REDUCING HIGH-FREQUENCY TIME SERIES DATA IN DRIVING STUDIES

Jeffrey D. Dawson, Amy Johnson O'Shea, Joyee Ghosh,
University of Iowa, USA

IASE, Kuala Lumpur, Malaysia

August 14, 2019

Acknowledgments:

NIA, NINDS, NHLBI, CDC, Nissan, Toyota;

UI Neuroergonomics Research Team;

UNMC Driving Research Team

Outline

- Intro to Driving Research
- Example of lane position data
- Comments about “Big Data” aspects of driving
- Specific model to handle “semi-reflective” data
- Methods to fit this model
- Simulations
- Lessons learned

Driving in U.S.—Public Health Issue

- 1.2 vehicles per drivers licenses in US
- 87% of those >=16 years old have licenses
- Crashes are ~7th most common cause of death (not grouped w/ other accidental deaths)
  - 1st in ages 15 to 24 yrs
  - 1st–2nd among accidental causes in all age groups >1 yrs
- High-risk groups
  - Young, inexperienced
  - Users of alcohol and other drugs
  - Elderly
  - Cognitively and/or physically impaired
- Trade-offs: safety, performance, quality of life, etc.

3 Pieces of Our Driving Research

Off-road Factors

- Demographics
- Disease status
  - Alzheimer's
  - Parkinson's
  - Sleep Apnea
  - Healthy
- Neuropsych tests
  - Vision
  - Cognition
  - Motor Skills
  - Interventions

