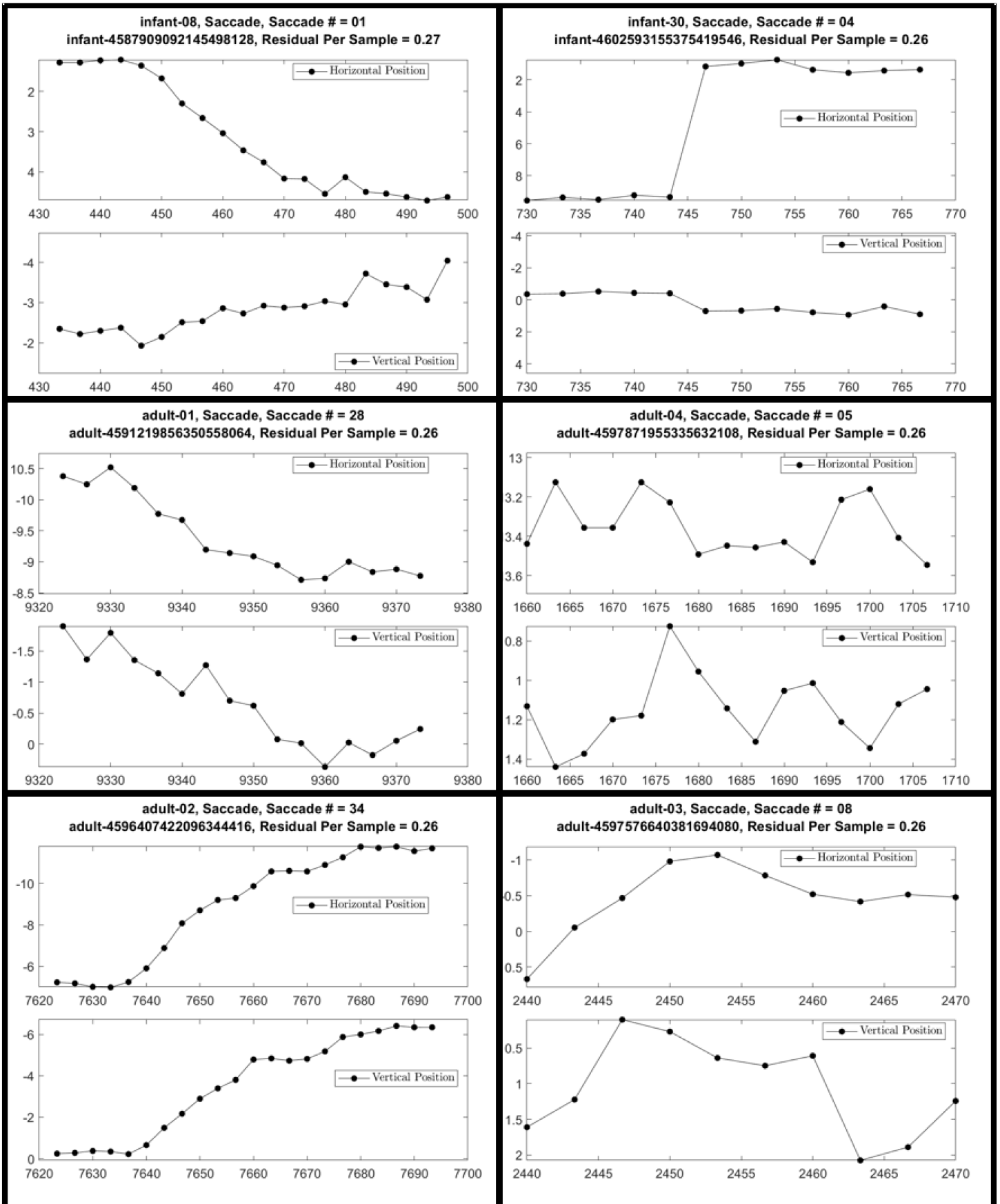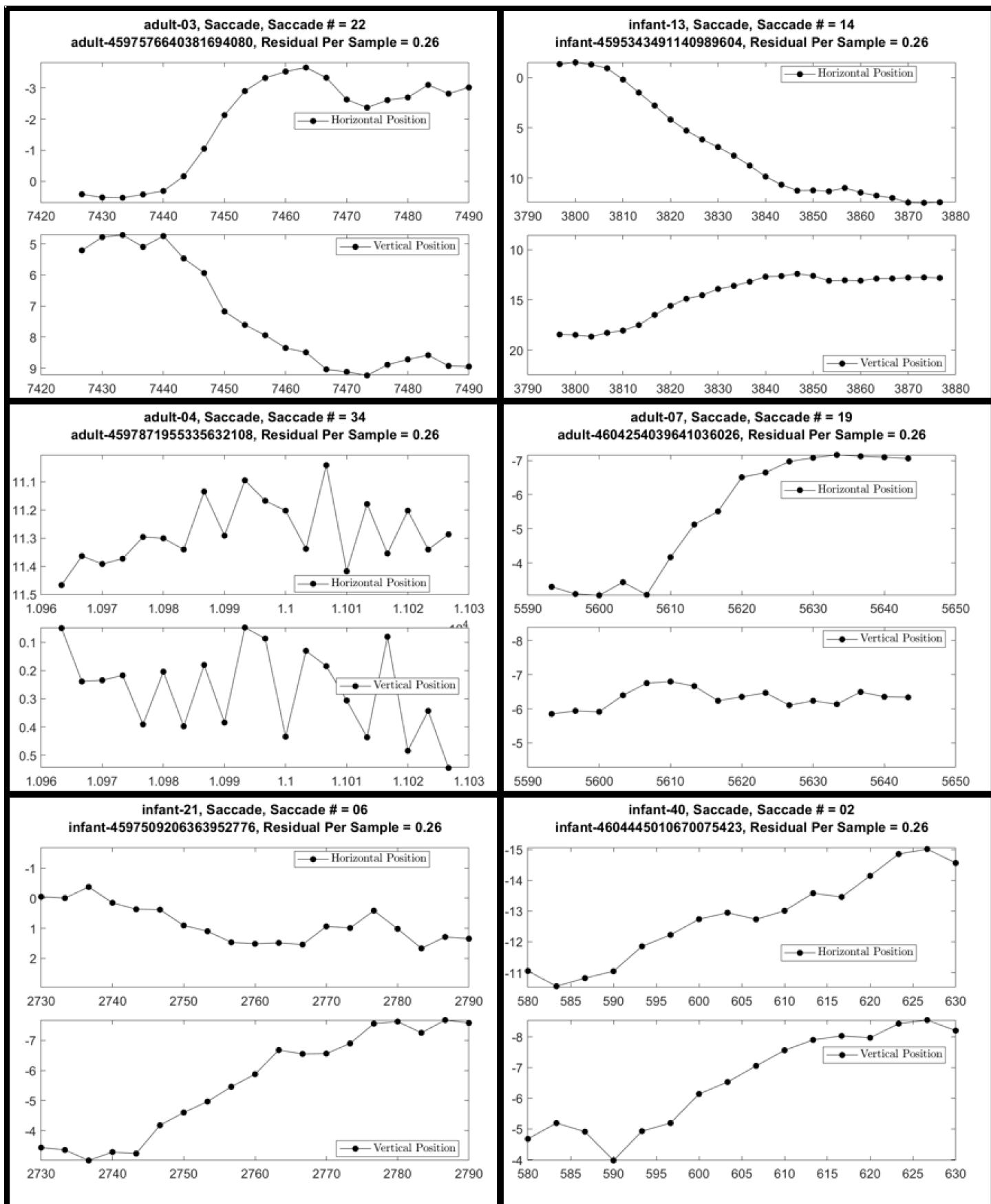
Figure 16:

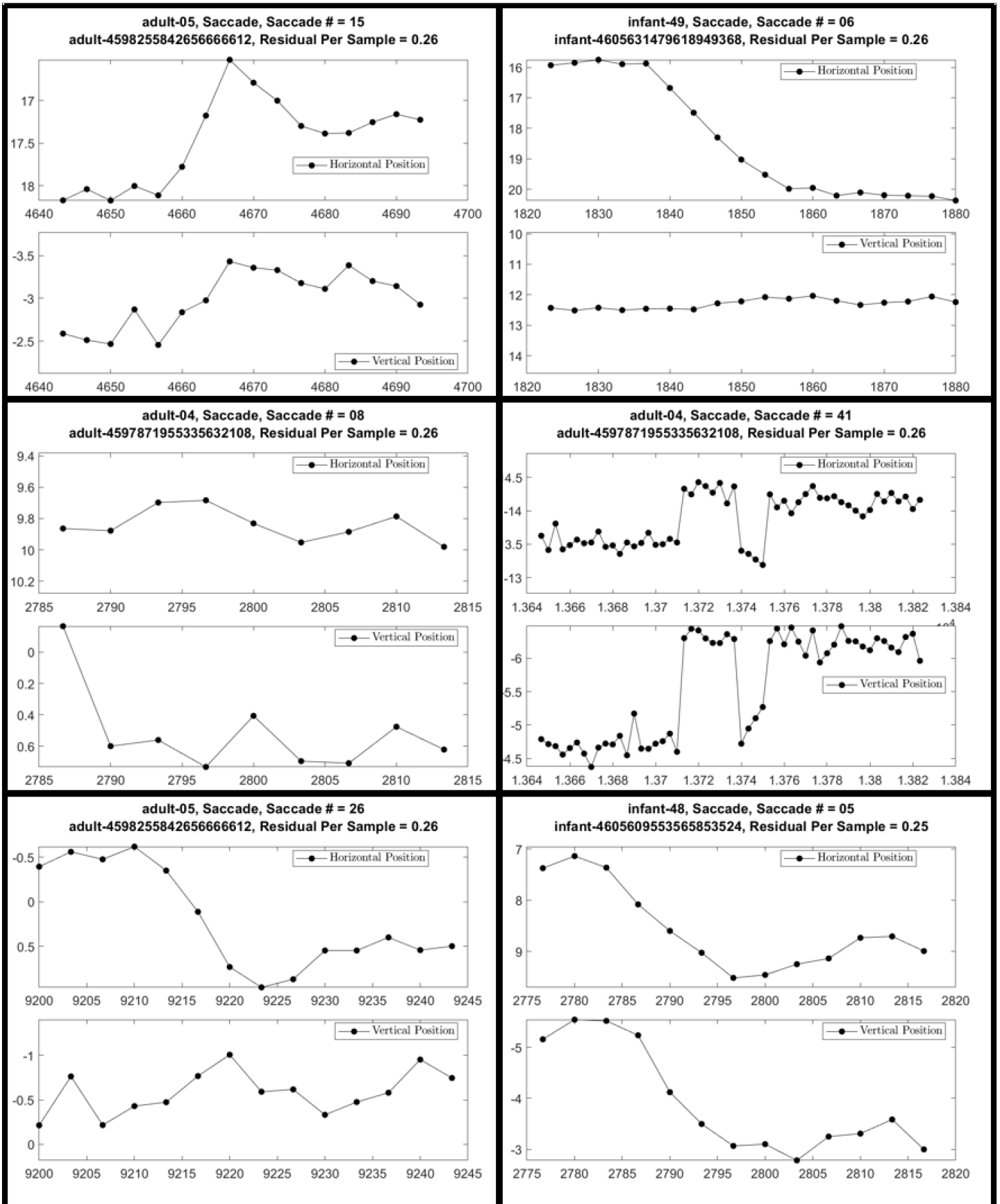Figure 17:

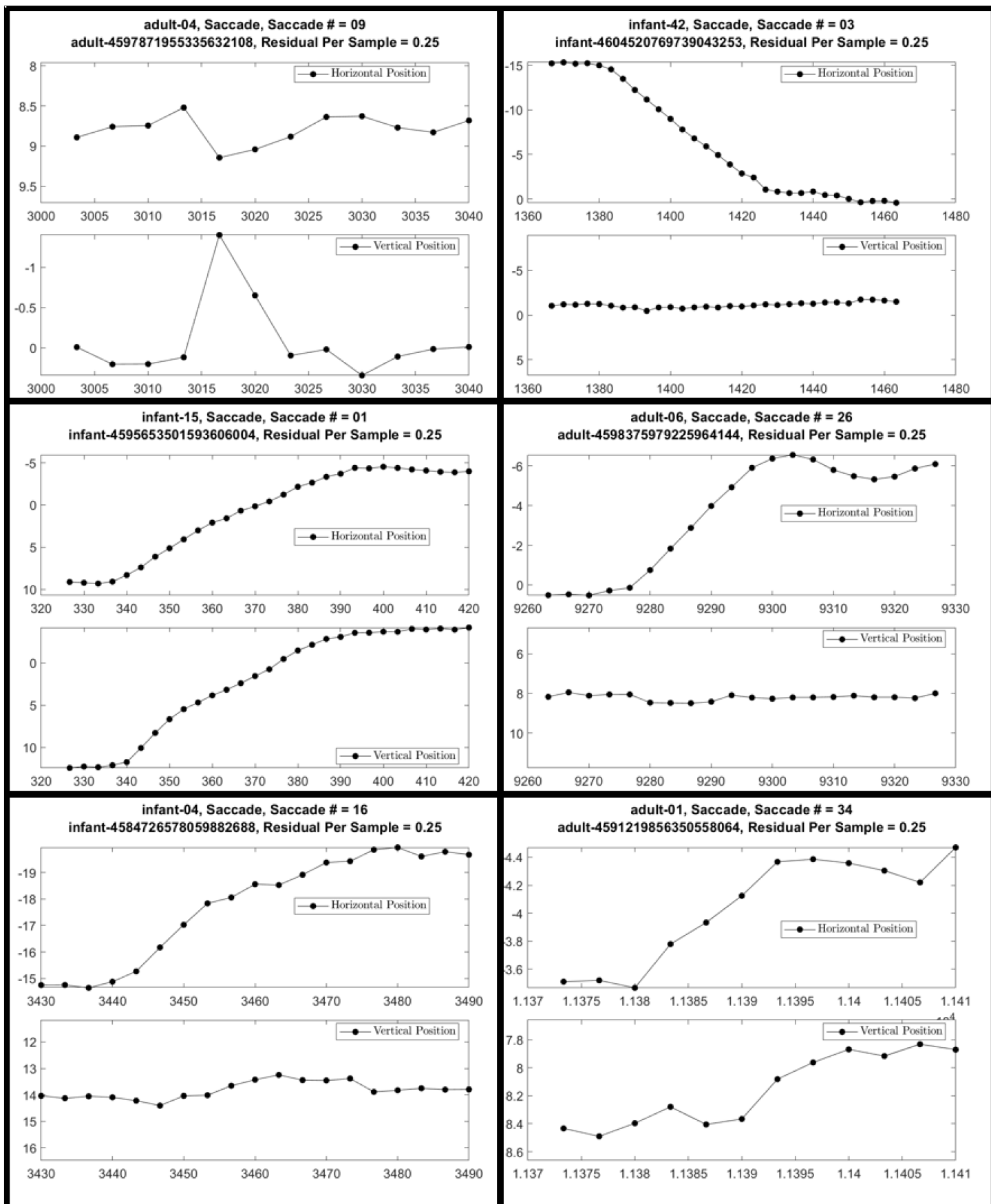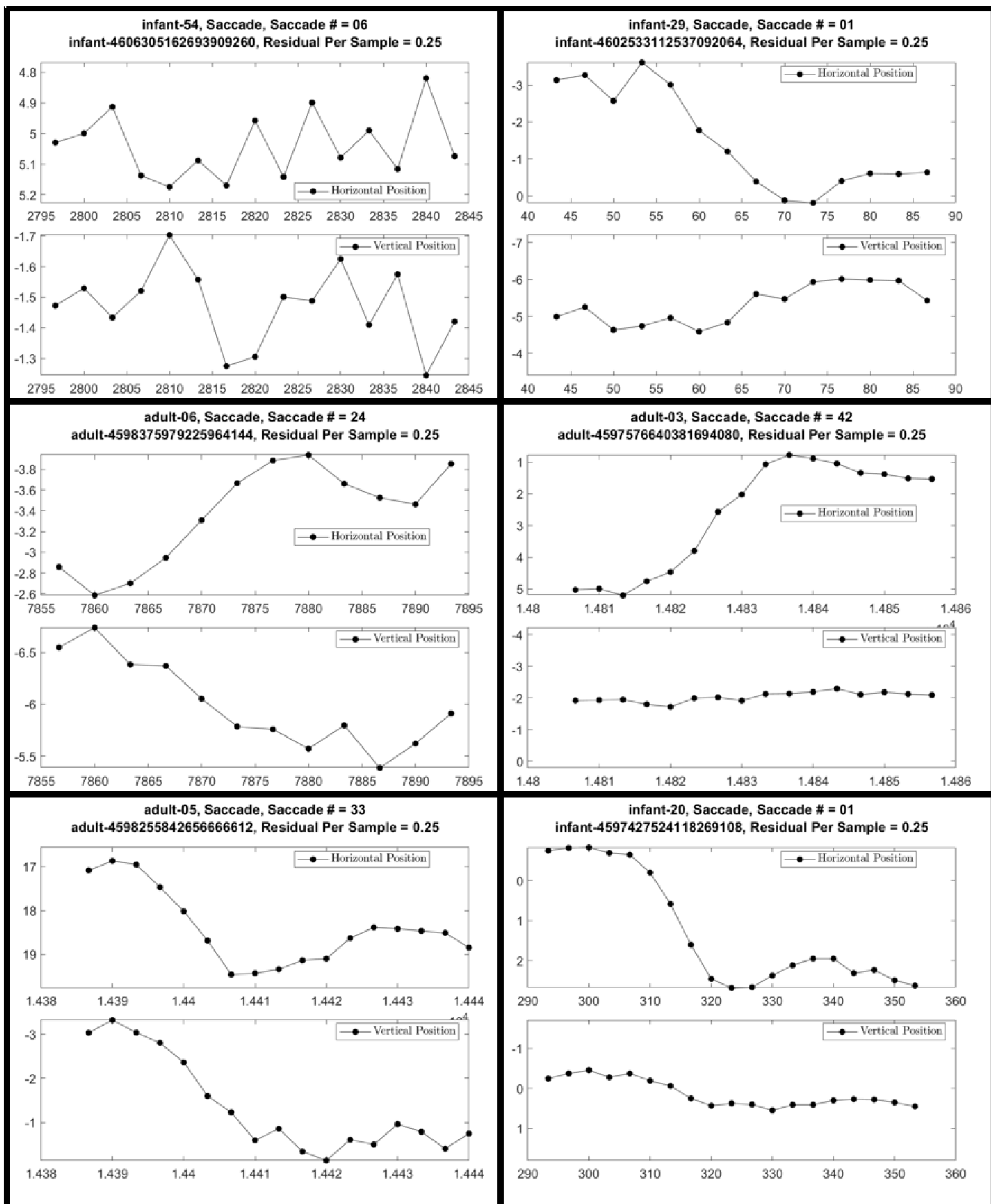Figure 18:

Figure 19:

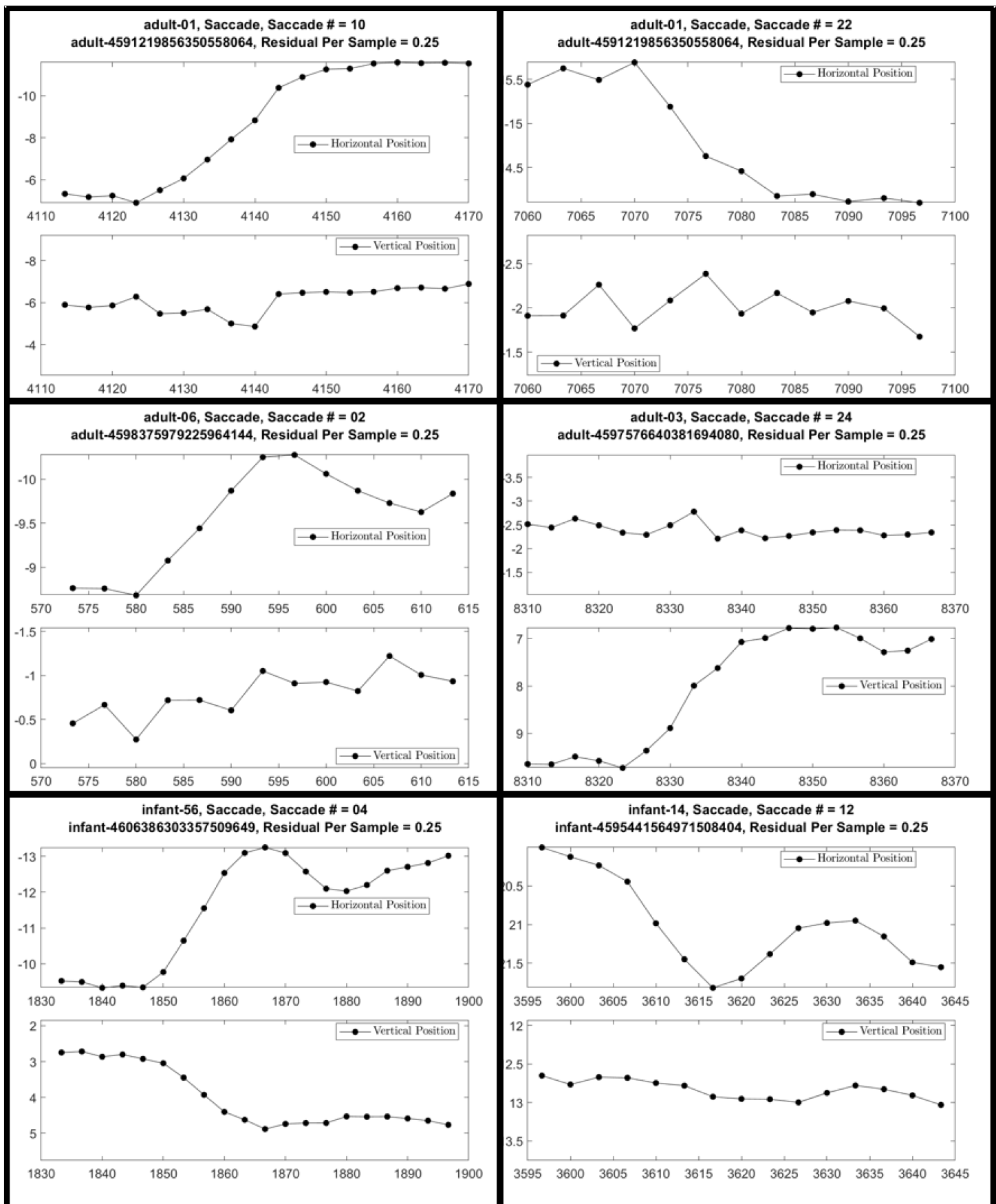Figure 20:

Figure 21:

Figure 22:
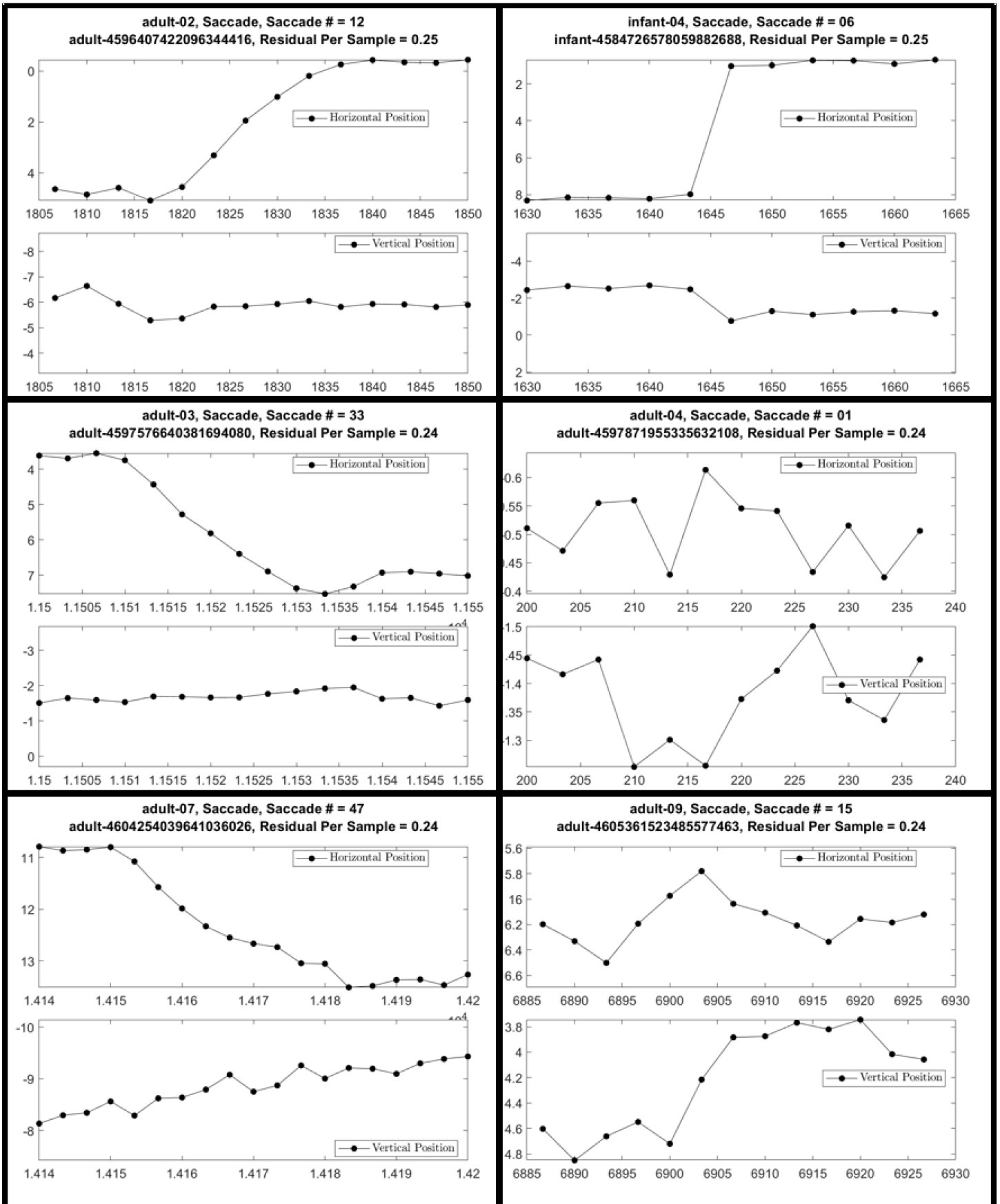
Figure 23:

Figure 24:

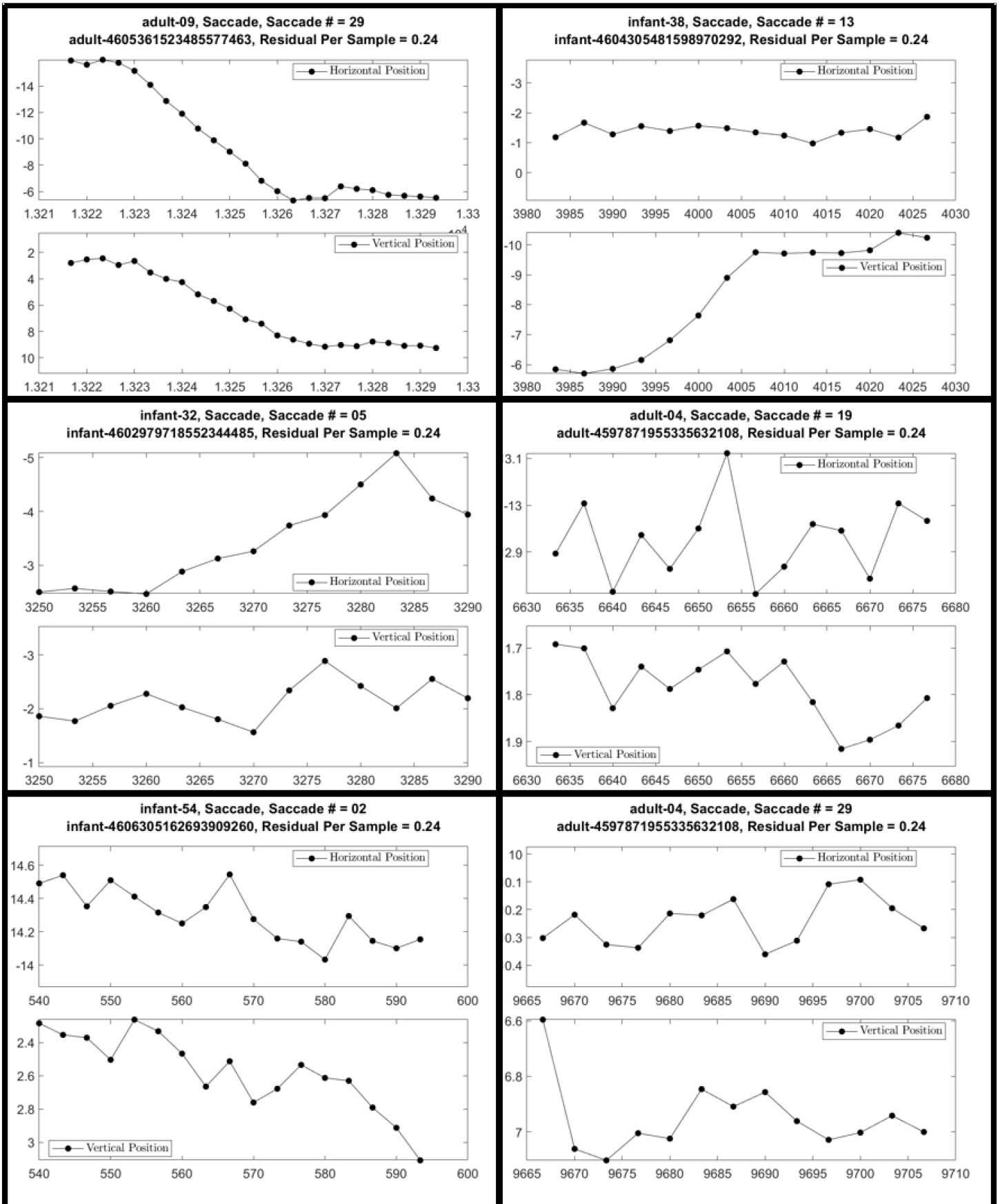