REDUCING HIGH-FREQUENCY TIME SERIES DATA IN DRIVING STUDIES

Jeffrey D. Dawson[1], Amy M. J. O'Shea[2], Joyee Ghosh[3]
University of Iowa Department of Biostatistics[1]
University of Iowa Department of Internal Medicine, VA Center at Iowa City[2]
University of Iowa Department of Statistics & Actuarial Science[3]
jeffrey-dawson@uiowa.edu

Driving behavior studies often capture electronic measures at 1-30 Hz for long intervals. It is important to find stochastic models that describe such data, with parameters that can be interpreted and accurately estimated. In this report, we review a family of models that are useful in describing the lateral position of a vehicle in a simulator. These models consist of "projection" and "signed error" pieces, with the latter containing a parameter representing the tendency for drivers to return the vehicles to a central position. We use ad hoc and likelihood-based methods to fit these models, but these all result in biased estimates. Fortunately, in two-group studies, simulations suggest that such biases may offset each other and hence that two-group comparisons may have acceptable accuracy. If we can resolve the bias issue, electronic data from a vehicle might be useful in predicting future errors and crashes.

INTRODUCTION

With advances in technology, an increasing amount of vehicular and driving simulator data are collected for research and other purposes in vehicles and driving simulators. Electronic measures, such as speed, acceleration, steering wheel angle, and lane position, are often captured at 1-30 Hz or even greater frequency, leading to huge datasets. For example, one 30-minute driving session in a 30-Hz simulator would result in 54,000 rows of data, while 90 days of driving for 30 minutes per day in an on-road naturalistic setting would result in 1.62 million rows of data if the capture rate is 10 Hz. With such volumes of data, it is imperative to reduce the data into meaningful metrics. To explore the statistical properties of such metrics, one needs models that could produce the data, as well as a specific method for estimating the parameters of such models. In this paper, we review a family of models that has been useful for reducing lateral position data in a driving simulator, which might also be used in real vehicles and with other variables. We then describe the distribution and likelihood functions based on these models, and present basic statistical properties of three methods used to estimate the model parameters.

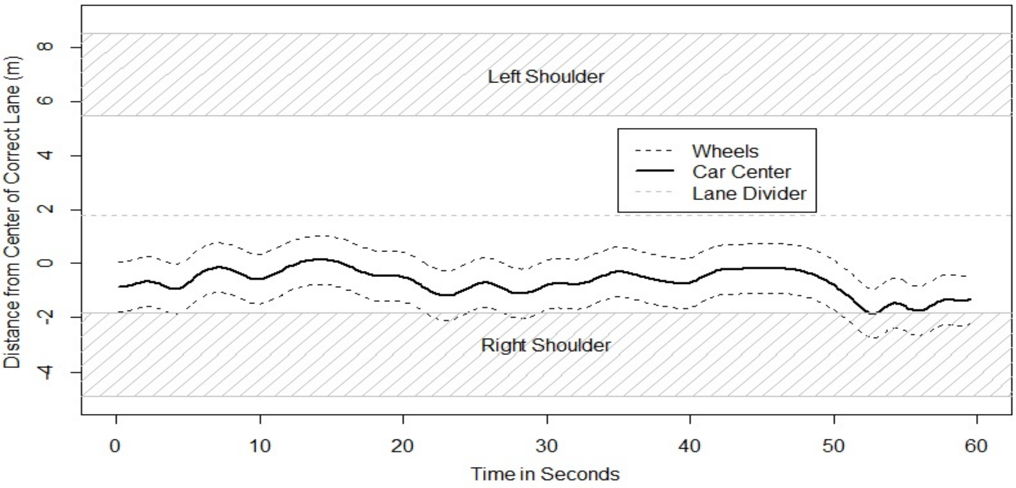

Figure 1. Example of Lateral Position Data in a Simulator

THE MODEL

The solid line of Figure 1 displays sample lane position data at the center of the vehicle, Y_t, at a given time, *t*. Note Y_t=0 when the physical center of the vehicle is in the center of the driving lane, while Yt > 0 and Yt < 0 correspond to being left and right of center, respectively. Drivers will tend to correct themselves toward the middle of the lane as they approach a boundary, but since such boundaries can