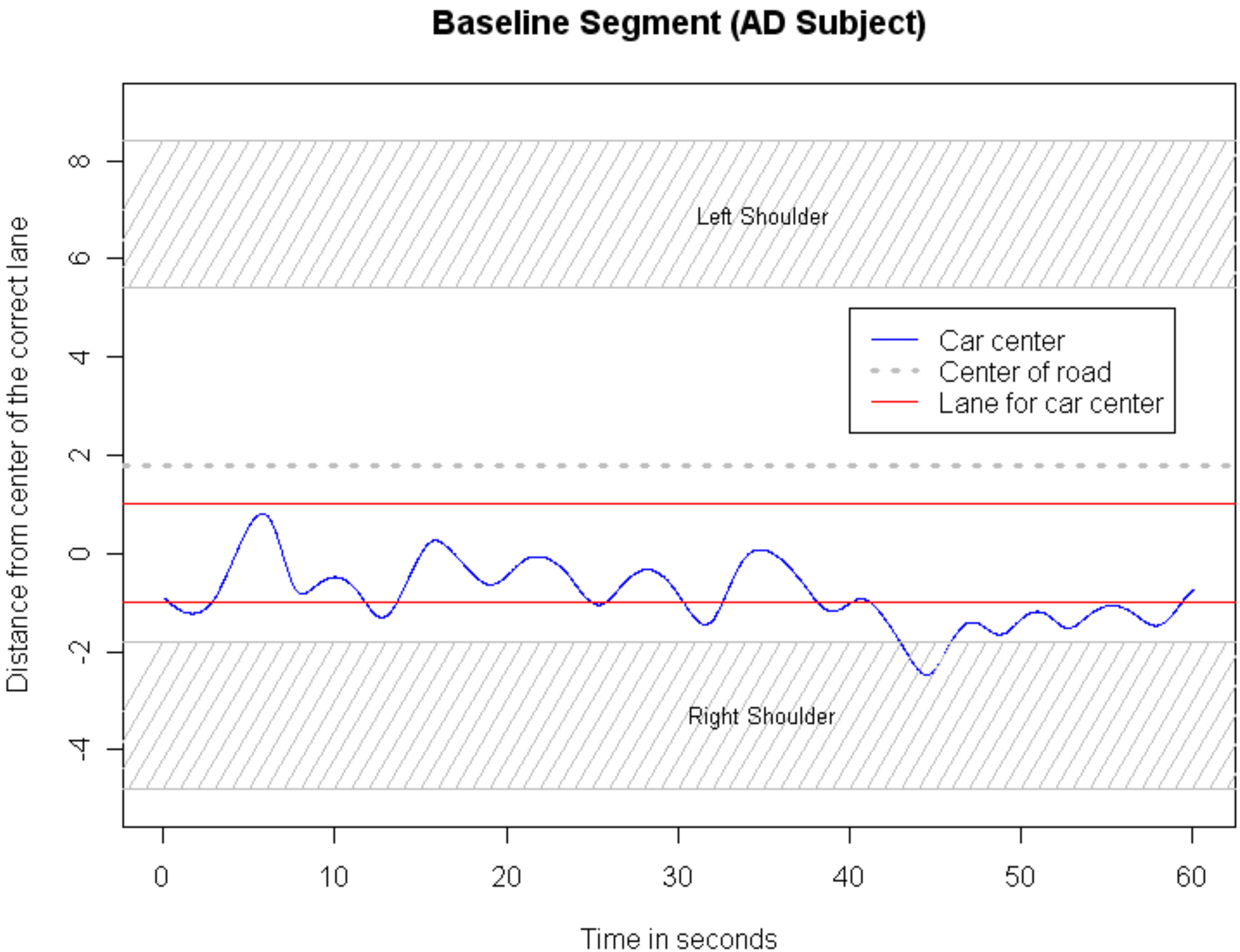
Driving Simulators

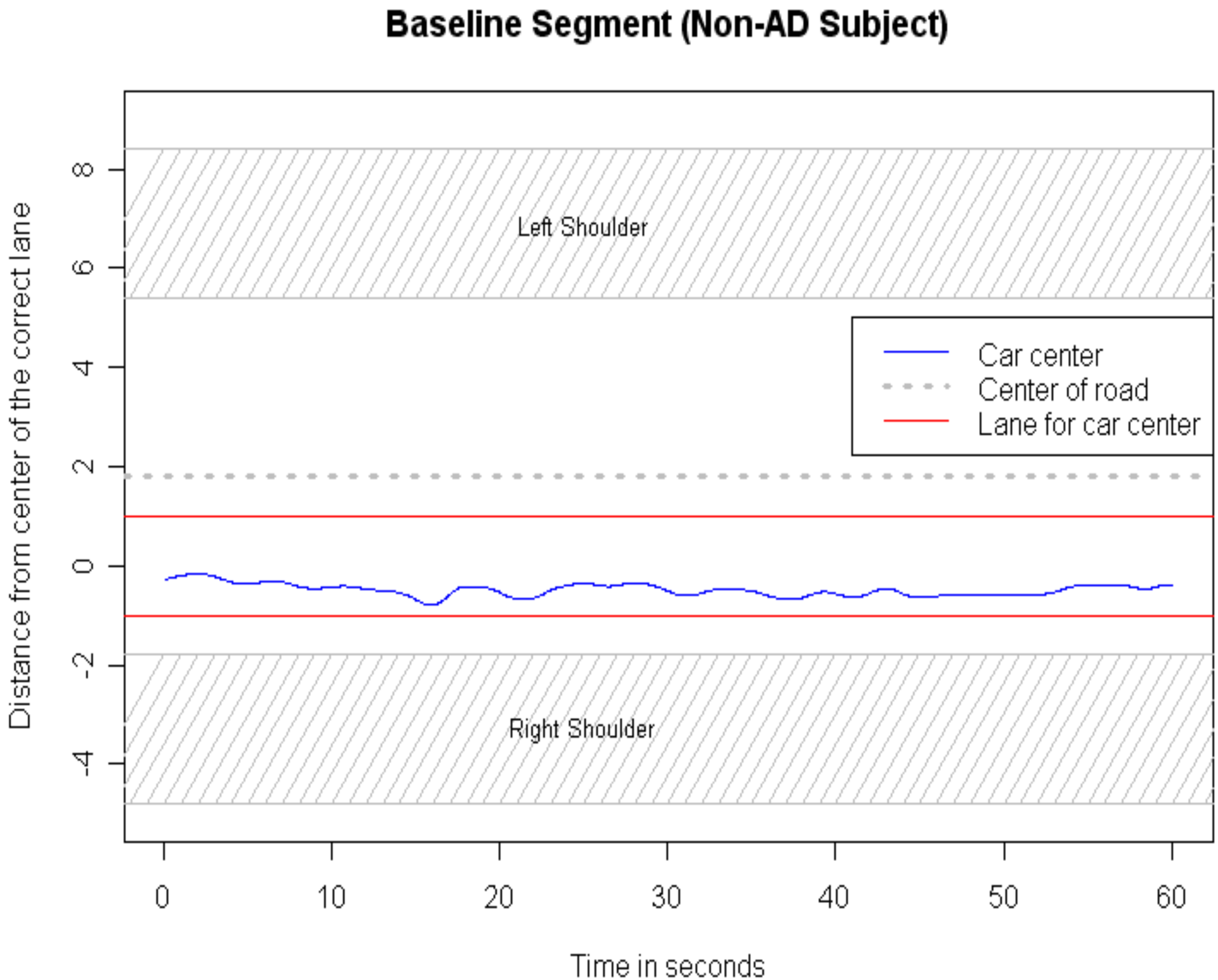
- Motion based
- Fixed base
- PC screen

On-road Outcomes

- (Closed track)
- Public fixed-route
- Naturalistic driving
- DOT/DMV records

Fixed Base Simulator: “SIREN” *(Rizzo et al, 2004)*


Baseline Segment (AD Subject)
Left Shoulder
Car center
Center of road
Lane for car center
Right Shoulder
Distance from center of the correct lane
-4
-2
0
2
4
6
8
0
10
20
30
40
50
60
Time in seconds

