IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2020, accepted September 16, 2020, date of publication September 24, 2020, date of current version October 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026392

Driving Stability Analysis Using Naturalistic Driving Data With Random Matrix Theory

KAI SONG[1], FUQIANG LIU[1], (Member, IEEE), CHAO WANG[1], (Member, IEEE), PING WANG[1], (Member, IEEE), AND GEYONG MIN[2], (Member, IEEE)

[1]School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[2]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, U.K.

Corresponding author: Chao Wang (chaowang@tongji.edu.cn)

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B090912002, and in part by the National Natural Science Foundation of China under Grant 61331009. The work of Chao Wang was supported in part by the National Natural Science Foundation of China under Grant 61771343.

ABSTRACT Driving behavior analysis has diverse applications in intelligent transportation systems (ITSs). The naturalistic driving data potentially contain rich information regarding human drivers' habits and skills in practical and natural driving conditions. But mining knowledge from them is challenging. In this paper, we propose a novel approach for analyzing driving stability using naturalistic driving data. Our method can extract features, based on the random matrix theory, to reflect the statistical difference between actual driving data and the data that would be generated by a theoretically ideal driver, and thus imply the skillful level of a driver in terms of vehicle control in both longitudinal and lateral directions. The execution of our method on a practical ITS dataset is conducted. Using the extracted features, a driving behavior analysis application that partitions drivers into clusters to identify common driving stability characteristics is demonstrated and discussed.

INDEX TERMS Driving behavior analysis, intelligent transportation systems, random matrix theory.

I. INTRODUCTION

With the ever-increasing number of vehicles on the road, traffic accident and jam become serious issues in most modern countries. The problems cannot be completely solved by conventional transportation engineering methods such as urban design and traffic control [1], [2]. The rapid development of information and communication technology (ICT) in the past decades enables the promising concept of intelligent transportation system (ITS). Equipping the transportation system with advanced sensing, communication, and computing capabilities can greatly enhance its capacity [3]–[5].

Most ITS services and applications target providing drivers with better knowledge regarding the driving environment to help decision making. Understanding human drivers' natural behaviors and habits is also important. Driver assistance services with driving behavior analysis functions can help a driver to be aware of the state of the vehicle, and of improper operations of his/hers own as well as surrounding drivers' [6]. Such high-level information is also beneficial to the design of ITS [7]. Having the knowledge of human driving behaviors can further enable artificial intelligence (AI) autonomous driving agents to learn from good drivers and evolve to better ones that ensure both safety and comfort for passengers [8].

Driving behaviors have been investigated from diverse perspectives. The analysis is typically performed by studying maneuvering actions (e.g., accelerating, turning, car-following, and lane changing) using vehicle states (including speed, acceleration level, steering angle, and yaw rate, etc), represented by various vehicle sensor data. Control skill serves as the basis for successfully taking proper driving actions in complex traffic environments [9], and thus is also important in driving behavior analysis. [10] points out that driver evaluation should take both their operational control actions and ability of understanding the environment into consideration. The characteristics of the operational control are inherently determined by driver's control skill [11]. Understanding such skill is valuable and can help develop customized ITS services and applications [12].

Human drivers and many driving assistance applications in general rely on predicting the future states of the ego-vehicle and surrounding vehicles to ensure safety. For example, time-to-collision (TTC) is a typical indicator used by active-safety applications to ensure safety distance between vehicles [13].

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Awais Javed.

To estimate TTC, the target vehicles' future states must be predicted. Normally the future states of a vehicle in a short period are assumed as constant. This simplifies the prediction process, by assuming that drivers tend to keep the states of their vehicles to be stable without abrupt changes. A driver who cannot do this becomes unpredictable. Therefore, one potential way of quantifying a driver's control skill is to measure the level to which he/she can maneuver the vehicle to keep the states unchanged. In this paper, this is termed *driving stability*. In general, the driving process can contain two types of phases, i.e. *smooth driving phases* (vehicles are controlled to have approximately fixed acceleration and steering angle), and *action phases* (vehicles are controlled to interact with surrounding vehicles, pedestrians, traffic signals, etc). Analyzing driving stability in both phases is of importance.

Naturalistic driving data have great potential in driving behavior analysis [14], since they contain the information of drivers' behaviors in natural driving conditions. One challenging task in data-driven driving behavior analysis is to extract features that can reflect maneuvering patterns from driving data. This requires properly tagging the data, i.e., partitioning the available driving data into segments, each of which corresponds to, ideally, a single objective maneuvering action. However, it is difficult to control the naturalistic driving data collection procedure to provide such label information. A number of existing works have proposed data partitioning solutions based on mathematics and signal processing techniques (e.g., see [6], [15]–[17]). But it is often difficult to guarantee each data segment to be the consequence of only a single complete maneuvering action. Using them to analyze driving behavior may bias the results.

In this paper, we investigate the method to analyze drivers' vehicle control skills without the need of partitioning naturalistic driving data into individual segments. We propose a novel algorithm to measure the driving stability level based on random matrix theory (RMT) [18], [19]. The data collected in the Safety Pilot Model Development (SPMD) program [20] are used to demonstrate the execution of our algorithm on a real-world ITS dataset. Specifically, the raw acceleration and steering angle data of each driver are first separated according to the speed level to roughly distinguish driving environments, and then respectively organized to matrices. From each data matrix, a series of sub-matrices are generated by sliding windows. The mean spectral radius (MSR) of every sub-matrix is calculated to reflect the distribution characteristics of the matrix entries. The differentiation of the MSR sequence, termed differential MSR (DMSR), is then taken. The concentration interval and outliers' dispersion level of the DMSR sequence are obtained as features to reflect the driving stability level of the driver, in the smooth driving and action phases respectively. Finally, based on these features, a density based spatial clustering of applications with noise (DBSCAN) clustering algorithm [21] is applied to partition all drivers into groups to summarize common driving stability characteristics.

The main contributions of our paper can be summarized as follows:

- We present a novel approach, based on RMT, to perform driving stability analysis using naturalistic driving data. Our method does not need to partition the data into segments according to individual maneuvering actions. Through synthetic data, we show that the output of the proposed algorithm can evaluate the statistical difference between the driving data and the data that would be generated by an ideally skillful driver who can maintain constant vehicle states. Hence features can be extracted from naturalistic driving data to reflect driver skills.
- We demonstrate the execution of the proposed method on a practical dataset produced by ITS technologies. Using the extracted features, a clustering algorithm is employed. The results show that the majority of drivers share a similar pattern of driving stability. But some drivers exhibit notable differences from others. Such observations can potentially be used to help better understand human drivers and facilitate further investigations.

The remainder of the paper is organized as follows. Section II reviews related works. Section III describes the dataset and pre-processing methods. Section IV presents the mathematical background of our RMT-based driving stability analysis algorithm. Section V discusses the results of executing the algorithm. Finally, Section VI concludes the paper.

Notations: Throughout the paper, $\mathbf{O}_{a\times b}$ denotes an $a \times b$ all-zero matrix. For matrix $\mathbf{X}$, $[\mathbf{X}]_i$ and $[\mathbf{X}]_{i,j}$ respectively denote the ith row and the element on the ith row and jth column. $\mathbf{X}^H$ denotes the conjugate transpose of $\mathbf{X}$.

II. RELATED WORKS

In addition to detecting the occasions that a driver does not fully concentrate on driving due to, e.g., sleepiness or distraction [22], driving behavior analysis normally refers to modeling drivers' habit of maneuvering vehicle to maintain the driving status (e.g., keeping the lane with constant speed) or interacting with traffic environment (e.g., turning, changing lane, or overtaking). Driver identification research works show that patterns extracted from driving data can be unique fingerprints. For instance, [23] extracts 12 features from turning maneuvers such that a random forest classifier can be used to distinguish drivers. Reference [24] estimates the distribution of accelerating maneuvers to identify drivers. Since different drivers exhibit different behavior characteristics, it is possible to distinguish abnormal drivers or even predict maneuvering intentions. For example, [25] uses random forest to classify drivers into low-risk, moderate-risk, and high-risk groups based on the transition probability between maneuvers. The safety evaluation method proposed in [26] measures driver risk by analyzing accelerating maneuvers using linear regression, decelerating maneuvers using Linde-Buzo-Gray algorithm, and turning maneuvers with kinematics analysis respectively. Reference [27] categorizes the performance of turning, accelerating

and decelerating maneuvers into four levels using support vector machine and topological anomaly detection. Reference [28] predicts driver maneuvering intentions using a hidden Markov model. Hence, driving behavior analysis has great potential in understanding human participants on the road.

Analyzing driving behaviors is a sophisticated task, since drivers' maneuvering pattern can be affected by many factors such as road environment (e.g., intersection or traffic congestion) and driver categories (e.g., age). For example, [29] applies kinematics analysis to extract features from driving data of several driving school instructors, general drivers, and elderly drivers when they drive passing urban intersections. It is shown that elderly drivers exhibit a common pattern that is different from young drivers. Reference [30] also proves this result by distinguishing elderly drivers with linear discriminant analysis. Using support vector machines and hidden Markov models, [31] suggests that the stopping maneuvers at intersections can be used to classify drivers into two categories, either compliant or violating. By clustering drivers' glance allocations, [32] points out that the behavior pattern at signalized intersections can be quite different from that at unsignalized intersections. Reference [33] shows that a post-congestion condition can cause drivers to be more aggressive.

Apart from simple control actions on the accelerator/brake pedals and steering wheel, a number of works have also studied relatively complex driving activities (such as car-following, lane changing, and overtaking) that can be deemed as the combination of multiple low-level maneuvers. Reference [34] models human driving patterns by estimating the distribution of parameters that describe the lane changing maneuver, to accelerate the verification of automated vehicles. Reference [35] uses one-class support vector machine to detect dangerous lane changing maneuvers. Reference [14] models the car-following maneuver with Gaussian kernel density estimation to discuss the impact of data volume on driving behavior analysis. Reference [36] identifies aggressive and cautious car-following maneuvers by analyzing the relationship between the vehicle's dynamics and the distance between the leading and the following cars with kinematics model. In [37], the car-following and approaching maneuvers are considered for driving behavior analysis. It uses the K-means algorithm to show that drivers can be partitioned into clusters according to their similarity levels of car-following time stability, prudence, conflict proneness, or skillfulness.

