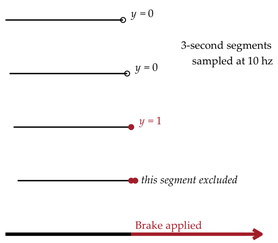
Fingerprinting individual driving behavior with vehicle kinematics data and dimension reduction regression

Brook Luers, Kerby Shedden

Department of Statistics, University of Michigan
MIDAS Center for Data-Intensive Transportation Research
luers@umich.edu

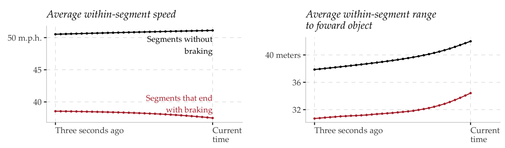
IVBSS naturalistic driving data

The Integrated Vehicle-Based Safety Systems (IVBSS) field trial (Sayer. et al., 2011) involved 108 drivers in southeast Michigan who drove instrumented study vehicles during their regular daily routines in a 40-day study period. These high-frequency, longitudinal data provide insight into variation in drivers' behavior when faced with a variety of driving contexts. In this case study, we focus on braking behavior by dividing each "trip" for each driver (a single ignition cycle) into three-second time segments, each ending in either an application of the brake pedal, denoted $y = 1$, or not ($y = 0$). Only segments that are entirely free of braking (y=0), or that are free of braking until the brake is applied at the last time point (y=1) are retained.

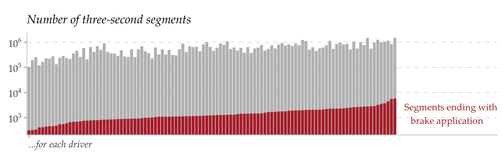


Vehicle kinematics

Driving behavior and context within each three-second segment are described by vehicle kinematic measurements recorded 10 times per second. In this analysis we include the vehicle's speed and radar-detected range (distance) to the lead vehicle, resulting in a 62-dimensional vector x of kinematic measurements in each time segment. We also exclude driving segments where the car's speed is less than 7 m/s (about 15 mph) during the final time point in that segment as well as segments in which there is no lead vehicle detected.



There are approximately 65 million three-second segments satisfying these criteria across the 20,317 driver-trips in this analysis.



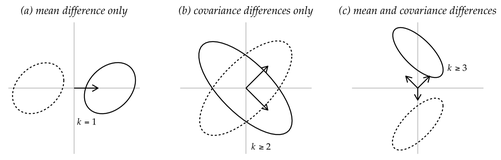
Dimension reduction regression

To understand heterogeneity in drivers' braking behavior in a given driving context, we apply the ideas of dimension reduction regression to construct interpretable, low-dimensional summaries of the relationship between braking probability and driving context.

Consider two populations of p-dimensional vectors x, labeled $y = 1$ and $y = 0$. We wish to characterize $\mathrm{P}(y = 1|x)$ using a small number of linear projections:

$$\mathrm{P}(y = 1|x) \approx f(b_1^\intercal x, \ldots, b_k^\intercal x) \quad \text{(for some function } f\text{)}$$

The number of projections, k, that adequately describes $\mathrm{P}(y = 1|x)$ depends on the extent to which the two populations differ in their location (means) and shape (covariances).



In (a), above, only one projection, or "index", is informative about $\mathrm{P}(y = 1|x)$, while in (b) and (c) multiple indices are informative. Many popular prediction methods including logistic regression and support vector machines are single-index models which make use of only one linear projection of the data.

Motivated by the techniques of *sufficient dimension reduction* (see below), we use the following dimension-reduction directions to characterize $\mathrm{P}(y = 1|x)$ (WLOG, $\mathrm{Cov}(x) = I$):

$$b_1 : \mathrm{E}(x|y = 1) - \mathrm{E}(x|y = 0)$$

$$b_2, \ldots, b_k : \text{dominant } k-1 \text{ eigenvectors of } \mathrm{Cov}(x|y = 1) - \mathrm{Cov}(x|y = 0)$$

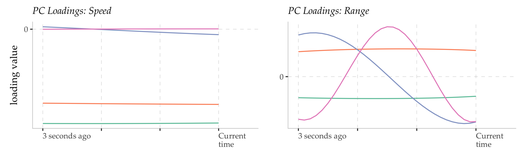
The first direction, b_1, the standardized mean difference, is the same direction used to classify points in linear discriminant analysis with two classes. The subsequent directions are the eigenvectors of the Difference in Covariances (DOC) matrix, which can capture additional structure in $\mathrm{P}(y = 1|x)$ not described by single-index methods.