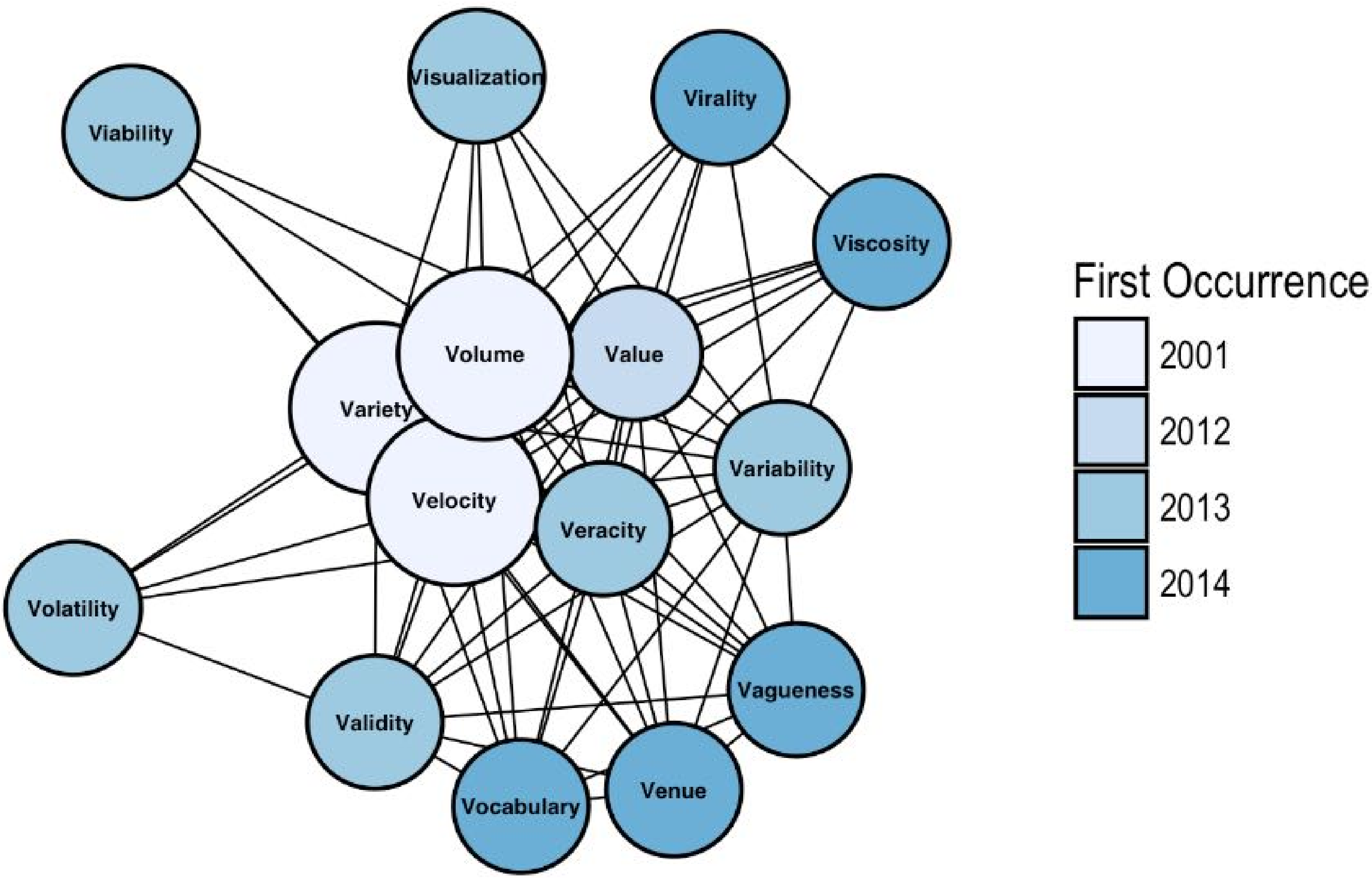

Baseline Segment (Non-AD Subject)
Left Shoulder
Right Shoulder
Car center
Center of road
Lane for car center
Distance from center of the correct lane
Time in seconds
8
6
4
2
0
-2
-4
0
10
20
30
40
50
60

Big Data Definitions

- Many exist--Gil Press of Forbes listed 5 and then added 7 more (https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#3c0167b013ae)
- Wikipedia (one of many)—"**Big data** is [sic?] data sets that are so voluminous and complex that traditional data-processing application software are inadequate to deal with them."

Big Data—Give me a V! (or 5 or 14 or 42)

https://www.elderresearch.com/blog/42-v-of-big-data ("voodoo")



Consider Five V Attributes/Issues

- **Volume**—Size of generated and stored data.
- **Velocity**—Could refer to capture rate but could also refer to timeliness of analysis
- **Variety**—Complexity (how many variables from how many sources, e.g., sensors and video).
- **Veracity** (truthfulness)—Accuracy (vs. noise and bias)
- **Value**—Benefit of analyzing the data (to businesses, individuals/society, etc.)

Arguments Against (some) Driving Data Being Called “Big”

- For some, Big Data problems are tied to idea of exploring data **without *a priori*** hypotheses (“data mining” or “analytics”), whereas our studies have specific aims and hypotheses.
- “Voluminous”, “complex”, and “traditional” in earlier definition are vague terms
  - To some, any tool newer than Excel is non-traditional
  - To others, high performance computing with parallel processors and code designed to make optimum use of them may be traditional.

Big Data Issues Seen in Driving Studies (departing from "V's")

1. Size (e.g., 90 days of driving → 1.62 million rows of data for one subject)
2. Disconnect between data creation & analysis (even if there are planned hypotheses, many analyses are not *a priori*)
3. Limitation of traditional methods (focus of this paper to reduce data)
4. Multidisciplinary aspects (make sure collaborators understand importance of accommodating random effects)