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. ©2019 ISI/IASE iase-web.org/Conference_Proceedings.php

still be crossed, they might be termed "semi-reflective". In this setting, lane position may be based on a third order autoregressive time series (Kendall & Ord, 1990), with an added sign term (Dawson et al., 2010). The general form of such a time series model for t > 3 is given by

$$Y_t = g(Y_{t-1}, Y_{t-2}, Y_{t-3}) \text{ ("projection piece")}$$
$$+ |e_t|\, I_t, \text{ ("signed error piece"), where}$$
$$e_t \sim N(0,\sigma^2), \text{ and}$$
$$p_t = prob\ (I_t = -1), \text{ else } I_t = 1.$$

To begin with, $g(.)$ is an unspecified function predicting Y_t with the three most recent observed lateral positions. The residual e_t is the difference between the observed and projected/predicted Y values, while I_t is a signed indicator equaling -1 and 1 with probabilities p_t and $1 - p_t$ respectively.

To aid interpretation, we reparameterize $(Y_{t-1}, Y_{t-2}, Y_{t-3})$ to (W_{1t}, W_{2t}, W_{3t}), by letting,

$$W_{1t} = Y_{t-1},$$
$$W_{2t} = Y_{t-1} + (Y_{t-1} - Y_{t-3})/2, \text{ and}$$
$$W_{3t} = 3Y_{t-1} - 3Y_{t-2} + Y_{t-3}.$$

Note these are flat, linear, and quadratic trends, projected from the most recent time point. Thus, the projected values can be codified as a linear combination of such trends, by specifying

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

This linear combination can be identified as a weighted average, by requiring $\beta_1 + \beta_2 + \beta_3 = 1$, as a "summation constraint" and requiring $\beta_1 \geq 0$, $\beta_2 \geq 0$, and $\beta_3 \geq 0$ as a "range constraint".

To give a functional form on p_t, we adopt a simple logistic model:

$$\log(p_t/[1-p_t]) = \lambda_0 + \lambda_1 Y_{t-1}.$$

A high positive value of λ_1 is generally desirable, as it indicates an increased tendency for the actual value to be closer to 0 (i.e., the middle of the driving lane for lane position data) than what the previous 3 points are predicting based on the β's. Thus, we sometimes refer to this as the re-centering parameter. In Dawson et al. (2010), for example, healthy elderly drivers had an estimated re-centering parameter that was 40% higher than that of drivers with mild Alzheimer's disease. Also, Johnson, Dawson, and Rizzo (2011) found this re-centering parameter to be associated with certain tests neuropsychological ability, as well as on-road safety errors on a fixed route.

THE CONDITIONAL FUNCTION AND THE LIKELIHOOD

It can be shown that, for $t > 3$, the joint conditional function of (Y_t, I_t) is

$$f(y_t, I_t \mid y_{t-1}, y_{t-2}, y_{t-3}, \boldsymbol{\beta})$$

$$= \begin{cases} \dfrac{2}{\sqrt{2\pi}\sigma_e} exp\left\{-\dfrac{1}{2}\left[\dfrac{y_t - \mu_t}{\sigma_e}\right]^2\right\} \times \dfrac{exp(\lambda_0 + \lambda_1 y_{t-1})}{1 + exp(\lambda_0 + \lambda_1 y_{t-1})}, & y_t < \mu_t \\ \dfrac{2}{\sqrt{2\pi}\sigma_e} exp\left\{-\dfrac{1}{2}\left[\dfrac{y_t - \mu_t}{\sigma_e}\right]^2\right\} \times [\,1 + exp(\lambda_0 + \lambda_1 y_{t-1})\,]^{-1}, & y_t \geq \mu_t \end{cases}$$

As described by Hamilton (1994), in many time series settings it is more feasible to apply a conditional likelihood based on the Markov property rather than the unconditional likelihood. In this context, it can be shown that the conditional log-likelihood is

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

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

METHODS OF ESTIMATION

We consider three methods for estimating the model parameters. The first is called the "Single-Pass" ("SP") approach, as it uses standard statistical techniques (multiple linear regression accommodating the summation and range constraints described above to estimate the β and σ terms, followed by simple logistic regression to estimate the λ terms), rather than an iterative process. This approach is almost identical to that originally proposed by Dawson et al. (2010), and described in detail by O'Shea & Dawson (2018). Briefly, the steps are as follows:

- Impose the summation constraint by replacing β_1 in the model with $1-(\beta_2 + \beta_3)$.
- Use multiple linear regression, with no intercept, to estimate the parameters of the model: $$Y_t - W_{1t} = \beta_2\,(W_{2t} - W_{1t}) + \beta_3\,(W_{3t} - W_{1t}) + e_t$$
- Check to ensure that estimates of β_2, β_3, and β_2+β_2 are all in the range of [0,1], remapping them to the closest point in the parameter space if they are out of range.
- Record the sign of the estimated errors from the regression model, and use simple logistic regression to model the probability of a negative sign (hence, obtaining estimates of λ_0 and λ_1).