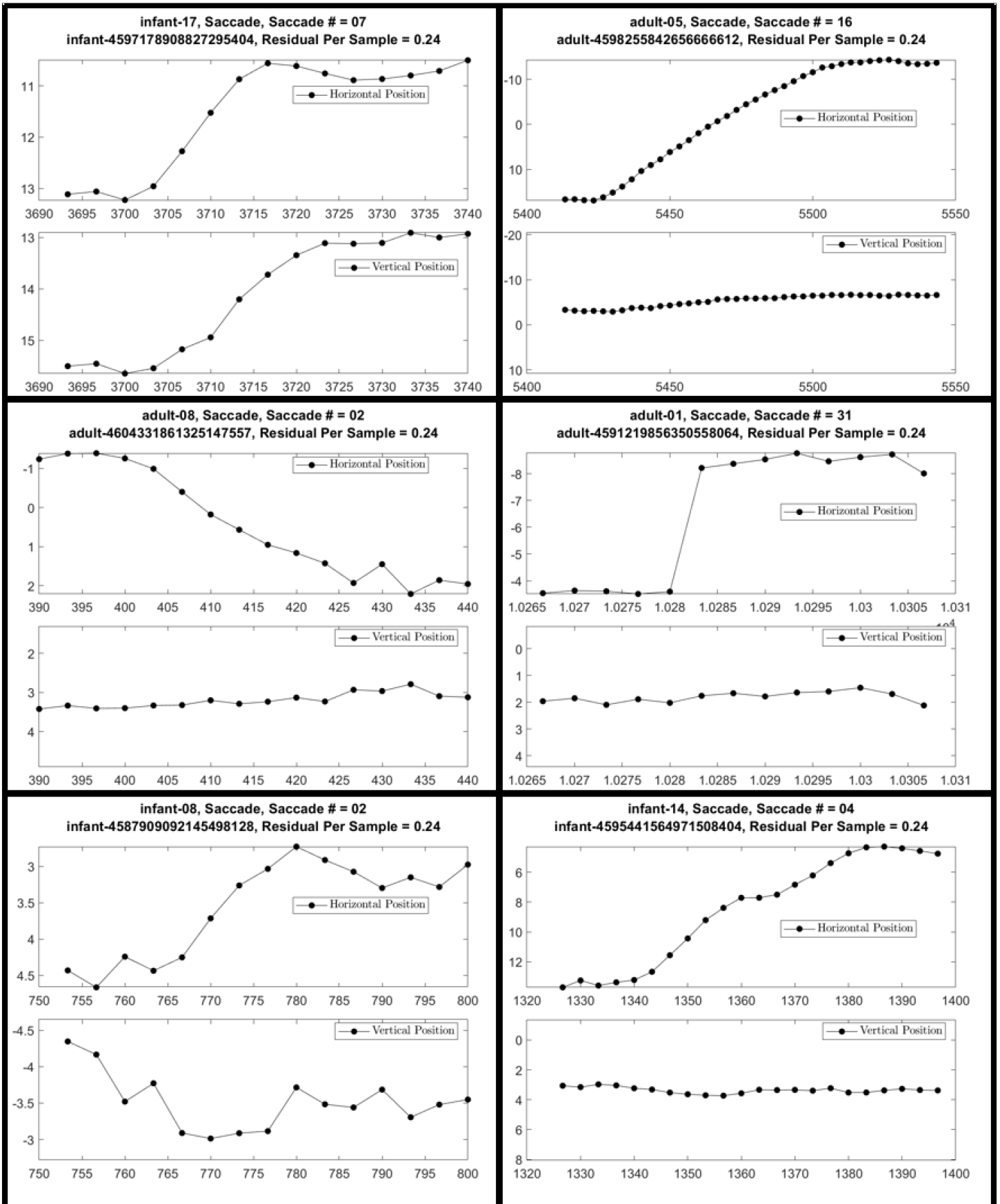Figure 25:

Figure 26:

Figure 27:
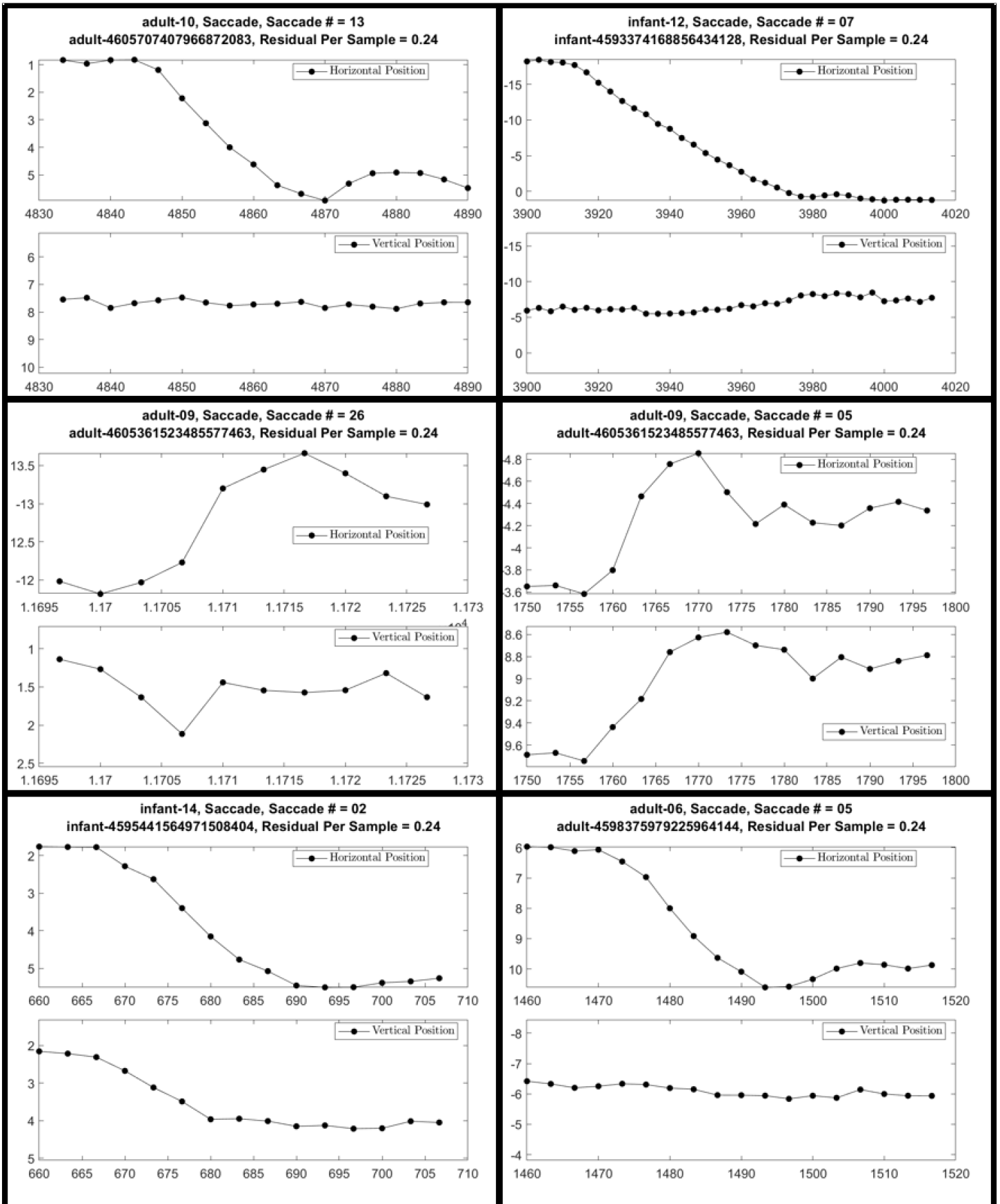
Figure 28:

Figure 29:

Figure 30:

7   Discussion

If the humanFixationEvaluation datset is processed according to the description found in the gazeNet paper, there are 929 potential saccades to be classified by an automatic algorithm for fixation rater MN. Of these, 332 or 36% of potential saccades were either preceded or followed by NaNs and are therefore are actually blink-saccades and not saccades. There are at least 162 saccades, and probably substantially more, that are very ill-formed. Some of these ill-formed potential saccades include well-formed saccades but either start too early or end too late. No trained and skilled human rater would consider these events to be properly identified saccades. Treating them as ground truth for saccades is insane. These events account for of a total 27% of 597 non-blink saccades. So, if we consider the total number of non-saccade events considered as saccades by gazeNet [1] (332+162) the total accounts for 36% of all potential saccades.

It would appear that if you wanted an algorithm to misclassify blink-saccades and ill-formed saccades as saccades that gazeNet is your algorithm. If you want to only identify true saccades as saccades it would probably be best to look for a superior algorithm.

References

[1] Zemblys, R., Niehorster, D. C., and Holmqvist, K. (2019). gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, 51(2), 840-864. https://doi:10.3758/s13428-018-1133-5.

[2] Hooge, I. T. C., Niehorster, D. C., Nystrom, M., Andersson, R., and Hessels, R. S. (2017). Is human classification by experienced untrained observers a gold standard in fixation detection? *Behavior Research Methods*. 50(5), 1864-1881 https://doi:10.3758/s13428-017-0955-x

[3] Hessels, R. S., Niehorster, D. C., Kemner, C., and Hooge, I. T. C. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior Research Methods*, 49(5), 1802–1823.

[4] Hessels, R. S., Hooge, I. T., and Kemner, C. (2016). An in-depth look at saccadic search in infancy. *Journal of Vision*, 16(8), 10–10.

[5] Dai, W., Selesnick, I., Rizzo, J.-R.,Rucker,J. C. and Hudson, T. E. (2016) A parameteric model for saccadic eye movement. In *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6.