Most of the above works require that the driving data can be partitioned or tagged according to individual maneuvering actions. This can be challenging if driving behavior analysis is carried out on a large amount of naturalistic driving data. In addition, individual maneuvering actions may not fully reflect the control skills of drivers. In what follows, we present a method to evaluate the driving stability, based on RMT, to address these issues.

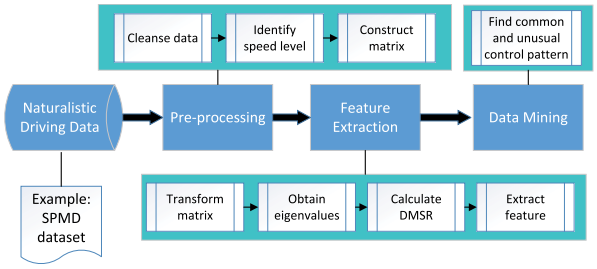

FIGURE 1. Flowchart of the proposed method.

III. DATASET AND PRE-PROCESSING

We apply our proposed method on an example naturalistic driving dataset. The execution follows a data-driven research framework which consists of four stages as shown in Fig. 1: data collection, pre-processing, feature extraction, and data mining. In this section, we introduce the dataset, explain the pre-processing stage, and then define driving stability.

A. NATURALISTIC DRIVING DATASET

Driving simulator is a commonly considered data source in driving behavior analysis (see, e.g., [12], [38], [39]). It is in general straightforward to extract different control operations so that individual target maneuvering actions can be analyzed. Dangerous and extreme road conditions can also be studied. The main drawback of using simulation data is that simulators cannot fully reflect the true driving condition. To collect driving data in practical traffic environments, carefully designed field-experiments are conducted in [23], [29], [40]. However, one potential issue of such methods is that drivers may be aware of the experiment purposes. Reference [41] reports that there is a significant difference between drivers' natural behavior and the behavior when they drive experiment vehicles with measurement devices, especially in the first 50 hours.

The data that potentially contain the best knowledge of drivers' true behaviors are the naturalistic driving data [41]–[44]. Naturalistic driving data are collected when the vehicles are driven under natural conditions, and data collection lasts a long period of time [14] (e.g, 12 to 13 months in the 100-Car Naturalistic Driving Study [41]). The long data collection time makes drivers oblivious to the data collection process so that the influence of data measurement devices on drivers' mental state and behaviors is minimized. One way of attaining such data is to equip vehicles with devices that can access the automobile bus through the OBD-II port, and acquire the measurements of various in-vehicle sensors [23], [24], [30]. Sensors of other devices, such as smartphones [27], can also be considered as a cheaper solution, at the cost of limited data types and measurement accuracy. Driving data can be continuously collected when the vehicle is driven and then analyzed after a certain period of time.

Large-scale naturalistic driving data collection is costly. The recent development of the concept of Internet of vehicles (IoV) [4] provides a feasible solution to this problem. IoV refers to using vehicle-to-anything (V2X) wireless

communication technologies to connect vehicles, roadside infrastructure, and other elements in the transportation system. Sensing data collected by various devices at different locations can be shared so that the environment awareness level of each individual vehicle can be significantly enhanced. A typical type of messages transmitted in IoV is the heartbeat basic safety message (BSM) [45]. A BSM contains the real-time status information of a vehicle and is broadcasted normally at frequency 10 Hz [45]. Currently, the dedicated short-range communication (DSRC) [46], LTE-V2X [47], and 5G technologies [48] are warmly discussed in both academia and industries as the V2X solutions. Large-scale experiments and field tests have also been conducted to verify the feasibility of IoV. It is expected that in the near future, all vehicles on the road will be equipped with sensing data collection and transmission devices. In addition to supporting real-time active-safety ITS applications, the BSMs stored in data centers may serve as the valuable naturalistic driving data, from a large number of drivers, in a variety of traffic environments, and over a long period of time.

To demonstrate the proposed algorithm, we use the data of the SPMD program [43] as our naturalistic driving dataset. The program was conducted to evaluate the performance of DSRC and communication-based active-safety applications. It was carried out in Michigan, USA, and lasted one and a half years since August 2012. A number of vehicles participated in the project and were equipped with data acquisition system with sampling frequency of 10 Hz. The data were organized in *trips*, each of which refers to one ignition cycle. The attributes include vehicle states (acceleration, steering, speed, etc.), road conditions (descriptions of lanes and intersections, etc.), and weather. The data have already been used for driving behavior analysis in, e.g., [14], [34].

Our method intends to summarize the driving skill of each driver from a large amount of driving data. Hence from the available dataset, we choose only the drivers who have sufficiently many (more than 40 for each speed scenario) long trip records (at least 6 minutes after pre-processing). This results in a total of 42 drivers, denoted by *driver*_01, *driver*_02, $\cdots$, *driver*_42. Finally, to demonstrate the unsupervised learning nature of the considered driving behavior analysis, we ignore the road condition and driving environment data, and use only the speed (m/s), acceleration (m/s^2), and steering angle (^{o}) readings. Acceleration is used for reflecting drivers' longitudinal control, steering angle is used for lateral control, and speed is considered to roughly infer driving condition, as explained in the next subsection.

B. DATA PRE-PROCESSING

The data pre-processing consists of the following steps.

1) DATA CLEANSING

The first step of data pre-processing is the detection and removal of missing and abnormal values through data cleansing [49]. The basic idea behind the proposed driving behavior analysis method is to quantify how much the statistical characteristics of the driving data matrix deviate from that in an ideal case. Occasional abnormal readings would not significantly affect the results. Sophisticated data cleansing algorithms are unnecessary. Hence we use simple interpolation to replace missing and abnormal readings.

If the speed data in a trip have continuous 0's, the vehicle might be immobile when these data were recorded. Since a driver does not exhibit any control skill in this case, the data should be removed from analysis. In our paper, when a speed segment of at least 10 continuous 0's (i.e., 1 second) is identified, the vehicle is believed to stay at the same location during that time. If the segment's length is relatively small, the associated data (including speed, acceleration, and steering angle) are removed. Otherwise, if a major portion of a trip has zero speed, the whole trip is discarded.

2) DATA SEPARATION

A driver may have different ways of taking the same maneuvering action in different driving conditions (e.g., different road types or traffic conditions). Identifying the driving stability level of each driver and comparing those of multiple drivers would be more meaningful under the same condition. Since the difficulty of controlling vehicle varies with speed, we consider the speed level to be a main factor that influences a drivers' behavior. Two different driving conditions are taken into consideration, i.e. low-speed scenario and high-speed scenario. To distinguish them, for each trip data, we find the median value of the speed. If the result is greater than a pre-defined threshold, V_{th}, the data of the trip are assumed to be collected in a high-speed scenario. A low-speed scenario is assumed if the median speed is lower than V_{th}. Considering that freeway and non-freeway are typical high-speed scenario and low-speed scenario respectively, the choice of V_{th} can be made according to the typical difference between the speed limitations of different road types. In our work, we take the speed limit policy in Michigan [50] as an example. There are two main types of road, freeway and non-freeway. The former has minimum speed limit of 55 miles/hour (i.e. 24 m/s). For non-freeway, different maximum speed limits are set for different levels of road, in general smaller than 45 miles/hour (i.e. 20 m/s). Therefore, we choose $V_{\text{th}} = 20$ m/s.

To further demonstrate the motivation of separating driving conditions into two scenarios, we apply a data visualization method, which was originally designed for discovering the difference between human and robot users in social media [51], to display the states of vehicle motion (speed, acceleration, and yaw rate) simultaneously. Fig. 2 illustrates the plot of three example drivers in our dataset, i.e., *driver*_15, *driver*_30 and *driver*_35. For each driver, we randomly select three hours of data from the low-speed scenario, and three hours from the high-speed scenario. The data of each hour are plotted as a plate, with a circumference, a kernel, and multiple threads pointing outward. The circumference represents the time information. For ease of illustration, we use the average data value of each second to summarize the driving data of that second. The one-hour data recording starts from the top center (i.e., 12 o'clock) and proceeds clockwise, with a total

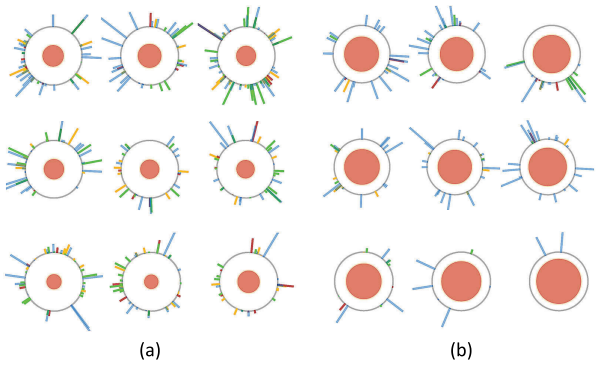

FIGURE 2. Driving behaviors in (a) low-speed scenario and (b) high-speed scenario.

of 3600 sample instants. The plate kernel represents the speed information: its area denotes the average speed of the hour.

The threads represent the driving actions. Take acceleration in the longitudinal direction as an example. When the difference between the (average) acceleration readings in two consecutive seconds is larger than a certain threshold, it is believed that the driver took a notable operation on the accelerator pedal. A green thread is plotted at the time instant to denote this action, and the length of the thread is proportional to the difference between the incremental readings and the threshold. The same approach applies to deceleration (blue threads), left turn (yellow threads), and right turn (red threads).

Each row in Fig. 2 shows one driver's behaviors. The left hand side (LHS) data are collected from the low-speed scenario and the data on the right hand side (RHS) are from the high-speed scenario. Clearly each driver has a similar pattern when the average speed is similar. But comparing the two scenarios, it is seen that more frequent actions were taken by the same driver in the low-speed scenario. Different drivers can also have quite diverse patterns. Thus, it is reasonable to carry out the data separation step so that driving behaviors can be individually analyzed in each driving scenario.