The other two methods are based on the conditional log-likelihood shown on the previous page. We first employed a grid search algorithm ("Grid") to attempt maximization of the conditional likelihood, which we would expect to be slow but reliable. We next used a modified Newton-Raphson ("NRmod" algorithm), which we anticipated to be faster. Unfortunately, due to the I_t being discontinuous and dependent on the β vector, this approach does not have the usual theoretical justification. It also required additional modifications, such as half-stepping and using the Single-Pass estimates as starting values. More details of these approaches have been documented by Johnson (2013).

We performed computer simulations to examine the mean, variance, mean bias, and confidence interval coverage probabilities at a specific setting of the parameters. Due to the anticipated slow speed of the grid search, we limited our study to 100 datasets, each with a sample size of 20 subjects, with each subject having 600 data points after 100 points of burn-in.

RESULTS

Table 1 shows the specific parametric settings for the simulation, as well as the performance results. Note that the single-pass approach had mean biases of magnitudes of 0.1 to 11.4%, depending on the parameter, with the re-centering parameter (λ_1) having a -3% mean bias. For most parameters, the grid search performed noticeably better in terms of less bias and higher coverage probability. Unfortunately, the main exception to this trend was that it performed worse than the single-pass approach for λ_1, which may be of primary interest. The modified Newton-Raphson approach performed worse than the grid search for all parameters.

To gain additional understanding regarding why the grid search was not a clear winner over the somewhat ad hoc single-pass approach, we examined the log-likelihood functions for more insight. To reduce the dimensionality, we set $\lambda_0 = 0$, which corresponds to an average location centered between the semi-reflective boundaries of the series. We additionally set $\beta_1 = 0.0546$ and $\sigma_e = 0.0046$, which are values provided as the base parametric setting for data generation and are not adjusted as we calculate the log-likelihood. Data for a single time series with an exploitable length $T = 1{,}000$ were generated, where in addition to the above values, we set $\beta_2 = 0.4666$, $\beta_3 = 0.4788$, (by subtraction), and $\lambda_1 = 2.2890$. Finally, the value of the conditional log-likelihood is calculated such that β_2 and λ_1 vary along the ranges $[0.000, 0.650]$ and $[-0.804,\ 5.221]$ respectively with 100 equally spaced values each. It is also noted that the value of β_3 is determined by subtraction via the summation constraint described earlier. The resulting conditional log-likelihood surfaces are shown in 3- and 2-dimensional graphs in Figures 2 and 3.

Table 1: Parametric Settings and Results of Simulation Studies

Parametric Setting ($\boldsymbol{\theta}$)		$\boldsymbol{\beta_1}$ 0.0546	$\boldsymbol{\beta_2}$ 0.4666	$\boldsymbol{\beta_3}$ 0.4788	$\boldsymbol{\sigma_e^2}$ 2.14e-5	$\boldsymbol{\lambda_0}$ 0.6340	$\boldsymbol{\lambda_1}$ 2.2890
Method		$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$	$\widehat{\boldsymbol{\beta}}_3$	$\widehat{\boldsymbol{\sigma}}_e^2$	$\hat{\boldsymbol{\lambda}}_0$	$\hat{\boldsymbol{\lambda}}_1$
Overall Mean	SP	0.061	0.433	0.506	2.14e-5	0.617	2.224
	Grid	0.057	0.462	0.481	2.14e-5	0.697	2.522
	NRmod	0.059	0.445	0.495	2.10e-5	0.594	1.530
Overall Variance	SP	0.0001	0.0007	0.0007	1.48e-12	0.021	0.179
	Grid	0.0002	0.0008	0.0006	1.51e-12	0.025	0.215
	NRmod	0.0002	0.0007	0.0007	1.41e-12	0.021	0.187
Mean Bias (%)	SP	11.402	-7.242	5.758	-0.080	-2.619	-2.860
	Grid	4.314	-0.887	0.373	-0.122	0.579	10.181
	NRmod	8.598	-4.531	3.435	-1.929	-6.304	-33.138
Coverage Probability (%)	SP	36.00	0.00	2.00	96.00	92.00	84.00
	Grid	93.00	88.00	95.00	94.00	94.00	41.00
	NRmod	57.00	10.00	24.00	68.00	75.00	0.00