Sufficient Dimension Reduction

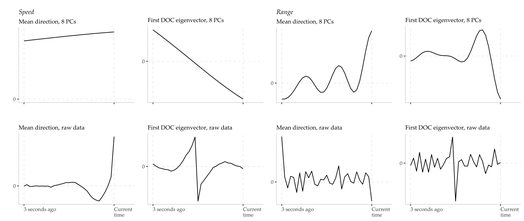
For a p-dimensional predictor vector x with response variable y, suppose we can find a matrix B such that x is independent of y given $s = B^\intercal x$ and that $\dim(s) < p$. Then we can replace x with the lower-dimensional vector s without losing any information about the conditional distribution $y|x$. In this way, s provides a "sufficient" reduction in the dimension of x. See Cook and Lee, 1999 for details.

Dimension reduction of kinematic data

To stabilize our estimates of the dimension reduction directions $b_1, \ldots, b_k$, we compute these estimates using the dominant q principal components (PCs) of the kinematic measurements rather than the raw measurements. The first four principal component loadings are displayed below.



The estimated directions b_1 (the standardized mean difference) and b_2 (first DOC eigenvector) are displayed below, using the first eight PCs and the raw kinematic data.



Computation

Converting the raw kinematic data to three-second segments and then computing the mean difference vector and DOC matrix requires a single "pass" through the data—the raw data files are read from disk only once—so this dimension-reduction technique requires modest amounts of memory even for large datasets. We stream through the raw data in small, in-memory subsets (e.g. 100,000 lines) to construct the conditional moments

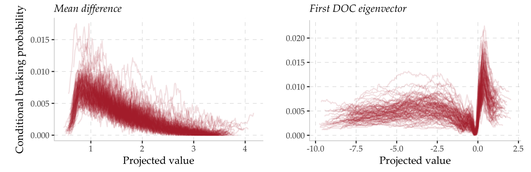
$$\mu_{y,j} \equiv \mathrm{E}(X|Y = y) \quad \Sigma_{y,j} \equiv \mathrm{Cov}(X|Y = y),$$

for the jth subset. These conditional moments are marginalized using the identities $\mu_y = \mathrm{E}(\mu_{y,j})$ and $\Sigma_y = \mathrm{E}(\Sigma_{y,j}) + \mathrm{Cov}(\mu_{y,j})$. All subsequent computations can be carried out using the μ_y and Σ_y.

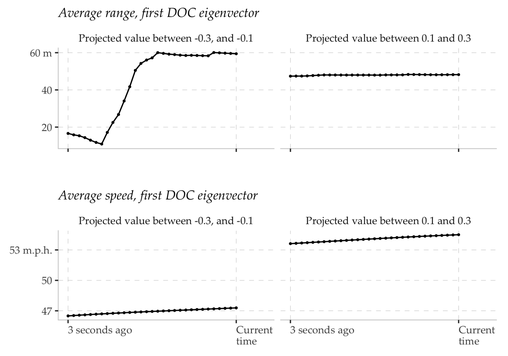
The data processing relies on a streaming data analytics library that we are developing called `dstream`, which facilitates complex manipulations of data streams by composing primitive operations into data processing pipelines. All calculations are carried out in Go (golang.org), a compiled, strongly typed language that supports high performance data processing and effective use of multi-core hardware. Source code is available from `https://github.com/kshedden/dimred`, and `https://github.com/kshedden/dstream`.

Individual braking behavior

We compute $b_1, \ldots, b_k$ by pooling the three-second segments from all drivers. The projections $b_1^\intercal x, \ldots, b_k^\intercal x$ were performed separately for each driver, and we estimated the conditional probability of braking along each direction, for each driver, by nearest-neighbors kernel regression. The estimated conditional braking probabilities for $b_1^\intercal x$ and $b_2^\intercal x$ are displayed below.



These estimates illustrate driver-to-driver variation in braking probability within the same driving contexts (at each fixed point along the horizontal axes). As an example of how one might interpret these dimension-reduction directions, we compute the average speed and range for each point x satisfying $b_2^\intercal x \in [-0.3, -0.1]$ or $b_2^\intercal x \in [0.1, 0.3]$:



Drivers have higher braking probabilities along $b_2^\intercal x$ during segments of near-constant average following distance and high average speed.

References

Cook, R. D. & Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association, 94*(448), 1187–1200.

Sayer., J., LeBlanc, D., Bogard, S., Funkhouser, D., Bao, S., Buonarosa, M., & Blankespoor, A. (2011, June). *Integrated vehicle-based safety systems field operational test final program report*. (Tech. rep. No. UMTRI-2010-36). Transportation Research Institute, University of Michigan. Ann Arbor, Michigan.

Acknowledgement: Building a transportation data ecosystem (MIDAS Challenge Initiative, PI: C. Flannagan)