The Proposed Model

- At time $t>3$, model the lane position as:

$$Y_t = g(Y_{t-1}, Y_{t-2}, Y_{t-3}) + |e_t| I_t,$$

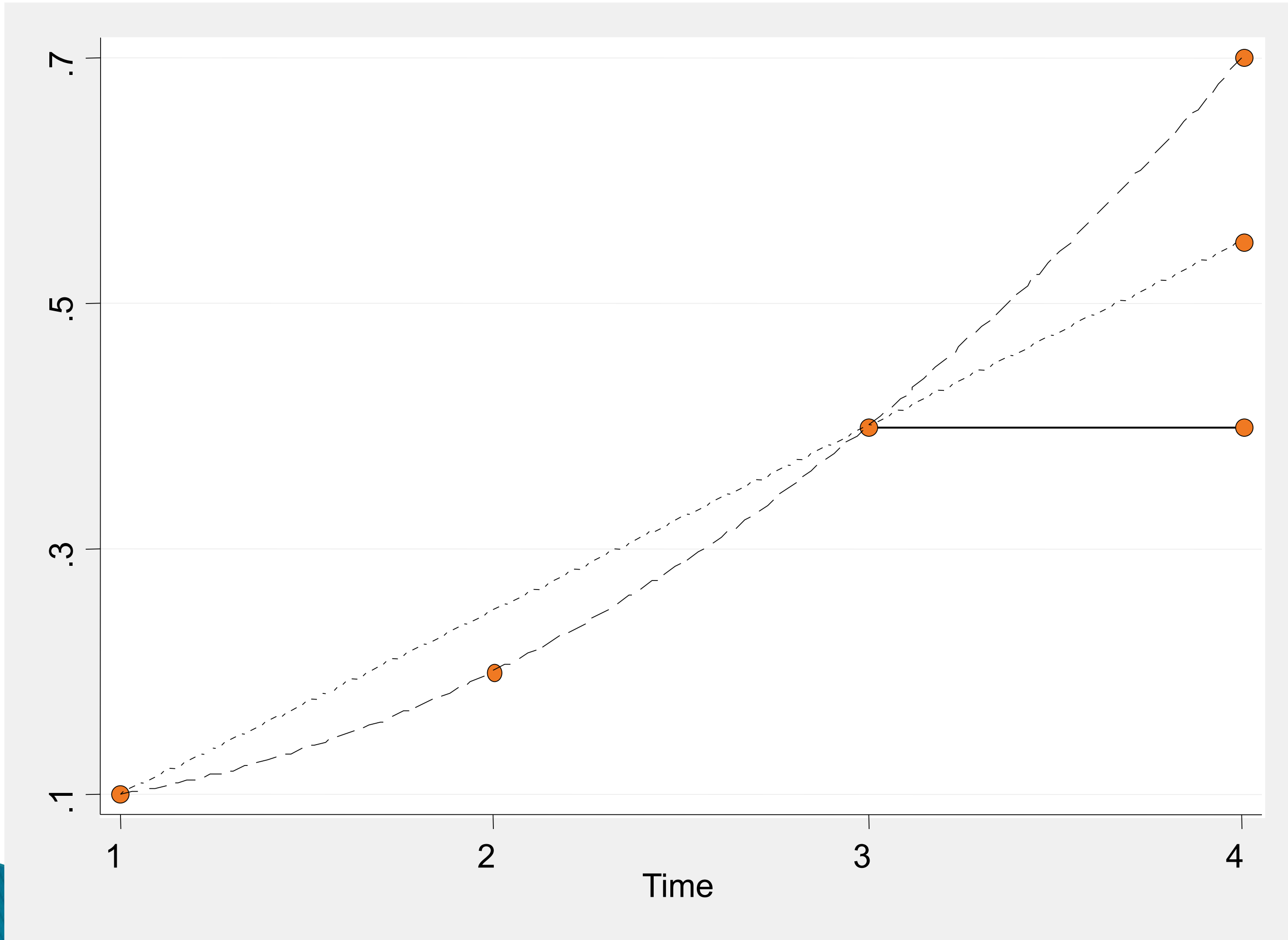
where $e_t \sim N(0, \sigma_e^2)$

and $\mathrm{Prob}(I_t=-1) = p_t$; $\mathrm{Prob}(I_t=1) = 1-p_t$

The Proposed Model (Con't)

- Parameterize $(Y_{t-1}, Y_{t-2}, Y_{t-3})$ as:
  - Flat Component: $W_{1t} = Y_{t-1}$
  - Linear Comp.: $W_{2t} = Y_{t-1} + (Y_{t-1} - Y_{t-3}) / 2$
  - Quad. Comp. : $W_{3t} = 3\, Y_{t-1} - 3\, Y_{t-2} + Y_{t-3}$
- Then, $g(Y_{t-1}, Y_{t-2}, Y_{t-3}) = \beta_1 W_{1t} + \beta_2 W_{2t} + \beta_3 W_{3t}$