3) DATA TRANSFORMATION

After separating each driver's data according to the speed scenario, we organize the driving data to matrix forms. Specifically, four matrices for each driver are generated, respectively denoted by $\mathbf{A}_\mathrm{l}$, $\mathbf{A}_\mathrm{h}$, $\mathbf{S}_\mathrm{l}$, and $\mathbf{S}_\mathrm{h}$. We randomly choose M_l trips from each driver's data in the low-speed driving scenario, and M_h trips from the high-speed scenario. Let T_l and T_h be two sufficiently large integers. The matrix $\mathbf{A}_\mathrm{l}$ is formed by the acceleration data of a driver in the low-speed scenario: Each row of $\mathbf{A}_\mathrm{l}$ is a segment of T_l acceleration readings (with unit m/s^2) chosen from the middle part of each trip (to avoid data of the starting-up and full-stop operations). Similarly, matrix $\mathbf{A}_\mathrm{h}$ is formed by the acceleration data in the high-speed scenario. Each row of $\mathbf{A}_\mathrm{h}$ is a segment of T_h acceleration readings in each of the M_h trips. $\mathbf{S}_\mathrm{l}$ and $\mathbf{S}_\mathrm{h}$ are steering angle data (with unit degree) matrices in the low- and high-speed scenarios respectively. The former consists of M_l trips, each of which has T_l readings. The latter consists of M_h trips, each of which has T_h readings. In general, M_l and M_h (resp. T_l and T_h) can be different. For ease of presentation, we choose $M = M_l = M_h = 40$ and $T = T_l = T_h = 3000$. The proposed driving stability analysis algorithm, to be introduced in the next section, is applied to each of the four matrices. Therefore, we measure the control skill of a driver in longitudinal and lateral directions, and in low-speed and high-speed driving scenarios, respectively. Let $\mathcal{D} = \{\mathbf{A}_\mathrm{l}, \mathbf{A}_\mathrm{h}, \mathbf{S}_\mathrm{l}, \mathbf{S}_\mathrm{h}\}$ denote the set of data matrices. Executing the proposed method on only one matrix $\mathbf{X} \in \mathcal{D}$ is discussed in the following sections.

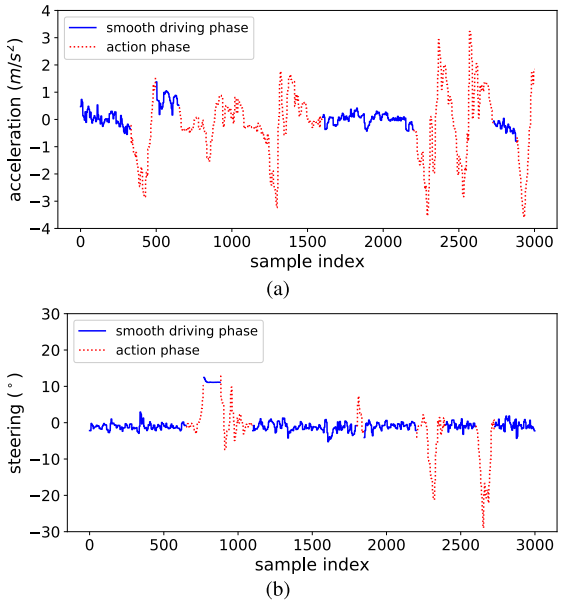

FIGURE 3. An example of naturalistic driving data in (a) acceleration (m/s^2) and (b) steering angle (°).

C. DRIVING STABILITY

Before introducing the feature extraction stage, we present the definition of driving stability considered in our paper. In general, a human driver or a driving assistance system perceives the surrounding driving environment and makes maneuvering plans according to the observations of the movements of the ego-vehicle and other vehicles, road condition, and traffic rule. If a vehicle is driven to move in a relatively stable way without abrupt changes, it is easier to predict and can be considered to be safer. Hence the ability of controlling vehicle in this way is deemed as *driving stability*.

Fig. 3 displays a sample row of the acceleration data matrix and a sample row of the steering angle data matrix. It is seen that a typical driving trip can roughly be divided into two types of phases. The first corresponds to the period that the data vary around a certain constant. This is termed *smooth driving phase*. In this type of driving phase, a driver, without being affected by surrounding traffic environment, intends to maintain the same vehicle state in either the longitudinal or lateral direction. We consider an ideally skillful driver to be one who can keep the longitudinal force (represented by acceleration) and lateral force (represented by steering angle) on vehicle to be constant. In practice, the forces in

both directions change continuously, the level of which may reveal the driver's skill. If the data change around an expected value with relatively high variation, the driving stability level is considered to be low, and the vehicle state is hard to predict.

Certainly, drivers have to interact with traffic environment and carry out intentional operations to change the vehicle state. This corresponds to the second type of driving phase, termed *action phase*. It can be deemed as the transition between two smooth driving phases. A more skillful driver is able to maneuver vehicle and complete the transitions prudently. If the transitions are often carried out abruptly with rapid oscillations, the driver's skill level in vehicle control is considered to be relatively low. Hence, in action phases, the average slope of the data statistics changes can be used to reflect the control skill.

Assume that vehicle kinetic data collection using in-vehicle sensors is always subject to measurement noise. That is, the driving data matrix $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{F}+\mathbf{N}$, in which $\mathbf{F}$ denotes the true force on the vehicle taken by the driver, and $\mathbf{N}$ represents random noise. The measurement noise of each trip is assumed to be a stationary Gaussian process. For an ideal driver (who can keep the true force in smooth driving phases to be constant and achieve extremely small data statistics change slopes, i.e. 0, in action phases), every row of $\mathbf{F}$ is a constant. The samples in each row of $\mathbf{X}$ are identically distributed, and there is no correlation between any two rows. We use $\overline{\mathbf{X}}$ to denote the row-normalized version of matrix $\mathbf{X}$, i.e., the ith row vector $[\overline{\mathbf{X}}]_i$ is attained by normalizing $[\mathbf{X}]_i$ to consist of entries with zero-mean and unit-variance. Then all entries of $\overline{\mathbf{X}}$ are independent and identically distributed (i.i.d.). In practice, due to the driver's skill and complex driving environment, it is hard to maintain fixed acceleration and steering angle, even for a limited period of time. This causes the statistical characteristics of $\mathbf{X}$ to be different from that in the ideal case. Therefore, we consider using an i.i.d. matrix as a theoretic benchmark data matrix (which is not practically achievable) and measuring how much the intrinsic characteristics of $\overline{\mathbf{X}}$ generated by the driver deviate from the ideal case as the driving stability level.

If the data segments corresponding to the two phases can be accurately separated, one possible approach of evaluating the driving stability in the smooth driving phases is to measure the average variances of the acceleration or steering angle data of multiple segments. That in the action phases may also be evaluated by finding the average data changing slopes. However, distinguishing the two phases from naturalistic driving data is involved. In the following sections, we propose our RMT-based algorithm to extract representative features from the data matrix $\mathbf{X}$ without the demand for driving phase separation.

IV. FEATURE EXTRACTION THROUGH RMT

In this section, we present an algorithm to extract features that can describe driving stability. Since the row-normalized data matrix of an ideally skillful driver has i.i.d. entries, we use the statistical difference between the normalized matrix and an i.i.d. random matrix to reflect the stability level of a real driver. To this end, we first follow the RMT and present indicators to reflect the statistical characteristics of an i.i.d. random matrix.

A. RING LAW FOR i.i.d. RANDOM MATRIX

In RMT, the *ring law* [52] is a common property of i.i.d. non-Hermitian random matrices. To explain the concept, we consider l independent $m \times n$ non-Hermitian random matrices $\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_l$. All their entries, $[\mathbf{R}_s]_{i,j}$, $\forall s \in \{1, 2, \cdots, l\}, i \in \{1, 2, \cdots, m\}, j \in \{1, 2, \cdots, n\}$, are i.i.d. with zero mean and unit variance. Define the *singular value equivalent matrix* of $\mathbf{R}_s$ as [53].

$$\widetilde{\mathbf{R}}_s = \sqrt{\mathbf{R}_s \mathbf{R}_s^H}\mathbf{U}, \quad s \in \{1, 2, \cdots, l\}, \tag{1}$$

where U is an $m \times m$ Haar-unitary matrix. Let

$$\mathbf{R} = \prod_{s=1}^{l} \widetilde{\mathbf{R}}_s. \tag{2}$$

Denote the standard deviation of $[\mathbf{R}]_i$, the ith row of $\mathbf{R}$, to be σ_i. We can define an $m \times m$ matrix $\hat{\mathbf{R}}$ such that the relationship between the ith row of $\hat{\mathbf{R}}$ and the ith row of $\mathbf{R}$ is

$$[\hat{\mathbf{R}}]_i = \frac{1}{\sqrt{m}} \frac{[\mathbf{R}]_i}{\sigma_i}, \quad i \in \{1, 2, \cdots, m\}. \tag{3}$$

Clearly, the variance of the entries of $\hat{\mathbf{R}}$ is $\frac{1}{m}$.

The matrix $\hat{\mathbf{R}}$ has m (complex) eigenvalues. Since $\hat{\mathbf{R}}$ is a random matrix, the eigenvalues are also random. If $m \to \infty$, $n \to \infty$, and $\lim_{m\to\infty} \frac{m}{n} = c$ is a constant ($0 < c \leq 1$), the probability density function (PDF) of the m eigenvalues converges to a *limiting spectral density* (LSD) [52]:

$$f_{\hat{\mathbf{R}}}(\lambda) = \begin{cases} \dfrac{1}{\pi c l} |\lambda|^{\frac{2}{l}-2}, & (1-c)^{\frac{l}{2}} \leqslant |\lambda| \leqslant 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This is the ring law, which says that on the complex plane, the eigenvalues are confined within a ring defined by an outer circle with unit radius and an inner circle with radius $(1-c)^{\frac{l}{2}}$. The property applies to any $l \geq 1$.

For instance, we generate a 40×80 synthetic random matrix $\mathbf{Y}_1$ with Gaussian i.i.d. random entries, row-normalize it to $\overline{\mathbf{Y}}_1$, and then calculate the matrix $\hat{\mathbf{R}}$ following (1)-(3) by setting $l = 1$ and $\mathbf{R}_1 = \overline{\mathbf{Y}}_1$. The eigenvalues of $\hat{\mathbf{R}}$ are plotted on the complex plane as purple dots in Fig. 4(a), together with the outer circle (with radius 1) and the inner circle (with radius $(\frac{1}{2})^{\frac{1}{2}} = 0.707$). Since the matrix size is sufficiently large, almost all the 40 eigenvalues locate within the ring belt.