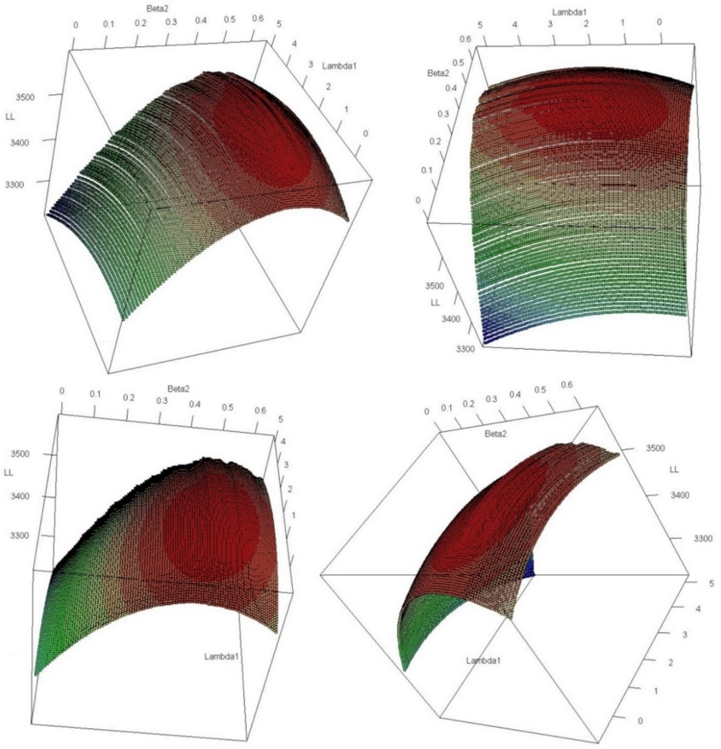

Figure 2. The Log-Likelihood Surface in Three Dimensions from Various Angles

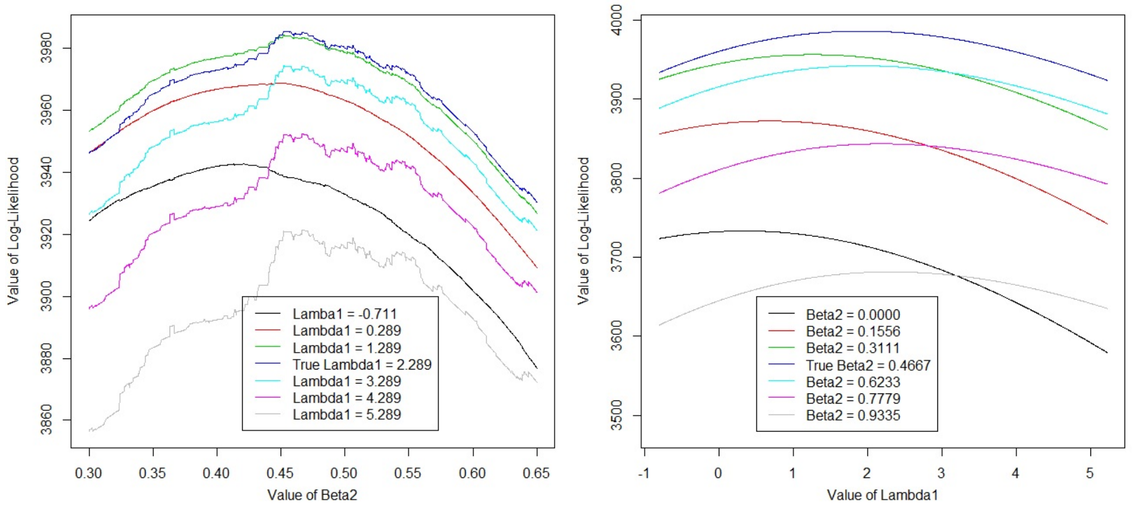

Figure 3. The Log-Likelihood Surface in Two Dimensions for a Range of Values of β_2 and λ_1

As can be seen in the graphs, the log-likelihood surfaces are neither smooth nor unimodal, and derivatives are undefined. Thus, Newton-Raphson techniques are not appropriate for maximizing this log-likelihood function. In fact, the multi-modal nature of the surfaces puts any likelihood-based technique, including grid searches, at risk of finding a local maximum instead of the overall maximum, leading to inaccurate results.