Projection Examples

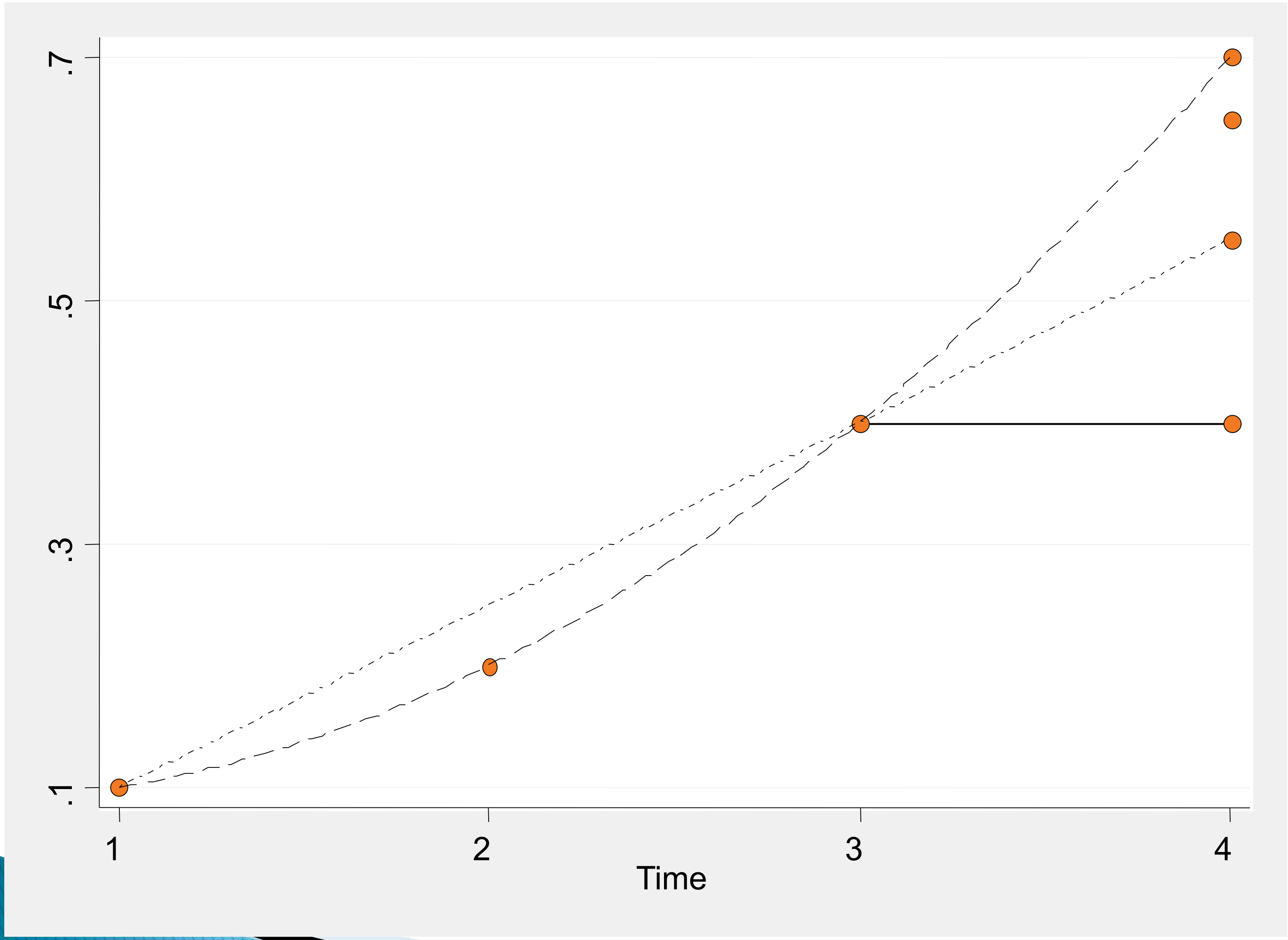


$\beta_3 = 1$
(Quad.)

$\beta_2 = 1$
(Linear)

$\beta_1 = 1$
(Flat)

Projection Example



$\beta_2, = .33;$

$\beta_3 = .67$

Getting Rid of One Parameter

- Recall that we have parameterized,

$$g(Y_{t-1}, Y_{t-2}, Y_{t-3}) = \beta_1 W_{1t} + \beta_2 W_{2t} + \beta_3 W_{3t}$$

- Add **constraints** that it is **weighted average**:

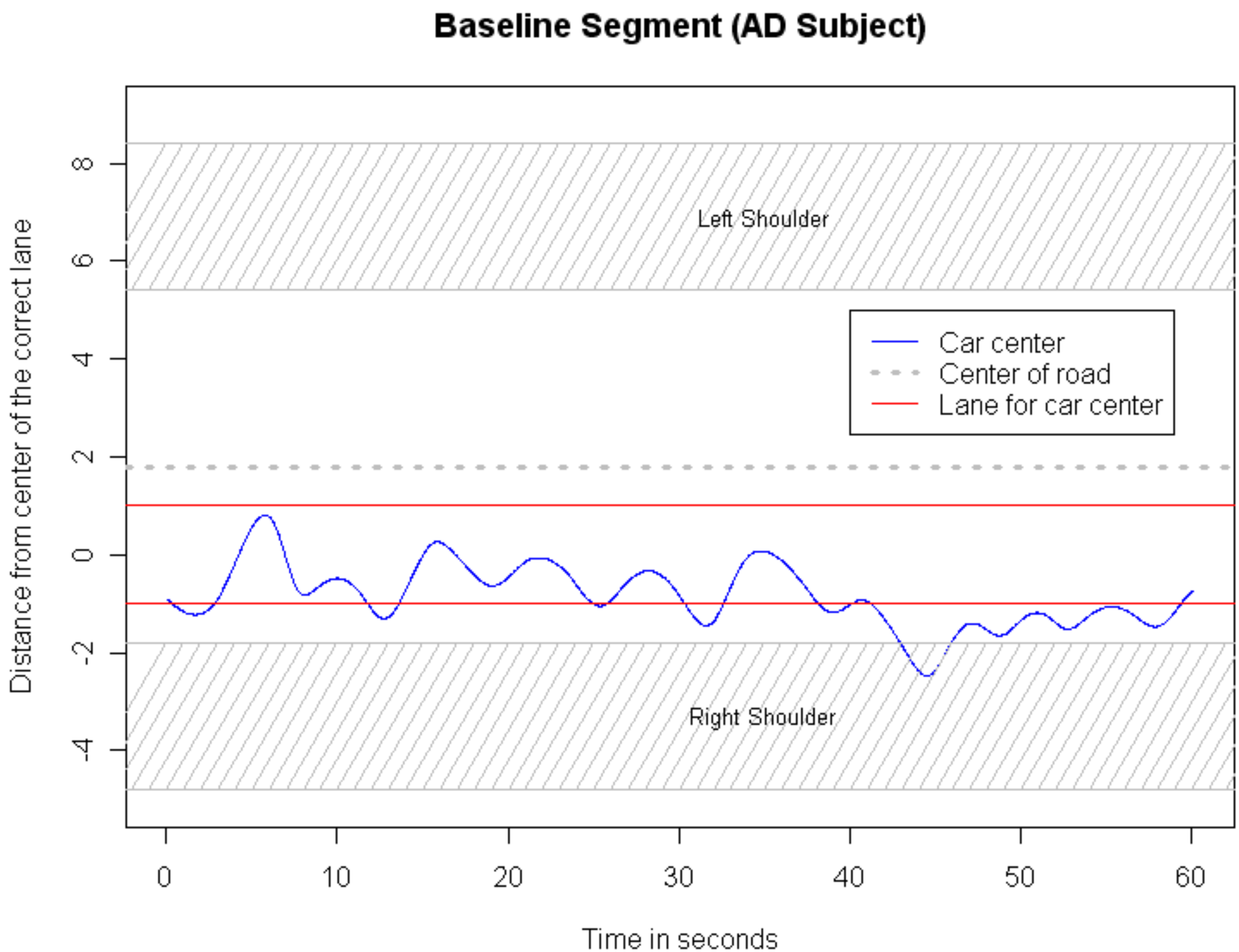
$$\beta_1 + \beta_2 + \beta_3 = 1, \text{ where all } \beta_i \geq 0$$