If the condition that all entries in the matrices $\mathbf{R}_1$, $\mathbf{R}_2, \cdots, \mathbf{R}_l$ are i.i.d. does not hold, the ring law would be violated. The eigenvalues tend to collapse to the origin of the complex plane. For example, consider a 40×80 all-zero matrix $\mathbf{Y} = \mathbf{O}_{40\times 80}$. We randomly select 10 rows from the matrix. For the ith selected row, one position index j_i ($20 \leq j_i \leq 40$) is randomly selected. Then we set the

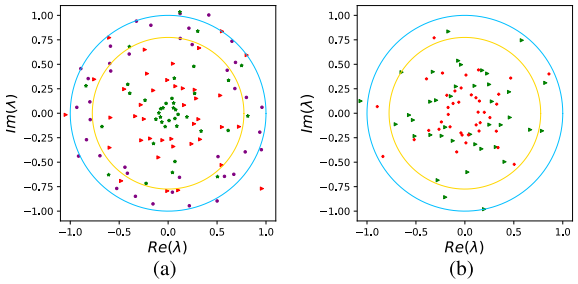

FIGURE 4. Eigenvalues of (a) three synthetic random matrices, and (b) two naturalistic driving data matrices, with the ring belt defined by the ring law.

elements $[\mathbf{Y}]_{i,j} = j - j_i$ for $j \in \{j_i+1, j_i+2, \cdots, j_i+10\}$, and $[\mathbf{Y}]_{i,j} = 10$ for $j \in \{j_i+11, j_i+12, \cdots, 80\}$. The resulting matrix is added to $\overline{\mathbf{Y}}_1$ to produce a new synthetic matrix $\mathbf{Y}_2$, whose entries are now non-i.i.d. We normalize $\mathbf{Y}_2$ to $\overline{\mathbf{Y}}_2$, and follow the steps of generating matrix $\hat{\mathbf{R}}$ by setting $l = 1$ and $\mathbf{R}_1 = \overline{\mathbf{Y}}_2$. The eigenvalues of $\hat{\mathbf{R}}$ are plotted as red triangles in Fig. 4(a), which clearly shows the violation of the ring law. If more than 10 rows of $\mathbf{Y}_2$ have changes of the entry statistics, the eigenvalues would tend to be even closer to the origin.

Now we again randomly select 10 rows from $\mathbf{Y} = \mathbf{O}_{40\times 80}$. For each of these rows, one position index j_i ($20 \le j_i \le 40$) is chosen. All elements in this row to the RHS of the position $j_i + 10$ are set to 30, i.e., $[\mathbf{Y}]_{i,j} = 30$ for $j \in \{j_i+11, \cdots, j_i+80\}$. A linear increase is set between columns j_i+1 and j_i+10, i.e., $[\mathbf{Y}]_{i,j} = 3(j - j_i)$ for $j \in \{j_i + 1, \cdots, j_i + 10\}$. The resulting matrix is added to $\overline{\mathbf{Y}}_1$ to produce another synthetic matrix $\mathbf{Y}_3$. Compared with $\mathbf{Y}_2$, the change of the statistics of the entries occurs to a greater extent. Generate the matrix $\hat{\mathbf{R}}$ by setting $l = 1$ and $\mathbf{R}_1 = \overline{\mathbf{Y}}_3$ and plot eigenvalues of $\hat{\mathbf{R}}$ as green asterisks in Fig. 4(a). It is seen that being more different from the original i.i.d. matrix leads to characteristics of eigenvalues farther away from that described by the ring law. Therefore, comparing the statistical behaviors of the eigenvalues of matrix $\hat{\mathbf{R}}$ can reflect how much a matrix is different from an i.i.d. matrix.

Finally, we randomly choose two drivers from our dataset. For each driver, a 40×80 sub-matrix is extracted from the acceleration data matrix in the low-speed driving scenario. Following (1)-(3), we calculate the matrix $\hat{\mathbf{R}}$ by setting $l = 1$ and $\mathbf{R}_1$ to be the row-normalized version of the sub-matrix. The eigenvalues of $\hat{\mathbf{R}}$ are plotted in Fig. 4(b). Clearly, their behaviors notably violate the ring law.

B. LINEAR EIGENVALUE STATISTIC (LES)

To statistically summarize the random behaviors of the m eigenvalues of matrix $\hat{\mathbf{R}}$, we define the LES as [54]:

$$p_{\text{LES}} = \sum_{i=1}^{m} \varphi(\lambda_i), \tag{5}$$

where $\varphi(\lambda_i)$ is a continuous function of the ith eigenvalue λ_i. p_{LES} is a statistic of the eigenvalues and is proved to satisfy the law of large numbers and the central limit theorem [54]. According to the law of large numbers, when $m \to \infty$, $\frac{1}{m}p_{\text{LES}}$ converges in probability to the expectation of $\varphi(\lambda)$:

$$\lim_{m\to\infty} \frac{1}{m} p_{\text{LES}} = \int \varphi(\lambda) f_{\hat{\mathbf{R}}}(\lambda) \mathrm{d}\lambda, \tag{6}$$

where $f_{\hat{\mathbf{R}}}(\lambda)$ is the LSD in (4). Based on the central limit theorem, [55] proves that, when the entries in matrix $\mathbf{R}_s$ are i.i.d., the samples of p_{LES} have a small confidence interval.

One way of defining the function $\varphi(\lambda_i)$ is to set $\varphi(\lambda_i) = \frac{|\lambda_i|}{m}$. The resulting LES is termed *mean spectral radius* (MSR) [19] and is denoted by p_{MSR}. It calculates the mean distance between the origin of the complex plane and the eigenvalues:

$$p_{\text{MSR}} = \frac{1}{m} \sum_{i=1}^{m} |\lambda_i|. \tag{7}$$

When the PDF of eigenvalues converges to LSD in (4), we can obtain the theoretical value of p_{MSR} using (4) and (6) as:

$$\begin{aligned} p^*_{\text{MSR}} &= \lim_{m\to\infty} \frac{1}{m}|\lambda_i| = \int_0^{2\pi} \int_{(1-c)^{l/2}}^{1} r \cdot \frac{1}{\pi c l} r^{\frac{2}{l}-2} \cdot r \mathrm{d}r\mathrm{d}\theta \\ &= \frac{2}{3cl} - \frac{2(1-c)^{\frac{3}{2}}}{3cl}. \end{aligned} \tag{8}$$

For instance, when $l = 1$ and $c = \frac{m}{n} = 0.5$, we have $p^*_{\text{MSR}} = 0.8619$. MSR describes the behaviors of the random eigenvalues using a single value. Comparing the MSR of a matrix with p^*_{MSR} provides a measurement of the difference between the matrix and an i.i.d. random matrix.

C. DIFFERENTIAL MSR (DMSR)

To describe the changes of statistics of each driver's driving data matrices, we follow [19] and separate the $M \times T$ data matrix $\mathbf{X} \in \mathcal{D}$ into a series of $M \times N$ ($M \le N \le T$) sub-matrices, using $T \times N$ sliding-window matrices $\mathbf{W}^{[k]} = [\mathbf{O}_{N\times(k-N)}, \mathbf{I}_N, \mathbf{O}_{N\times(T-k)}]^T$, for $1 \le k \le T - N + 1$:

$$\mathbf{X}^{[k]} = \mathbf{X} \times \mathbf{W}^{[k]}, \quad k = \{1, 2, \cdots, T - N + 1\}. \tag{9}$$

For each $\mathbf{X}^{[k]}$, we can find its row-normalized matrix $\overline{\mathbf{X}}^{[k]}$ and then follow (1)-(3) by setting $l = 1$ and $\mathbf{R}_1 = \overline{\mathbf{X}}^{[k]}$ to attain matrix $\hat{\mathbf{R}}$. Since the entries of $\overline{\mathbf{X}}^{[k]}$ are non-i.i.d., the eigenvalues of $\hat{\mathbf{R}}$ tend to collapse to the origin of the complex plane. Denote the resulting MSR by $p^{[k]}_{\text{MSR}}$. For each k, $p^{[k]}_{\text{MSR}}$ is a random value that is almost always less than p^*_{MSR}. The value $p^{[k]}_{\text{MSR}}$ is likely to be smaller when the data sub-matrix $\overline{\mathbf{X}}^{[k]}$ is more different from an i.i.d. matrix.

We use an example to show that the sequence of $p^{[1]}_{\text{MSR}}$, $p^{[2]}_{\text{MSR}}$, $\cdots$, $p^{[T-N+1]}_{\text{MSR}}$ can be used to measure the strength of driving operations. Let $\mathbf{N}_0$ be a 40×3000 synthetic i.i.d. noise matrix with standard Gaussian entries. Set $\mathbf{Y} = \mathbf{O}_{40\times 3000}$ to be an all-zero matrix and randomly select two rows. For these rows, we further set $[\mathbf{Y}]_{i,j} = 0$ for columns $j \in \{1, 2, \cdots, 700\}$, $[\mathbf{Y}]_{i,j} = 0.4(j - 700)$ for $j \in \{701, 702, \cdots, 800\}$, $[\mathbf{Y}]_{i,j} = 40$ for $j \in \{801, 802, \cdots, 1400\}$, $[\mathbf{Y}]_{i,j} = 40 - 0.2(j - 1400)$ for $j \in \{1401, 1402, \cdots, 1600\}$, $[\mathbf{Y}]_{i,j} = 0$ for $j \in \{1601, 1602, \cdots, 2300\}$, $[\mathbf{Y}]_{i,j} = 0.2(j - 2300)$

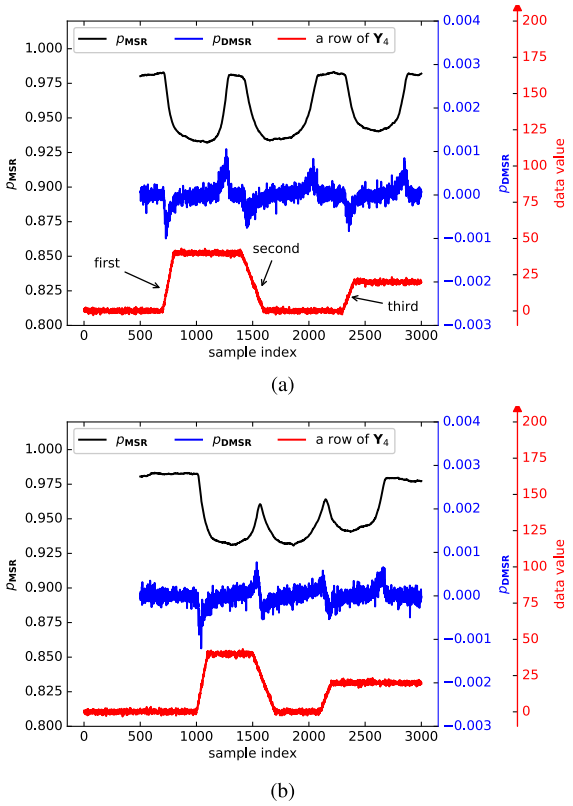

FIGURE 5. MSR and DMSR sequence for synthetic matrices (a) $\mathbf{Y}_4$, and (b) $\mathbf{Y}_5$.

for $j \in \{2301, 2302, \cdots, 2400\}$, and $[\mathbf{Y}]_{i,j} = 20$ for $j \in \{2401, 2402, \cdots, 3000\}$. Define $\mathbf{Y}_4 = \mathbf{Y} + \mathbf{N}_0$, and display one of the selected rows of $\mathbf{Y}_4$ in Fig. 5(a) (red curve). We use these rows to mimic three driving operations, i.e., change of acceleration/steering angle from one constant value to another. (The changes in the two rows lead to correlation and significant deviation from i.i.d. matrix.) The first two changes have the same magnitude, but the former is more rapid (with a larger absolute value of the slope). The third change has the same slope as the second, with a smaller magnitude.