DISCUSSION/CONCLUSION

High-frequency electronic data from driving studies possess many of the attributes, as well as the challenges, of “Big Data”. In this paper, we emphasize the importance of reducing the volume of data, both to facilitate analysis as well as to provide interpretation. The model which we have proposed has interpretable parameters, but it is not clear how best to estimate them. For the intuitive re-centering parameter, it appears that our ad hoc “single-pass” method, despite its known bias, may be better than likelihood-based methods. Recently, O’Shea and Dawson (2018) found that a modified single-pass method performed better than the original approach, by imposing the summation constraint in a different part of the algorithm. In the original SP approach, the summation constraint was used before using multiple linear regression, so that only two β terms had to be estimated. In the modified SP approach, three β terms were freely estimated by multiple linear regression, and then the summation constraint was used to adjust the resulting estimates. It should also be noted that they found that, when doing simulations based on two-group situations, the bias somewhat canceled out between groups.

The reparameterization of the projection piece of the model into a linear combination of flat, linear, and quadratic terms may give better insight than the original model. For example, Dawson et al. (2010) found that the coefficient corresponding to the quadratic projection was 33% higher in drivers with Alzheimer’s disease compared to elderly controls. This suggests when impaired subjects attempt to make corrections in the lane position, they may over-correct in an exaggerated fashion rather than in a more tempered manner that may be more desirable. Dawson et al. (2010) also pointed out that the respective weights of 0, 1/3, and 2/3 for the flat, linear, and quadratic projections correspond to a 2nd-order Taylor series expansion of the process. This enabled them to make a more valid comparison between their method and an “entropy” method previously proposed (Boer, 2000).

The summation and range constraints were initially considered because a) the lane position in the driving simulator was well approximated by a model based on weighted polynomial projections, and b) using the summation constraint cut down on the dimensionality of the model, making the grid search more feasible. However, due to the problems with the grid search, it may be appropriate to remove those constraints altogether from the single-pass algorithms.

Other approaches to fitting the model might be considered. Since our SP approaches use multiple linear regression that ignores the serial correlation of the adjacent projections, an auto-regressive technique might be helpful. It may be possible to use a Bayesian approach to overcome the non-smoothness of the likelihood function.

Despite the bias that we have found with our estimation techniques, using a modified SP approach may be reasonable until better alternatives can be found. Our model should also be used for other types of data, including real-world driving, and other electronic measures besides lane position. Ultimately, having effective metrics, models, and estimation methods to distinguish between safe and unsafe driving patterns could result in better in-vehicle safety devices, and could also enable the vehicle to help diagnose cognitive and motor impairment in drivers.

There are number of educational lessons that can be highlighted based on the development, application, and investigation of our time series model in the context of driving research. First, it is essential that reasonable metrics for data reduction are proposed and investigated. Although it is important to use raw data for exploration and visualization, one generally must greatly reduce the data before formal statistical tests can be applied. Second, if the data reduction is done in subintervals of data (e.g., one-minute segments, or for individual drives when participants have multiple drives in the database), it is imperative that random effects models are used. This may seem obvious to most statisticians, but we have found that non-statistician members of research teams (e.g., engineers, computer scientists, physicians, etc.) are prone to forget this vital issue, and may attempt to present or publish results that have a high likelihood of Type I errors. Third, because of the high volume of data that are collected in driving studies, and because such data often take the form of multiple datasets (e.g., hundreds of datasets per driver in some on-road studies), it is necessary for statisticians to familiarize themselves with looping algorithms that import and process such files in sequence automatically. Fourth, a method for estimating the parameters of a proposed model, that appears to work empirically, may or may not exhibit appropriate statistical properties when investigated via simulations. Finally, some relatively simple models may end up with non-smooth likelihood functions, which are extremely difficult to maximize.

ACKNOWLEDGMENTS

This work was supported in part by the US National Institute on Aging of the National Institutes of Health, under award number RO1 AG017177.

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.

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.