- Therefore:

$$Y_t = (1 - \beta_2 - \beta_3) W_{1t} + \beta_2 W_{2t} + \beta_3 W_{3t} + \text{err}$$

$$Y_t - W_{1t} = \beta_2 (W_{2t} - W_{1t}) + \beta_3 (W_{3t} - W_{1t}) + \text{err}$$

- Thus, the model can be **re-parameterized in terms of two β's**.


Baseline Segment (AD Subject)
Left Shoulder
Car center
Center of road
Lane for car center
Right Shoulder
Distance from center of the correct lane
Time in seconds
8
6
4
2
0
-2
-4
0
10
20
30
40
50
60

The Proposed Model (con't)

- Recall: $l_t = \log[p_t / (1 - p_t)] = \lambda_0 + \lambda_1 Y_{t-1}$
  - The intercept, λ_0, accommodates a subject's natural driving "center"
    - $\lambda_0 = 0$: subject's mean position is lane center
    - $\lambda_0 < 0$: subject's mean position is left of center
    - $\lambda_0 > 0$: subject's mean position is right of center
  - **The higher λ_1, the greater the probability that a subject turns back to center as the vehicle nears a lane boundary ("semi-reflective", since boundaries can be breached)**

Method of Fitting 1: "SP" (Single Pass)

- Create polynomial components
  - Flat Component: $W_{1t} = Y_{t-1}$
  - Linear Component: $W_{2t} = Y_{t-1} + (Y_{t-1} - Y_{t-3}) / 2$
  - Quad. Component : $W_{3t} = 3\,Y_{t-1} - 3\,Y_{t-2} + Y_{t-3}$
- (Ignoring usual assumptions), use **linear regression** to find **β_2, and β_3**
- Find **β_1 by subtraction**
- Calculate residuals and note the sign
- Use sign of residuals, the flat component, and **logistic regression** to get λ_0 and λ_1
- Use residuals to estimate **σ_e^2**

Methods 2 and 3 (likelihood-based)

- Letting θ be the vector of all parameters of interest, it can be shown that the **conditional log-likelihood** for the model (starting at 4th observation for one person) is

$$\sum_{t=4}^{T} log[f(y_t, I_t | y_{t-1}, y_{t-2}, \cdots, y_1; \boldsymbol{\theta})]$$

$$= \sum_{t=4}^{T} \Big\{ \log(2) - \frac{1}{2}\log[2\pi] - \log[\sigma_e] - \log[1 + exp(\lambda_0 + \lambda_1 y_{t-1})] - \frac{1}{2}\frac{(y_t - \mu_t)^2}{\sigma_e^2} + [\lambda_0 + \lambda_1 y_{t-1}] 1_{y_t < \mu_t} \Big\}.$$

Method 2: Grid search ("Grid")

- For each parameter (6–1=5 parameters)
  - Choose a min and max.
  - Have 5 equally spaced parameter settings (4 intervals)
- Calculate conditional log-likelihood for all combos
- Choose values which gave max.
- Use those values plus/minus one interval length to get new min and max (hence, total width reduced by 50% in each iteration)
- Repeat until converged.