Now, we set each sliding-window matrix $\mathbf{W}^{[k]}$ to be a 3000×500 matrix, for $k \in \{1, 2, \cdots, 2501\}$. Multiplying $\mathbf{W}^{[k]}$ by matrix $\mathbf{Y}_4$ generates a total of 2501 sub-matrices with dimension 40×500, denoted by $\mathbf{Y}_4^{[1]}$, $\mathbf{Y}_4^{[2]}$, $\cdots$, $\mathbf{Y}_4^{[2501]}$, respectively. For each value of k, we derive the row-normalized matrix $\overline{\mathbf{Y}}_4^{[k]}$, and follow (1)-(3) using $l = 1$ and $\mathbf{R}_1 = \overline{\mathbf{Y}}_4^{[k]}$ to find the MSR, $p_{\text{MSR}}^{[k]}$, of the resulting matrix $\hat{\mathbf{R}}$. The MSR sequence is plotted in Fig. 5(a) (black curve), where the kth MSR and the $(k+499)$th element of the selected row (i.e., the last entry selected by the window) are aligned.

It can be seen that when k increases from 1 to 200, the values of $p_{\text{MSR}}^{[k]}$ fluctuate around a constant. This is because all the normalized sub-matrices $\overline{\mathbf{Y}}_4^{[k]}$ are i.i.d. matrices. The MSR is then a random variable with expected value $p_{\text{MSR}}^* = 0.9797$, attained using (8) with $l = 1$ and $c = \frac{40}{500} = 0.08$, and a small confidence interval [55]. When k exceeds 200, the sub-matrix $\mathbf{Y}_4^{[k]}$ starts to include the data corresponding to the first change. Due to the data correlation and deviation from an i.i.d. matrix, the eigenvalues begin to shrink towards the origin of the complex plane. The resulting MSR deviates from p_{MSR}^*. As k increases, more entries within $\overline{\mathbf{Y}}_4^{[k]}$ exhibit different statistical characteristics, and $\overline{\mathbf{Y}}_4^{[k]}$ is more different from an i.i.d. matrix. The value of $p_{\text{MSR}}^{[k]}$ tends to further reduce, until the change of data statistics occurs in the middle of the windowed sub-matrix. Afterwards, $p_{\text{MSR}}^{[k]}$ starts increasing. When all the data regarding the first change leave the sliding window, i.e., $k > 800$, the entries of $\overline{\mathbf{Y}}_4^{[k]}$ are i.i.d.. The values of $p_{\text{MSR}}^{[k]}$ again variate slightly around $p_{\text{MSR}}^* = 0.9797$. These lead to the first U-shape as shown in Fig. 5(a). Similarly, when k continues to increase, the sliding window passes the other two changes, which results in two more U shapes in the figure.

From Fig. 5(a), it is seen that different types of data statistics changes can lead to different U-shapes of MSR. The first and second changes have the same magnitude. The associated U-shapes have the similar depth. But since the second change has a smaller slope, it demands more time to complete the change. This causes a greater width of the U-shape. The third change has a smaller magnitude compared with the second change. A smaller depth of the U-shape is observed. Therefore, by comparing the depth and width of the U-shapes of the MSR sequence, one can roughly measure the magnitude and the time instant of statistical changes in a large data matrix [19].

However, such a method demands the U shapes of MSR to be isolated so that their depth and width can be identified. This means that each sub-matrix contains only a single change of data statistics. Although this is possible, through a proper selection of the window size, for data collected from certain domains such as smart grid [19], it is difficult to satisfy the requirement with naturalistic driving data due to frequent operations of drivers. For instance, we generate another synthetic matrix $\mathbf{Y}_5$, which has the same three data changes as $\mathbf{Y}_4$. But the changes occur closer to each other, as shown in Fig. 5(b), so that two changes can be included in the same sliding window. The figure shows that this leads to overlapped U shapes. Measuring the depth and width of each U shape becomes hard to accomplish, especially if the changes of data statistics appear frequently in different rows.

To address this issue, based on MSR, we propose a new parameter, termed *differential mean spectral radius* (DMSR):

$$p_{\text{DMSR}}^{[k]} = p_{\text{MSR}}^{[k+1]} - p_{\text{MSR}}^{[k]}, \quad k \in \{1, 2, \cdots, T - N\}. \tag{10}$$

Essentially, the DMSR sequence measures the descending or ascending speed of the U-shapes of the MSR. For two U-shapes with the similar depth, the one that has a smaller width is likely to have a steeper edge, i.e., some large values of $p_{\text{DMSR}}^{[k]}$. For U-shapes with the similar width, a greater depth is more likely to cause some large values of DMSR.

Fig. 5(a) illustrates the DMSR sequence of $\mathbf{Y}_4$ (blue curve). It is seen that DMSR can clearly reflect the characteristics of the U shapes of the MSR sequence. The DMSR exhibits

a random behavior. But most realizations locate within a certain interval, with frequent occurrence of large values corresponding to the edges of the U shapes. In addition, deeper and narrower U shapes of MSR sequence (e.g., caused by the first change) lead to DMSR values more significantly deviating from common values. Since the DMSR sequence exploits the information contained in only a small number of MSR samples to measure the changing level of the data statistics, identifying the complete U shape is not necessary. In Fig. 5(b), we also show the DMSR of $\mathbf{Y}_5$. The pattern shown in Fig. 5(a) can still be observed, even though the U-shapes of the MSR have overlaps.

D. CASE STUDIES

To obtain features that reflect driving stability from the DMSR sequence, we first use two case studies to explain how the DMSR sequence contains driving stability information in the two driving phases respectively. The data used in our paper were collected from an IoV research project and thus do not provide any label information on whether a driver has a higher or lower driving stability level. Therefore, the case studies are based on synthetic data that simulate the characteristics of driving data. In a smooth driving phase, a driver operates the vehicle in order to maintain a constant state of the vehicle. Such operations in general occur frequently, with limited strength. Consequently, each sub-matrix attained using (9) has entries with multiple small changes in data statistics. Drivers with good driving skills tend to generate data matrices with less significant statistical changes compared with drivers with poor skills. In an action phase, a driver makes control operations that lead to changes of statistics with much larger magnitude than those in smooth driving phases. A skillful driver tends to make the operations prudently and steadily. In what follows, we use simulated data to explore the influence of these two kinds of changes on DMSR respectively.

1) SMOOTH DRIVING PHASES SIMULATION

Generate a 40×3000 synthetic control data matrix $\mathbf{Y}$. For each row i ($i \in \{1, \cdots, 40\}$), let the data have a step change with random magnitude for every 5 elements to simulate frequent operations in the smooth driving phases. Specifically, for each row i, we sample 600 random values from a uniform distribution between -0.5 and 0.5. Let the $(5(j-1)+1)$th to the $(5j)$th elements, i.e., $[\mathbf{Y}]_{i,5(j-1)+1}, \cdots, [\mathbf{Y}]_{i,5j}$, be the jth sample. Finally, we set the synthetic driving data matrix $\mathbf{Y}_6 = \mathbf{Y} + \mathbf{N}_1$, where $\mathbf{N}_1$ is a Gaussian noise matrix with mean 0 and standard deviation 0.5. By this means, we simulate frequent changes of data statistics around constants.

Set the sliding-window matrices $\mathbf{W}^{[k]}$ to be 3000×80 matrices. Apply (1)-(3) and (7) to the row-normalized sub-matrices of $\mathbf{Y}_6$ to obtain the MSR sequence, and apply (10) to obtain the DMSR sequence, as plotted in Fig. 6. Due to the frequent changes of data statistics in each sub-matrix, the U-shapes of MSR heavily overlap and thus no complete U shape is observable. However, based on the results shown in Fig. 6, the behavior of DMSR can still reflect the data

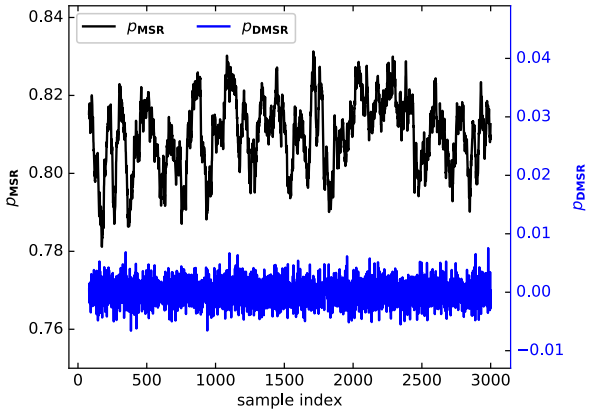

FIGURE 6. MSR and DMSR sequence for synthetic data matrix $\mathbf{Y}_6$.

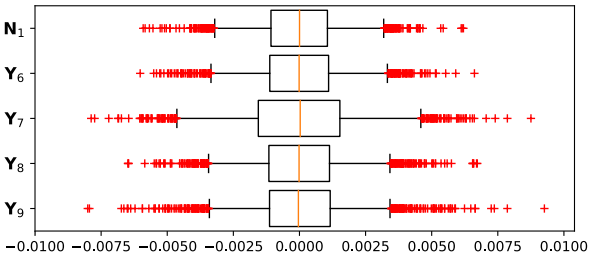

FIGURE 7. DMSR comparison for synthetic data matrices.

statistics in the original data matrix $\mathbf{Y}_6$. The frequent changes in matrix $\mathbf{Y}$ cause a large number of $p_{\text{DMSR}}^{[k]}$ to be relatively far from the center value 0. In other words, the observed DMSR values would be dispersed compared with those attained from ideal i.i.d. matrices. Larger changes of statistics would cause even dispersed behavior of the DMSR.

To show this, we generate another data matrix $\mathbf{Y}_7$ following the same way as $\mathbf{Y}_6$, except that the elements in the control data matrix $\mathbf{Y}$ are sampled from a uniform distribution between -1 and 1. This reflects both larger magnitude and speed of data statistical changes, compared with $\mathbf{Y}_6$. In Fig. 7, we use box plots to visualize the dispersion characteristics of the DMSR sequence corresponding to the two synthetic data matrices. The box size is determined by the interquartile range (IQR). The boundaries are the positions whose distance to the nearby box edge equals the IQR. Data samples outside the boundaries are treated as outliers. Clearly, the concentration interval size, i.e. the box size, of the DMSR sequence corresponding to $\mathbf{Y}_6$ is smaller than that corresponding to $\mathbf{Y}_7$. Therefore, using the box size of the box plot of DMSR can help reflect the driving stability levels in smooth driving phases.

2) ACTION PHASES SIMULATION

To explore the influence of driving operations in action phases on the DMSR, we further study two synthetic data matrices. Let $\mathbf{Z}$ initially be an all-zero matrix $\mathbf{O}_{40\times3000}$. For each row i ($i \in \{1, 2, \cdots, 40\}$) of $\mathbf{Z}$, we first randomly find a position index r_i ($1000 \leq r_i \leq 2000$). The next 50 elements on the right side of $[\mathbf{Z}]_{i,r_i}$, i.e., $[\mathbf{Z}]_{i,r_i+1}, \cdots, [\mathbf{Z}]_{i,r_i+50}$, then increase or decrease linearly from 0. The absolute values of

slopes are randomly sampled from a Gaussian distribution with mean 0.2 and standard deviation 0.1. The remaining entries $[\mathbf{Z}]_{i,r_i+j}$, $\forall j \in \{51, \cdots, 3000 - r_i\}$ are set to be the same as $[\mathbf{Z}]_{i,r_i+50}$. By this means, each change simulates a relatively large control operation. Finally, the synthetic data matrix $\mathbf{Y}_8$ is attained by $[\mathbf{Y}_8]_{i,j} = [\mathbf{Y}_6]_{i,j} + [\mathbf{Z}]_{i,j}$ when $j \leq r_i$ or $j > r_i + 50$, and $[\mathbf{Y}_8]_{i,j} = [\mathbf{Z}]_{i,j} + [\mathbf{N}_1]_{i,j}$ when $r_i < j \leq r_i + 50$. The data matrix contains both long, significant changes and frequent, small changes of data statistics compared with an i.i.d. matrix.

The second synthetic data matrix $\mathbf{Y}_9$, is generated similarly, except that the absolute values of slopes of the data changes in $\mathbf{Z}$ are sampled from a Gaussian distribution with mean 0.4. Fig. 7 also shows the box plots of the DMSR sequence corresponding to $\mathbf{Y}_8$ and $\mathbf{Y}_9$. We can notice that their box sizes are almost the same. This is because both matrices are generated using $\mathbf{Y}_6$, and have the same level of frequent changes. However, the DMSR sequence of $\mathbf{Y}_9$ has more outliers farther away from the box center. This is in line with the observation made in Fig. 5: steeper changes of data statistics lead to deeper U shapes of MSR and possible DMSR values significantly dispersing from the center. Therefore, one can use the dispersion of outliers to reflect the driving stability level in action phases.

E. DRIVING STABILITY FEATURE EXTRACTION ALGORITHM

As shown in the above case studies, we can use the concentration interval and outliers' dispersion level to summarize the statistics of DMSR sequence, which reflects the driving stability in two driving phases respectively. To facilitate a quantitative analysis, denote q_a the ath percentile of the DMSR sequence. For a certain value of $0 < a < 50$, the feature for representing the concentration level of the DMSR sequence can be defined by

$$C_{\text{DMSR}} = q_{100-a} - q_a. \tag{11}$$

In this paper, we let $a = 25$ and C_{DMSR} is the IQR (the box size) of the box plot in Fig. 7.

Outliers can be defined as those DMSR samples greater than $(\alpha+1)q_{100-a} - \alpha q_a$ or smaller than $(\alpha+1)q_a - \alpha q_{100-a}$ for some $\alpha > 0$. In our paper, we choose $\alpha = 1$ so that DMSR samples whose distance from the median value, denoted by $\tilde{p}_{\text{DMSR}}$, is much lager than C_{DMSR}, are deemed to be outliers. Denote the set of outliers by $\mathcal{P}_{\text{out}}$. The feature to represent the dispersion level of the outliers can be defined as the average distance between the outliers and $\tilde{p}_{\text{DMSR}}$, i.e.,

$$O_{\text{DMSR}} = \frac{1}{|\mathcal{P}_{\text{out}}|} \sum_{p \in \mathcal{P}_{\text{out}}} (p - \tilde{p}_{\text{DMSR}}), \tag{12}$$

where $|\mathcal{P}_{\text{out}}|$ is the cardinality of $\mathcal{P}_{\text{out}}$.

Now, for each driver, we can calculate the features C_{DMSR} and O_{DMSR}. Small values of them show that the driving data matrix (after normalization) is similar to an i.i.d. matrix, which is attained by perfect skill of maintaining constant vehicle state. Larger values of C_{DMSR} and O_{DMSR} imply that the data matrix is more different from an i.i.d. matrix.

Algorithm 1 Extraction of C_{DMSR} and O_{DMSR}

Input: Driving data matrix $\mathbf{X}$

For $k = 1 : T - N + 1$ **do**
  Obtain sub-matrix $\mathbf{X}^{[k]}$ using (9)
  Calculate row-normalized matrix $\overline{\mathbf{X}}^{[k]}$ from $\mathbf{X}^{[k]}$
  Set $l = 1$, $\mathbf{R}_1 = \overline{\mathbf{X}}^{[k]}$, and calculate $\tilde{\mathbf{R}}_1$ using (1)
  Calculate $\mathbf{R}$ using (2)
  Calculate $\hat{\mathbf{R}}$ using (3) and find its eigenvalues
  Calculate $p_{\text{MSR}}^{[k]}$ using (7)
end for
Calculate $p_{\text{DMSR}}^{[1]}, \cdots, p_{\text{DMSR}}^{[T-N]}$ using (10)
Calculate C_{DMSR} using (11)
Calculate O_{DMSR} using (12)

Output: Driving stability features C_{DMSR} and O_{DMSR}

Hence one can compare the features of two drivers to evaluate their skills. The complete algorithm to derive these features from the driving data is shown in Algorithm 1. Using such features to describe the different levels of driving stability can facilitate grouping drivers into clusters according to their skills, to allow further investigations on driving behaviors. In the next section, we apply the proposed algorithm to our dataset. Each individual driver's stability level can be measured, according to which the drivers are clustered into groups to study the common driving stability pattern. Drivers isolated from the majority can be identified and sent for additional inspection.

V. NATURALISTIC DRIVING DATA ANALYSIS

A. INDIVIDUAL DRIVER ANALYSIS

We use the SPMD dataset to demonstrate the results. Only the low-speed scenario is shown. For the acceleration data, the mean value of C_{DMSR} for all 42 drivers in our dataset is 2.80×10^{-3}, and the standard deviation is 2.28×10^{-4}. This leads to the coefficient of variation (CV) to be 8.13%. Fig. 8 shows the acceleration DMSR box plots of four drivers. A clear difference between the box size of *driver*_42 and those of others can be observed. The average level of frequent and small data statistical changes of *driver*_42 in smooth driving phases is the smallest, which implies the highest driving stability in the longitudinal direction. Further, *driver*_11 has the largest C_{DMSR}. But the difference between those of *driver*_12 and *driver*_20 is not sufficiently large. Later we will see that these three drivers can be clustered into the same group according to the stability levels reflected by their driving data.

The mean, standard deviation, and CV of O_{DMSR} for all 42 drivers are 5.50×10^{-3}, 5.11×10^{-4}, and 9.28%, respectively. From Fig. 8, it is seen that, although *driver*_12 and *driver*_20 have the similar box sizes, the outliers for *driver*_20 are more dispersed and farther away from $\tilde{p}_{\text{DMSR}}$ than those of *driver*_12. This leads to a larger value of O_{DMSR}. The result implies that the average level of long and large data statistical changes of *driver*_20 corresponding to action phases is greater than that of *driver*_12.

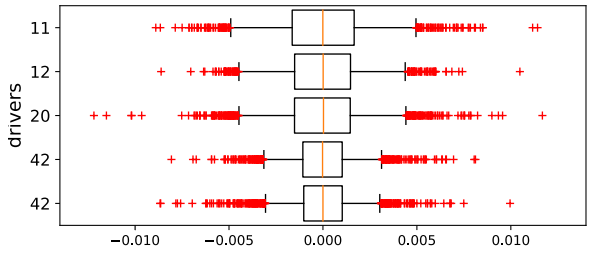

FIGURE 8. Acceleration DMSR characteristics of four drivers in the low-speed scenario. The C_{DMSR} values are respectively 3.30×10^{-3}, 2.97×10^{-3}, 2.97×10^{-3}, 2.10×10^{-3}, and 2.06×10^{-3}. The O_{DMSR} values are 6.36×10^{-3}, 5.27×10^{-3}, 5.70×10^{-3}, 4.10×10^{-3}, and 4.18×10^{-3}.

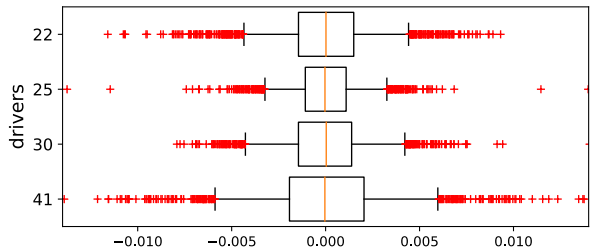

FIGURE 9. Steering DMSR characteristics of four drivers in the low-speed scenario. The C_{DMSR} values are respectively 2.95×10^{-3}, 2.17×10^{-3}, 2.83×10^{-3}, and 3.97×10^{-3}. The O_{DMSR} values are 5.95×10^{-3}, 4.36×10^{-3}, 5.32×10^{-3}, and 9.55×10^{-3}.