Method 3: Modified Newton-Raphson ("NRmod")

- Likely problematic, there is the usual theoretical justification is not there (with likelihood is not smooth, and I_t being discontinuous and dependent on βs)
- Used SP method for starting values
- Used "half-stepping" approach to maximizing

Simulation settings

- $\beta_1=0.0546$, $\beta_2=0.4666$, $\beta_3=0.4788$
- $\sigma_e^2=0.0000214$ (i.e., $\sigma_e=0.00463$)
- $\lambda_0 = 0.634$, $\lambda_1=2.289$
- This was setting for all subjects (n=20)
- Each subject had 700 data points with first 100 being a **burn-in** after first 3 data points coming from simple random walk
- We looked at mean, variance, **% bias**, and **confidence interval coverage** of estimates

Simulation results

- All methods had some **bias**
  - SP had 0.1 to 11% in magnitude
  - Grid had 0.1 to 10% in magnitude
  - NRmod had 1.9 to 33% in magnitude
- All had **<95% coverage** for some parameters
  - SP: <50% for βs; ~95% for σ_e^2, λ_0; 84% for λ_1
  - Grid: 88–95% for all but λ_1 (which had 41%)
  - NRmod had 75% for λ_1, **0% for λ_1**, others 10–68%
- Interpretation: Since λ_1 is often most important, SP is "best", but still needs improvement.

Lessons Learned

- 1. We must find good metrics to reduce complicated data into meaningful parameters
  - The “re-centering” parameter has reasonable interpretability
  - Our model has shown good empirical properties (e.g., illustrating difference between drivers with and without Alzheimer’s disease)
  - Unfortunately, this simulation study showed bias and a range of coverage properties for all estimation methods considered.

Lessons Learned (cont'd)

- 2. Random effects must be accommodated
  - P-values can be inappropriately reduced by a factor of 10^{14} if you don't
  - We accommodated by doing separate analysis for each person, but with the same structure

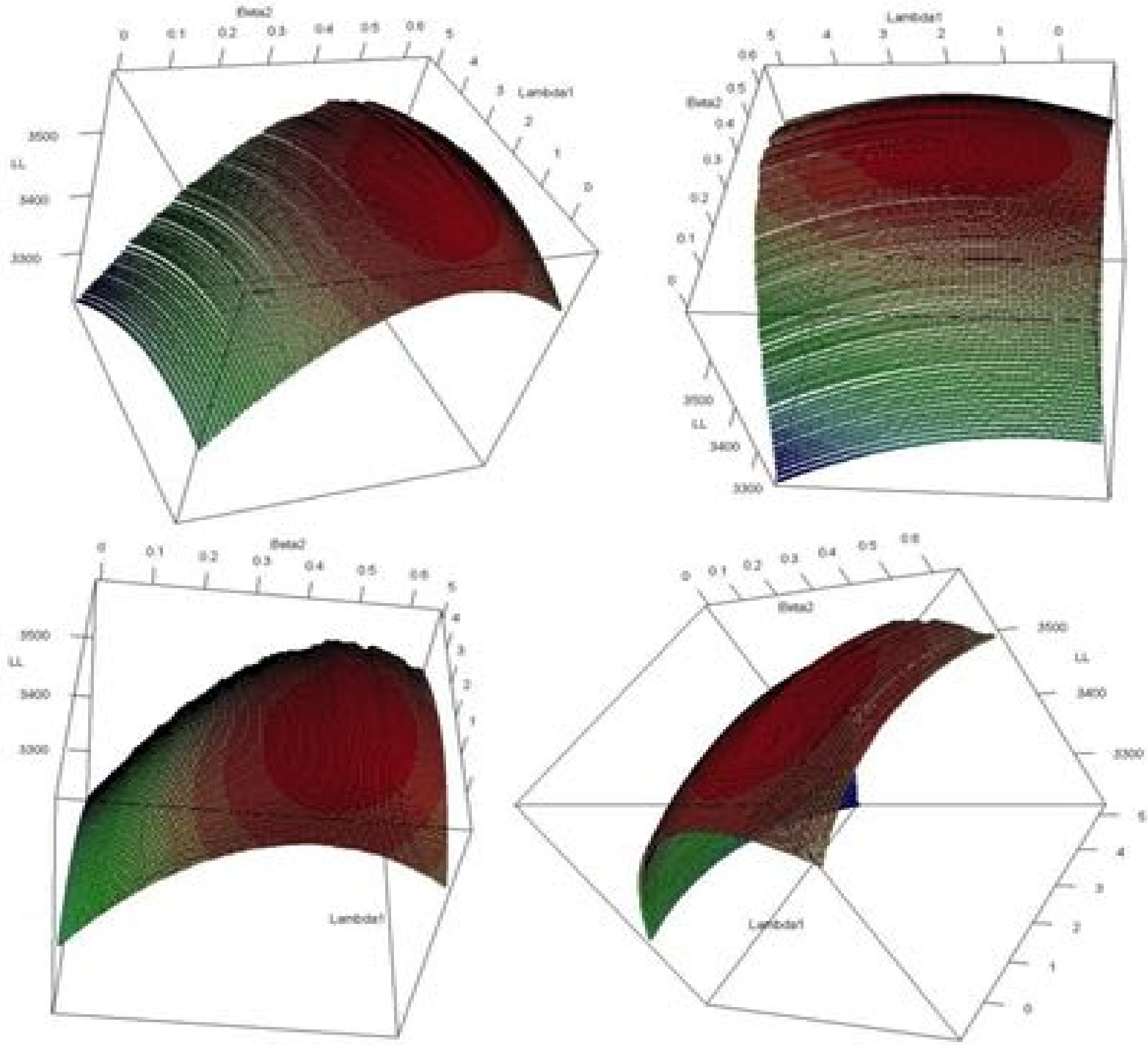
Lessons Learned (cont'd)

- 3. Important to know how to do "looping algorithms" to read in and process data
  - With 10,000 files of data, you do not want to type in all of those filenames!