The similar analysis can also be conducted for the control in the lateral direction. We attain 42 different values of C_{DMSR}, with mean value 2.75×10^{-3}, standard deviation 3.80×10^{-4}, and CV 13.82%, and 42 O_{DMSR}, with mean value 5.53×10^{-3}, standard deviation 9.36×10^{-4}, and CV 16.91%. Both CV values are larger than those derived from acceleration data. This may imply that drivers tend to behave more differently for lateral-direction operations. The box plots of four example drivers' steering DMSR are shown in Fig. 9. Different levels of box size and outliers' dispersion can also be observed. The results imply that *driver*_25 has higher driving stability in smooth driving phases. The steering angle data of *driver*_41 exhibit large-magnitude variations than other drivers, which presents a relatively lower level of driving stability.

Finally, to illustrate the consistency of the proposed features, we choose the driver, *driver*_42, who has a sufficiently large number of data trips in the low-speed scenario. Two acceleration data matrices are generated using randomly sampled non-overlapping trips. The box plots of DMSR of both matrices are shown in Fig. 8. Their characteristics are very similar, since they reflect the behavior of the same driver. All the above observations demonstrate the effectiveness of our method.

B. DRIVING STABILITY CLUSTERING

The above driving stability analysis results can potentially be adopted to facilitate further investigations on driving behaviors. Due to the lack of label information in naturalistic driving data that can be matched to the safety level, it is in general hard to supervise a learning algorithm to determine the exact knowledge that can distinguish safe and dangerous behaviors.

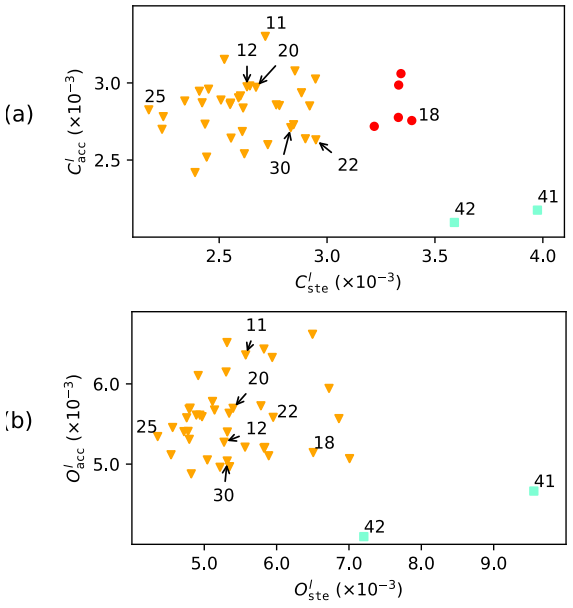

FIGURE 10. Clustering results with (a) $[C^l_{\text{acc}}, C^l_{\text{ste}}]$, and (b) $[O^l_{\text{acc}}, O^l_{\text{ste}}]$.

However, it is commonly accepted that if a driver's behavior is significantly deviated from that of the majority of normal drivers (e.g., driving over-cautiously or over-aggressively), he/she may become a dangerous factor to others and thus should be identified for further inspection [26], [56]. Therefore, we demonstrate a potential application of the proposed features by applying the DBSCAN [21] to help find the common and uncommon driving patterns, from the perspective of our driving stability analysis.

The basic idea behind the DBSCAN algorithm is to partition data points into individual groups such that the density of data points inside a cluster is much higher than that of points outside. Hence a cluster always contains at least a certain number (three in our experiment) of closely-located data points. The algorithm can discover clusters of arbitrary shape, without the necessity of determining the number of clusters in advance. Data points that are not included in any cluster are considered to be *noise objects*. Noise objects and clusters far from the majority can be further studied.

We denote the DMSR concentration intervals for acceleration and steering angle data in the low-speed scenario as C^l_{acc} and C^l_{ste} respectively. The DMSR dispersion degrees are denoted by O^l_{acc} and O^l_{ste}. In the high-speed scenario, the corresponding features are represented by C^h_{acc}, C^h_{ste}, O^h_{acc} and O^h_{ste}. For the ease of illustration, we separate the smooth driving phases and action phases so that our DBSCAN algorithm is applied individually to four two-dimensional feature vectors, $[C^l_{\text{acc}}, C^l_{\text{ste}}]$, $[O^l_{\text{acc}}, O^l_{\text{ste}}]$, $[C^h_{\text{acc}}, C^h_{\text{ste}}]$, and $[O^h_{\text{acc}}, O^h_{\text{ste}}]$. More generally, one can execute an clustering algorithm according to the 4 features in each speed scenario, or even all 8 features to discover complicated patterns in the high-dimensional space.

For instance, Fig. 10(a) shows the clustering result based on $[C^l_{\text{acc}}, C^l_{\text{ste}}]$, which partitions the 42 drivers according to the changing levels of data statistics for smooth driving

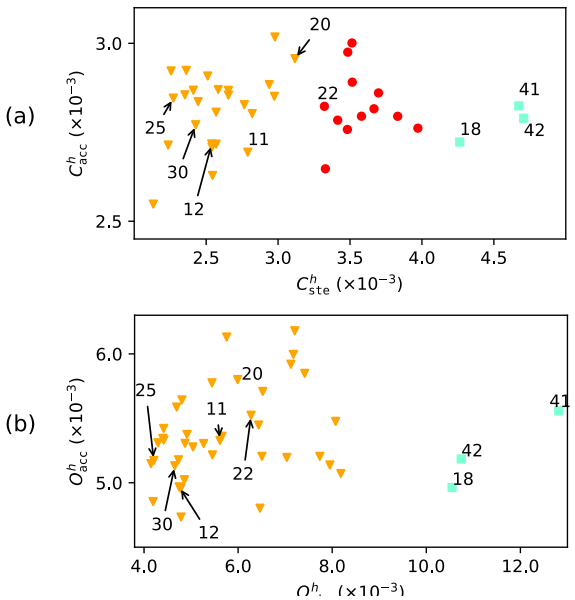

FIGURE 11. Clustering results with (a) $[C^h_{acc}, C^h_{ste}]$, and (b) $[O^h_{acc}, O^h_{ste}]$.

phases in the low-speed scenario. We can see that the variation of C^l_{acc} among the drivers is smaller than that of C^l_{ste}, which is in line with the CV result. This implies that the lateral control skill may be more distinguishable than the longitudinal control skill. The majority of the drivers are partitioned into two clusters, marked by yellow triangles and red circles respectively. The main difference between the two groups is the lateral driving stability level. The two drivers, *driver*_41 and *driver*_42, cannot be included in any group and hence are labeled as noise objects (blue squares). Both drivers exhibit notably smaller values of C^l_{acc} but larger values of C^l_{ste}, compared with the others in the dataset. Further investigations on these two drivers can be conducted to identify the causes of such a pattern. The similar analysis can be conducted on the clustering results using other feature pairs. Fig. 10(b) illustrates the stability pattern of action phases. *driver*_41 and *driver*_42 are again regarded as noise objects. The former has a significantly larger value of O^l_{ste} than other drivers, and the latter has notably lower value of O^l_{acc} compared with the majority. Other drivers are all in one group, because they do not have significantly different driving stability levels, both in the longitudinal and lateral directions.

Fig. 11 shows the results in the high-speed scenario. Compared with the low-speed scenario, the difference of driving stability in the lateral direction becomes more notable. In both phases, three drivers are identified to show clear difference from other drivers. They may deserve additional attention. Apart from finding uncommon drivers, the summary of the common driving stability patterns may also be useful in optimizing future ITS applications, e.g., allowing robot drivers to mimic human drivers.

VI. CONCLUSION

We have proposed a novel approach for analyzing driving stability using naturalistic driving data. On the assumption that sensor measurement noise is a stationary Gaussian process and a theoretically ideal driver can maintain constant vehicle states when the road condition is not taken into account, our method can extract two features by evaluating the statistical difference between the driving data and the data that would be generated by the ideal driver. Specifically, the acceleration and steering angle data of each driver have been organized in matrix forms, to respectively represent the control operations in the longitudinal and lateral directions. Based on RMT, we have presented an algorithm that can derive a parameter termed DMSR according to the LES. Through a number of case studies with synthetic data, we have shown that the concentration level and outliers' dispersion level of the DMSR can help measure the data statistical changes in the driving data matrices and thus can imply the driving skill of a driver in both smooth driving and action phases. The execution of the proposed method on a practical dataset produced by ITS IoV technologies has been demonstrated. Using the extracted features, driver clustering can be applied to discover patterns of drivers. The results can potentially be used to help better understand human drivers.

REFERENCES

[1] B. Hamilton-Baillie and P. Jones, "Improving traffic behaviour and safety through urban design," *Proc. Inst. Civil Eng.*, vol. 158, no. 5, pp. 39–47, May 2005.
[2] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," *Proc. IEEE*, vol. 91, no. 12, pp. 2043–2067, Dec. 2004.
[3] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.
[4] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2377–2396, 4th Quart., 2015.
[5] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.
[6] A. Sathyanarayana, P. Boyraz, and J. H. L. Hansen, "Driver behavior analysis and route recognition by hidden Markov models," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, Sep. 2008, pp. 276–281.
[7] W. Wang, J. Xi, and H. Chen, "Modeling and recognizing driver behavior based on driving data: A survey," *Math. Problems Eng.*, vol. 2014, pp. 1–20, Feb. 2014.
[8] S. Lefevre, A. Carvalho, and F. Borrelli, "A learning-based framework for velocity control in autonomous driving," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 1, pp. 32–42, Jan. 2016.
[9] M. Hatakka, E. Keskinen, N. P. Gregersen, A. Glad, and K. Hernetkoski, "From control of the vehicle to personal self-control; broadening the perspectives to driver education," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 5, no. 3, pp. 201–215, Sep. 2002.
[10] A. Liu and D. Salvucci, "Modeling and prediction of human driver behavior," in *Proc. Int. Conf. Hum.-Comput. Interact.*, Aug. 2001, pp. 1479–1483.
[11] E. Boer and M. Hoedemaeker, "Modeling driver behavior with different degrees of automation: A hierarchical decision framework of interacting mental models," in *Proc. Eur. Annu. Conf. Hum. Decis. Making Manual Control*, Dec. 1998, pp. 63–72.
[12] Y. Zhang, W. C. Lin, and Y.-K.-S. Chin, "A pattern-recognition approach for driving skill characterization," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 905–916, Dec. 2010.
[13] M. V. D. Hulst, "Anticipation and the adaptive control of safety margins in driving," *Ergonomics*, vol. 42, no. 2, pp. 336–345, Feb. 1999.

[14] W. Wang, C. Liu, and D. Zhao, "How much data are enough? A statistical approach with case study on longitudinal driving behavior," *IEEE Trans. Intell. Veh.*, vol. 2, no. 2, pp. 85–98, Jun. 2017.

[15] E. Olsen and W. Wierwille, "A unique approach for data analysis of naturalistic driver behavior," SAE, Warrendale, PA, USA, Tech. Rep. 2001-01-2518, 2001.

[16] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1609–1615.

[17] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot, "An unsupervised approach for inferring driver behavior from naturalistic driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3325–3336, Dec. 2015.

[18] Z. Bai and J. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*. Cham, Switzerland: Springer, 2010.

[19] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 674–686, Mar. 2017.

[20] [Online]. Available: https://www.its.dot.gov/data/

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Aug. 1996, pp. 226–231.

[22] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3017–3032, Dec. 2015.

[23] D. Hallac, A. Sharang, R. Stahlmann, A. Lamprecht, M. Huber, M. Roehder, R. Sosic, and J. Leskovec, "Driver identification using automobile sensor data from a single turn," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 953–958.

[24] B. Wallace, R. Goubran, F. Knoefel, S. Marshall, M. Porter, and A. Smith, "Driver unique acceleration behaviours and stability over two years," in *Proc. IEEE Int. Congr. Big Data (BigData Congr.)*, Jun. 2016, pp. 230–235.

[25] G. Li, S. E. Li, B. Cheng, and P. Green, "Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities," *Transp. Res. C, Emerg. Technol.*, vol. 74, pp. 113–125, Jan. 2017.

[26] C. Miyajima, H. Ukai, A. Naito, H. Amata, N. Kitaoka, and K. Takeda, "Driver risk evaluation based on acceleration, deceleration, and steering behavior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1829–1832.

[27] Y. Zheng and J. H. L. Hansen, "Unsupervised driving performance assessment using free-positioned smartphones in vehicles," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Rio de Janeiro, Brazil, Nov. 2016, pp. 1598–1603.

[28] V. Gadepally, A. Krishnamurthy, and U. Ozguner, "A framework for estimating driver decisions near intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 637–646, Apr. 2014.

[29] Y. Akagi and P. Raksincharoensak, "An analysis of an elderly driver behaviour in urban intersections based on a risk potential model," in *Proc. Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Nov. 2015, pp. 1627–1632.

[30] N. C. Fung, B. Wallace, A. D. C. Chan, R. Goubran, M. M. Porter, S. Marshall, and F. Knoefel, "Driver identification using vehicle acceleration and deceleration events from naturalistic driving of older drivers," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, May 2017, pp. 33–38.

[31] G. S. Aoude, V. R. Desaraju, L. H. Stephens, and J. P. How, "Driver behavior classification at intersections and validation on large naturalistic data set," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 724–736, Jun. 2012.

[32] G. Li, Y. Wang, F. Zhu, X. Sui, N. Wang, X. Qu, and P. Green, "Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China," *J. Saf. Res.*, vol. 71, pp. 219–229, Dec. 2019.

[33] G. Li, W. Lai, X. Sui, X. Li, and Y. Li, "Influence of traffic congestion on driver behavior in post-congestion driving," *Accident Anal. Prevention*, vol. 141, pp. 1–10, Jun. 2020.

[34] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 595–607, Mar. 2017.

[35] S. Ramyar, A. Homaifar, A. Karimoddini, and E. Tunstel, "Identification of anomalies in lane change behavior using one-class SVM," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, pp. 4405–4410.

[36] D. Filev, J. Lu, K. Prakah-Asante, and F. Tseng, "Real-time driving behavior identification based on driver-in-the-loop vehicle dynamics and control," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2009, pp. 2020–2025.

[37] J. Wang, M. Lu, and K. Li, "Characterization of longitudinal driving behavior by measurable parameters," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2185, no. 1, pp. 15–23, Jan. 2010.

[38] X. Wang, M. Chen, M. Zhu, and P. Tremont, "Development of a kinematic-based forward collision warning algorithm using an advanced driving simulator," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2583–2591, Sep. 2016.

[39] X. Wang, M. Zhu, M. Chen, and P. Tremont, "Drivers' rear end collision avoidance behaviors under different levels of situational urgency," *Transp. Res. C, Emerg. Technol.*, vol. 71, pp. 419–433, Oct. 2016.

[40] I. El-Shawarby, H. Rakha, V. W. Inman, and G. W. Davis, "Evaluation of driver deceleration behavior at signalized intersections," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2018, no. 1, pp. 29–35, Jan. 2007.

[41] T. Dingus, S. Klauer, V. Lewis, A. Petersen, S. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. Doerzaph, J. Jermeland, and R. Knipling, "The 100-car naturalistic driving study, phase II—Results of the 100-car field experiment," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. DOT HS 810 593, Apr. 2006.

[42] S. C. Marshall, M. Man-Son-Hing, M. Bédard, J. Charlton, S. Gagnon, I. Gélinas, S. Koppel, N. Korner-Bitensky, J. Langford, B. Mazer, A. Myers, G. Naglie, J. Polgar, M. M. Porter, M. Rapoport, H. Tuokko, B. Vrkljan, and A. Woolnough, "Protocol for candrive II/ozcandrive, a multicentre prospective older driver cohort study," *Accident Anal. Prevention*, vol. 61, pp. 245–252, Dec. 2013.

[43] D. Bezzina and J. Sayer, "Safety pilot model deployment: Test conductor team report," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. DOT HS 812 171, 2015.

[44] M. Maile, F. Ahmed-Zaid, L. Caminiti, J. Lundberg, P. Mudalige, and C. Pall, "Cooperative intersection collision avoidance system limited to stop sign and traffic signal violations," Nat. Highway Traffic Safety Admin., Washington, DC, USA, Tech. Rep. DOT HS 811 048, Oct. 2008.

[45] D. Committee, "Dedicated short range communications (DSRC) message set dictionary," SAE, Warrendale, PA, USA, Tech. Rep. J2735, Nov. 2009.

[46] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011.

[47] *Study on LTE-Based V2X Services (Release 14)*, document 36.885 V14.0.0, 3rd Generation Partnership Project (3GPP), Sophia Antipolis Valbonne, France, Jun. 2016.

[48] *NR; Study on Vehicle-to-Everything (Release 16)*, document 38.885 V2.0.0, 3rd Generation Partnership Project (3GPP), Sophia Antipolis Valbonne, France, Mar. 2019.

[49] H. Muller and J. Freytag, "Problems, methods, and challenges in comprehensive data cleansing," Humboldt Univ. Berlin, Berlin, Germany, Tech. Rep. HUB-IB-164, 2003.

[50] Michigan Department of Transportation. *Public Act 445–Increased Speed Limits on Freeways & Non-Freeways, Hospital Zones, Truck & Bus Speed Limits*. [Online]. Available: https://www.michigan.gov/mdot/0,4616,7-151-9615_79223- - -,00.html

[51] N. Cao, C. Shi, S. Lin, J. Lu, Y. Lin, and C. Lin, "Targetvue: Visual analysis of anomalous user behaviors in online communication systems," *IEEE Trans. Visual. Comput. Graph.*, vol. 22, no. 1, pp. 280–289, Jan. 2016.

[52] A. Guionnet, M. Krishnapur, and O. Zeitouni, "The single ring theorem," *Ann. Math.*, vol. 174, no. 2, pp. 1189–1217, Sep. 2011.

[53] R. Qiu and P. Antonik, *Smart Grid using Big Data Analytics*. Hoboken, NJ, USA: Wiley, 2015.

[54] A. Lytova and L. Pastur, "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries," *Ann. Probab.*, vol. 37, no. 5, pp. 1778–1840, Sep. 2009.

[55] Y. Chen, A. J. Goldsmith, and Y. C. Eldar, "Backing off from infinity: Performance bounds via concentration of spectral measure for random MIMO channels," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 366–387, Jan. 2015.

[56] D. Miles and G. Johnson, "Aggressive driving behaviors: Are there psychological and attitudinal predictors?" *Transp. Res. F, Traffic Psychol. Behav.*, vol. 6, no. 2, pp. 147–161, Jun. 2003.

KAI SONG received the B.E. degree in communication engineering from Tongji University, Shanghai, China, in 2014, where he is currently pursuing the Ph.D. degree in control science and engineering. In 2017, he was a Visiting Student with the College of Engineering, Mathematics and Physical Sciences, University of Exeter, U.K. His research interests include intelligent transportation systems, vehicular data analysis, and random matrix theory.

FUQIANG LIU (Member, IEEE) received the bachelor's degree from Tianjin University, Tianjin, China, in 1987, and the Ph.D. degree from the China University of Mining and Technology, Xuzhou, China, in 1996. In 2005, he was a Visiting Scholar with University Erlangen-Nürnberg, Erlangen, Germany. He is currently a Professor with the School of Electronics and Information Engineering, Tongji University, Shanghai, China. He also serves as the Director of the Broadband Wireless Communication and Artificial Intelligence Laboratory, Tongji University. His research interests include information and communications technologies and innovation applications in automotive and intelligent transportation systems.

CHAO WANG (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2003, and the M.Sc. and Ph.D. degrees from the University of Edinburgh, Edinburgh, U.K., in 2005 and 2009, respectively. In 2008, he was a Visiting Student Research Collaborator with Princeton University, Princeton, USA. From 2009 to 2012, he was a Postdoctoral Research Associate with the KTH-Royal Institute of Technology, Stockholm, Sweden. From 2018 to 2020, he was a Marie Curie Fellow with the University of Exeter, Exeter, U.K. He is currently an Associate Professor with Tongji University, Shanghai, China. His research interests include information theory and signal processing for wireless communication networks, and data-driven research and applications for smart city and intelligent transportation systems.

PING WANG (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, in 2007. He is currently an Associate Professor with the Department of Information and Communication Engineering, Tongji University. His main research interests include routing algorithms and resource allocation of wireless networks (especially for VANETs). He also focuses on developing prototype systems for connected vehicles and building test bed for intelligent vehicles based on MEC.

GEYONG MIN (Member, IEEE) received the B.Sc. degree in computer science from the Huazhong University of Science and Technology, China, in 1995, and the Ph.D. degree in computing science from the University of Glasgow, U.K., in 2003. He is currently a Professor of High Performance Computing and Networking with the Department of Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, U.K. His research interests include computer networks, wireless communications, parallel and distributed computing, ubiquitous computing, multimedia systems, and modeling and performance engineering.