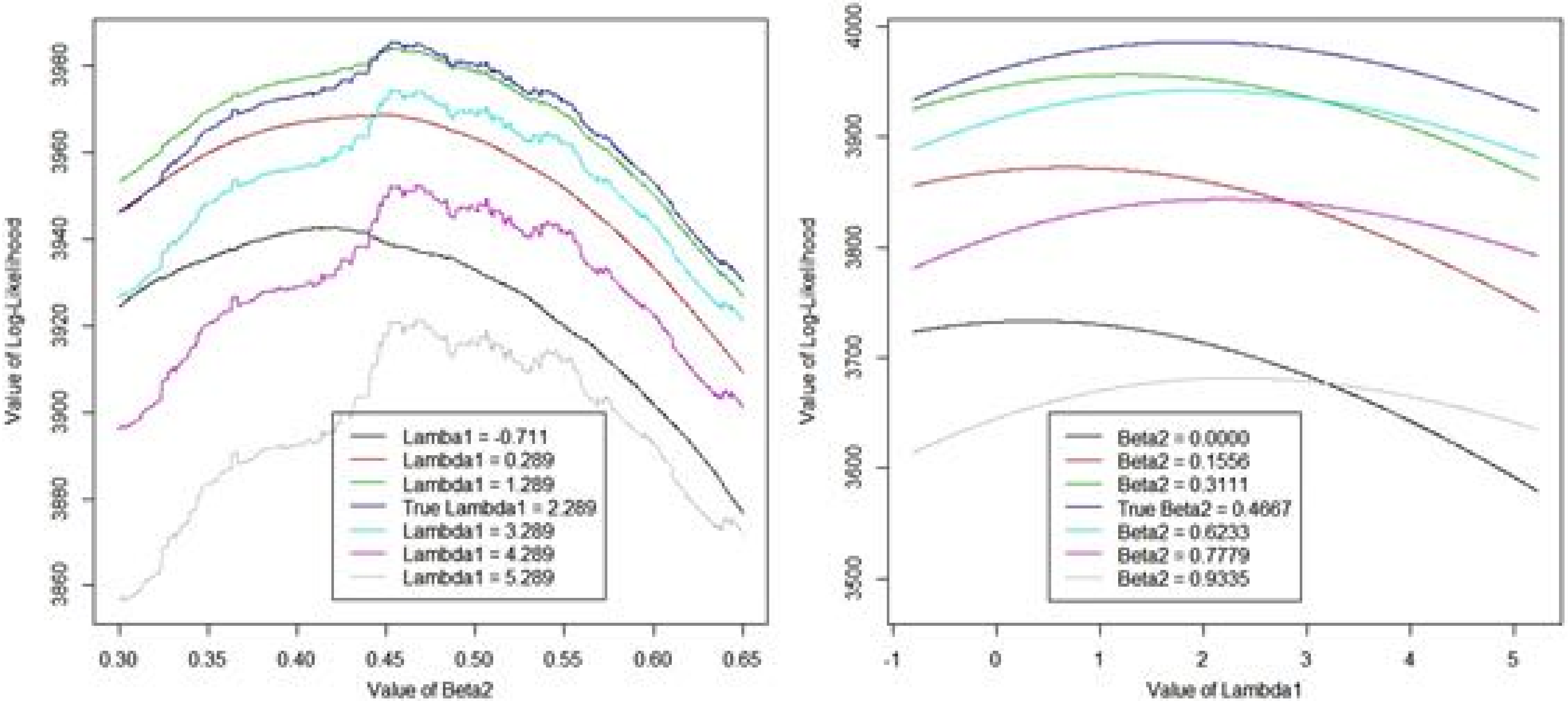
Lessons Learned (cont'd)

- 4. Even the "slow but sure" Grid searches are not guaranteed to find global maximum when there are several local maxima caused by "bumps".

Bumpy likelihood caused problems!



2D Graphs for λ_1 and β_2



References

- Boer, E. R. (2000). Behavioral entropy as an index of workload. *44th Annual Meeting of the Human Factors and Ergonomics Society (HFES2000)*. San Diego, CA.
- Dawson, J. D., Cavanaugh, J. E., Zamba, K. D. & Rizzo, M. (2010). Modeling lateral control in driving studies. *Accident; Analysis and Prevention, 42*(3), 891–897.
- Hamilton, J.D. (1994). *Time series analysis.* Princeton, NJ.: Princeton University Press.
- Johnson, A. M. (2013). *Modeling time series data with semi-reflective boundaries*, PhD thesis, University of Iowa. https://ir.uiowa.edu/cgi/viewcontent.cgi?article=4995&context=etd.
- Johnson, A. M., Dawson, J. D., & Rizzo, M. (2011). Lateral control in a driving simulation: Correlations with neuropsychological tests and on-road safety errors. *Proceedings of Driving Assessment 2011: The Sixth International Driving Symposium on Human Factors in Driving Assessment, Training, and Vehicle Design*.
- Kendall, M.G., & Ord, J.K. (1990). *Time series (3rd ed.).* London: Edward Arnold.
- Rizzo M (2004). Safe and unsafe driving. In: Rizzo M, Eslinger PJ, (Eds.). Principles and Practice of Behavioral Neurology and Neuropsychology (pp. 197–222). Philadelphia, Pennsylvania: WB Saunders.
- O'Shea, A. M. J., & D. Dawson, J. D. (2018). Modeling time series data with semi-reflective boundaries. *Journal of Applied Statistics*. 1–13. 10.1080/02664763.2018.1561834.

“Thanks!”